Métodos de predicción de precios para los productos agrícolas en Santander, mediante técnicas de aprendizaje automático.

Juan David Márquez González

Laura Vanessa Montaño Pérez

Trabajo de grado para optar al título de Ingeniero Industrial

Director:

Henry Lamos Díaz Ph.D. en Matemática - Física

Codirector:

Leonardo Hernán Talero Sarmiento M.Sc. Ingeniería Industrial

Universidad Industrial de Santander

Facultad de Ingenierías Fisicomecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2019

AGRADECIMIENTOS

A Dios y a nuestros padres, por apoyarnos y permitirnos terminar este proceso de carrera universitaria, gracias por su amor, incondicionalidad y consejos.

Al grupo de Investigación OPALO, por permitirnos participar y aprender de sus conocimientos para el desarrollo de este proyecto.

A nuestro director de proyecto Henry Lamos Díaz, por aceptar dirigir nuestro proyecto, gracias por su guía y enseñanzas.

Agradecemos a nuestro codirector Leonardo Hernán Talero Sarmiento por creer en nosotros, por su apoyo, guía, paciencia, comprensión, por compartirnos sus conocimientos y sacar lo mejor de nosotros durante este proceso.

A nuestros amigos más cercanos, por su colaboración, motivación y compartir nuestras alegrías y triunfos.

A la Universidad Industrial de Santander y a todos los docentes, por su guía, enseñanzas y ser parte de nuestra formación como Ingenieros Industriales

¡Muchas Gracias!

DEDICATORIA

A Dios, Familia y Amigos, ¡Muchas Gracias!

Juan David Márquez González

A Dios, por su sobrenatural amor, por ser mi fortaleza y guía en todo este proceso

A mis padres Laid Perez y Franklin Montaño, por su ejemplo, apoyo y paciencia

A mi familia y amigos, por animarme y compartir conmigo la alegría de este triunfo

Todo el amor para ustedes.

Laura Vanessa Montaño Perez

Tabla de contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

Resumen

Título: Métodos de predicción de precios para los productos agrícolas en Santander, mediante técnicas de aprendizaje automático*

Autores:

Juan David Márquez González**

Laura Vanessa Montaño Pérez**

Palabras clave: Modelos Autorregresivos, Productos Agrícolas, Pronóstico de Precios, Redes Neuronales.

Descripción:

El análisis de los precios del sector agropecuario constituye un campo de estudio importante, debido a los diversos factores que influyen en el comportamiento de sus precios y la dificultad que estos presentan para ser predichos; por lo tanto, pronosticar los precios del sector es un tema de gran complejidad ya que factores como la relación demanda-oferta, las necesidades del agricultor, el clima, periodos de cosecha y elementos inherentes a los datos (de los cuales son factores de los que no siempre se posee información histórica) no permiten generar modelos precisos. De esta forma, se desarrolla el presente proyecto el cual se enfoca en el estudio de series de tiempo de los precios de venta de productos agrícolas registrados en la central de abastos de Bucaramanga: banano, mandarina, naranja, papa, plátano, piña, tomate y yuca, en los días martes, jueves y viernes durante el periodo comprendido entre el 12 de junio del 2012 y el 30 de abril del 2018, con el objetivo de soportar la toma de decisiones por parte de los involucrados (organizaciones, familias campesinas, gobierno etc.). En el primer objetivo del proyecto se lleva a cabo una revisión de literatura que permite identificar a las RNA y el modelo ARIMA como los modelos más utilizados; en el segundo objetivo se analizan los datos previo al procesamiento en los modelos y son desarrollados los modelos propuestos ARIMA-RNA *feedforward*, RNA *backpropagation* y ANFIS; en el tercer objetivo se comparan los modelos de pronósticos mediante los indicadores de desempeño de error RMSE, MAPE, el tiempo de procesamiento de los modelos y un análisis de varianzas con el que se identifica que los resultados no dependen del modelo; siendo el mejor modelo el obtenido por el modelo ARIMA (seleccionado a través del tiempo de procesamiento) seguido del ANFIS.

*Trabajo de grado

**Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales.

Director PhD (c) Henry Lamos Díaz. Co-director M.Sc. Leonardo Hernán Talero Sarmiento

Abstract

Title: Price prediction methods for agricultural products in Santander, using machine learning techniques*

Authors:

Juan David Márquez González**

Laura Vanessa Montaño Pérez**

Keywords: Autoregressive Models, Farm products, Price Forecasting, Neural Networks.

Description:

Price analysis of agricultural sector is an important field of study due to many factors that make influence in the behavior of their prices and the difficulty to be predicted. Therefore, agricultural price forecasting is a subject of great complexity due to factors as the demand-supply relationship, farmer´s needs, the weather, harvest periods and elements inherent in the data (factors that do not always have historical information) that do not allow to generate precise models. In this way the present project is developed focusing on the study of time series of the sale prices of agricultural products registered in the Bucaramanga stock exchange: banana, tangerine, orange, potato, green plantain, pineapple, tomato and yucca, on tuesdays, thursdays and fridays during the period between june 12, 2012 and april 30, 2018, with the objective of supporting decision-making by those involved (organizations, peasant families, government, etc.). The first objective of the project is a literature review that identify that ANN and ARIMA model are the most used models. In the second objective, is analyzed the data prior to the processing in the proposed models (ARIMA-ANN *feedforward*, ANN backpropagation and ANFIS) and the models are developed. In the third objective, the forecast models are compared through the error performance indicators RMSE, MAPE, the processing time of the models and an analysis of variance (ANOVA) which is identified that the results do not depend on the model. It is determined that the best model is obtained by the ARIMA model (selected through the processing time) followed by the ANFIS model.

*Trabajo de grado

**Facultad de Ingenierías Físico-Mecánicas. Escuela de Estudios Industriales y Empresariales.

Director PhD (c) Henry Lamos Díaz. Co-director M.Sc. Leonardo Hernán Talero Sarmiento

Introducción

Colombia es un país en vía de desarrollo, el cual basa su economía principalmente en la explotación de materias primas y en el sector agropecuario, donde este último representa la ocupación de 3’768.700 campesinos al 2017 (Departamento Administrativo Nacional de Estadística, 2017); encontrando que inmediatamente en el año anterior, según cálculos del Ministerio de Agricultura y Desarrollo Rural:

> *"el 83.5% del total de los alimentos que consumen los colombianos lo producen las manos de nuestros agricultores, registrando un índice de autosuficiencia alimentaria positivo, margen significativo e importante para el sector”* (Ministerio de Agricultura, 2016).

Se evidencia el fuerte impacto que tiene la producción del sector en el desarrollo económico, empleo y suministro de alimentos a nivel nacional, departamental y municipal, lo que ilustra la gran importancia que representa para la nación y sus divisiones territoriales; en Santander hay un comportamiento similar, donde este contribuye al “Producto Interno Bruto (PIB) Departamental con un 6.8%” (Camara de Comercio de Bucaramanga, 2016), además de que de los “87 municipios que conforman el departamento, 78 basan sus actividades en la agricultura representándolo como el principal renglón de la economía” (Secretaría de Agricultura, 2017).

La realidad con respecto a los precios de los productos agrícolas que se manejan en el sector no es clara ni definida, debido a su inestabilidad y no tener un comportamiento específico. La dificultad para la predicción de los precios en gran medida se justifica por las condiciones meteorológicas que entorpecen las labores logísticas, diversas estructuras productivas, cambios en la política agrícola, fluctuaciones en la demanda y el número de intermediarios que existen entre

el agricultor y el mercado, dando como resultado que predecir los **"**(…) precios provenientes de un sector agrícola-ganadero no es lo mismo que hacerlo sobre cualquier otro bien, dado que tienen características especiales" (Masaro, 2016).

En consecuencia, desde hace un tiempo existen un conjunto de investigaciones enfocadas en el estudio de los precios de los productos agrícolas y los factores que enmarcan y definen su comportamiento; los estudios se encuentran liderados por investigadores de países como Estados Unidos, China, India y Brasil, los cuales se han enfocado en el análisis de series de tiempo y determinadas variables exógenas para la obtención de pronósticos mediante el uso de múltiples modelos estadísticos, empleando para ello técnicas de aprendizaje automático. A continuación, se exponen algunos ejemplos de las investigaciones realizadas:

- Un modelo de pronóstico híbrido del precio de la yuca basado en una Red Neuronal Artificial y Máquinas de Soporte Vectorial (Polyiam & Boonrawd, 2017), el cual busca predecir el precio de la yuca por medio del uso de 4 técnicas de Aprendizaje Automático (RNA, Maquina de Soporte Vectorial, el vecino más próximo y la técnica híbrida) aplicadas a una seria temporal de precios de 11 años (2005 – 2015)
- Modelo de pronóstico y validación del precio de productos acuáticos basado en series de tiempo GA-SVR, (Duan, Zhang, Wei, Xiao, & Wang, 2017). Los autores aplican un Algoritmo Genético para optimizar los parámetros del Regresor de Soporte Vectoria, y este último para la predicción de los precios, utilizando los datos de los precios de los peces mandarín, camarones y cangrejos, de enero de 2011 a diciembre de 2015 del sitio web del mercado Beijing Xinfadi.

Colombia por su parte, por medio del Departamento Nacional de Planeación (DNP) se encuentra desarrollando un proyecto que consta de realizar un seguimiento a los precios de los productos del sector agropecuario a nivel nacional, con el que pretende conocer las condiciones que enmarcan su comportamiento, denominado: "Big Data para la predicción de productos agrícolas" (Departamento Nacional de planeacion, 2017), lo que permite demostrar la disposición que tiene el gobierno en el análisis de precios del sector primario.

Por lo tanto, en el presente proyecto se propone el análisis de los precios de algunos productos agrícolas de la central de abastos de Bucaramanga, buscando contribuir al mejoramiento de la región y a la meta de "hacer del campo Santandereano un sector competitivo, eje fundamental para el crecimiento económico y social de sus comunidades" (Secretaría de Agricultura, 2017) a través de la formulación, prueba y comparación de tres modelos basados en Redes Neuronales Artificiales (RNA), seleccionando el mejor modelo de predicción según criterios de precisión en instancias de Benchmarking para cada uno de los productos escogidos. El proyecto igualmente plantea desde el ámbito académico el uso de competencias estadísticas e ingenieriles relacionadas con la programación de modelos para la predicción y el análisis de datos históricos para la generación de información confiable, con el objetivo de contribuir a la toma de decisiones relacionadas con inversiones y periodos de planeación de producción y a la mejora de la productividad en empresas del sector basándose en argumentos reales y acciones concretas (Acevedo Borrego & Linares Barrantes, 2012) demostrando la importancia de la ingeniería industrial y del proyecto en el campo, asistiendo de esta manera a la región, organizaciones y familias campesinas en la gestión y control de recursos en el pequeño, mediano y largo plazo.

Tabla de cumplimiento de objetivos

Tabla 1
Cumplimiento de objetivos de la investigación

Objetivos específicos	Apartado relacionado
1. Realizar una revisión de la literatura sobre la aplicación de modelos basados en máquinas de aprendizaje automático para la predicción de precios.	Capítulo 3
2. Seleccionar diversos modelos de predicción de precios basados en máquinas de aprendizaje automático para cada uno de los productos seleccionados.	Capítulo 3 y capitulo 7
3. Identificar el modelo más preciso para la predicción de los precios, mediante un conjunto de métricas del Benchmarking para los modelos desarrollados.	Capítulo 8
4. Generar un artículo de carácter publicable sobre los resultados obtenidos a través de la investigación.	Apéndice C

1. Generalidades del Proyecto

1.1 Planteamiento del Problema

La actividad agrícola en los últimos años ha adquirido en el país gran importancia para el desarrollo económico, debido a la implementación de leyes que buscan el desarrollo del campo y sus actividades como son: la Ley 101 de 1993 que promueve el "otorgar especial protección a la producción de alimentos, adecuar el sector agropecuario y pesquero a la internacionalización de la economía, sobre bases de equidad, reciprocidad y conveniencia nacional y promover el desarrollo del sistema agroalimentario nacional" (Ministerio de Agricultura y Desarrollo Rural, 1993); y la Ley 160 de 1994, la cual busca la adquisición de tierras a las familias campesinas y el desarrollo económico y equitativo del campo (Ministerio de Agricultura y Desarrollo Rural, 1994), en conjunto con los tratado de Paz y el posconflicto, generando esto un fuerte impacto en la economía de Colombia, un país que depende en gran medida de los productos que cultiva para el sostenimiento alimenticio de millones de familias (Urna de Cristal, 2016).

En Santander la actividad agrícola presenta un comportamiento similar al nacional, donde se posiciona como una actividad económica de gran importancia, ya que representa el 5.7% del PIB del departamento según la Cámara de Comercio de Bucaramanga (2016); además, se encuentra una de las centrales de comercialización de productos agrícolas más importantes del país (Centro Abastos de Bucaramanga), la cual obtuvo el mayor porcentaje de aumento en el abastecimiento de alimentos según ciudad y mercado mayorista entre la primera quincena de noviembre y diciembre de 2017, con una variación del 44.69%, seguido por las Flores de Bogotá con el 34.56% y la Nueva Sexta de Cúcuta con el 29.84% (SIPSA, 2017).

Para el desarrollo de la actividad agrícola a nivel nacional debe tenerse en cuenta aspectos de tipo social, de infraestructura, económicos, etc.; teniendo en este último los precios un papel fundamental, ya que poseen influencia directa en la oferta y la demanda potenciales, los canales de distribución de los productos agrícolas y la economía en la agricultura (Mishra & Singh, 2013), lo que permite evidenciar su relevancia en el crecimiento económico de la nación, pues un trato descuidado a este elemento reflejará un impacto sobre la economía, la seguridad alimentaria, y la ocupación de trabajadores.

De esta forma, conocer el comportamiento de los precios del sector agrícola es fundamental, puesto que, al ser un fenómeno dinámico en el tiempo, variable y con una naturaleza incierta y aleatoria, afecta en gran medida a fenómenos del mercado como los mencionados anteriormente, y a determinados elementos macroeconómicos, como la inflación. Reina, Zuluaga, Bermúdez, & Oviedo (2011) enuncian que los "precios actuales de la mayoría de productos agrícolas son más elevados que los niveles promedio registrados en años recientes. Los precios de muchos productos agrícolas llegaron a un punto máximo en 2008, disminuyeron y luego se recuperaron con fuerza", siendo este el caso de las frutas tropicales y de productos de gran interés para la nación como el cacao, el café, la caña de azúcar y la palma de aceite; igualmente sucede con insumos industriales como la goma y el algodón, algunos alimentos como la soya, el azúcar, el trigo y el maíz.

Cabe resaltar que aunque se observa una fuerte influencia de la actividad agrícola y los precios de sus productos en la economía de Colombia, la mayoría de estudios que se han realizado en la nación con el uso de técnicas de aprendizaje automático se enfocan en otros campos de estudio (con base en la revisión de literatura), relegando a la predicción de precios agrícolas el uso de análisis de series de tiempo, formulas econométricas, modelos regresivos y auto regresivos,

contemplando de esta forma que la Inteligencia Artificial (*machine learning*) ha sido poco abordada, desperdiciando el potencial que podrían ofrecer estos mecanismos a las soluciones de los problemas mencionados, los cuales tratan de dar un mejor panorama en el estudio del comportamiento de los precios de los productos tratados y la obtención de resultados más precisos, permitiendo así la minimización de medidas a tomar en el futuro a corto y largo plazo.

Conforme a lo anterior, y con el propósito de comprender el comportamiento de los precios del sector, el gobierno se ha enfocado en desarrollar una herramienta para suministrar una fuente de precios confiables y constante, basado en los boletines de precios mayoristas que se manejan en la principales centrales de abastos de Colombia, lo cual se encuentra plasmado en el proyecto que está realizando en conjunto con el Departamento Nacional de Planeación (DNP) "Big Data para la predicción de precios agrícolas" cuyo objetivo es el de "Desarrollar una herramienta que permita monitorear el comportamiento de los precios de los productos agropecuarios comercializados en las centrales mayoristas del país" (Departamento Nacional de planeacion, 2017), según información del DANE.

En vista de lo anterior, y con base en el enfoque que el país ha puesto en el estudio del precio de los productos agrícolas y las perspectivas futuras para realizar previsiones de su comportamiento, es prioridad identificar los factores que los pueden afectar, con la intención de generar medidas de control y gestión por parte de los afectados y en mayor medida por parte del gobierno. La identificación de factores como los cambios de la demanda, la cantidad de productos ofrecidos, la influencia de productos suplementarios (FAO, 2001), los altos precios de la energía, los bajos niveles de inventario en el mundo, el clima (M. Reina et al., 2011), entre otros; podrían ayudar a determinar las causas que expliquen estas variaciones, y de esta manera aplicar modelos

de forma acertada que expresen la tendencia de un conjunto de observaciones y predecir con un grado de confiabilidad óptimo su comportamiento futuro (Pernaut Ardanaz & Ortiz Felipe, 2008), permitiendo así disminuir el grado de incertidumbre en los resultados y su posterior toma de decisiones. Por lo tanto, se percibe que el tema de los pronósticos y sus formas de estudio se encuentran en constante reinvención y desarrollo, con el objetivo de obtener un modelo que logre predecir con mayor precisión el futuro de los precios del sector, utilizando diversas metodologías para lograrlo (todo esto nacido de la necesidad de entender lo que conlleva el cambio de los precios en un futuro incierto), como la aplicación de las técnicas basadas en Máquinas de Aprendizaje Automático, las cuales son herramientas que hacen parte de "una rama de la inteligencia artificial que se ocupa del diseño y aplicación de algoritmos de aprendizaje" (Mena, 1999, p. 480), buscando mejorar resultados obtenidos a través de otros mecanismos; dentro de estas máquinas se encuentra una técnica aplicada en mayor medida, las Redes Neuronales Artificiales (RNA).

Sin embargo, dado que en Colombia no existen fuentes de información completas relacionadas con la actividad agrícola (lo que dificulta la definición de variables que afectan los precios y su respectiva información) se plantea el uso de los precios de cada producto como variable predictora, con el objetivo de obtener un modelo que logre predecir con mayor precisión el precio futuro de los productos del sector aportando a la seguridad económica de los involucrados. Se propone con el presente proyecto realizar un estudio enfocado al comportamiento de los precios de los productos agrícolas del departamento de Santander, mediante la comparación de 3 modelos predictivos basados en RNA, seleccionando el mejor según criterios de precisión, que en conjunto con los resultados arrojados en otras investigaciones, contribuya al crecimiento de la actividad agrícola, a la competitividad del sector agropecuario como también a todos los sujetos involucrados (familias campesinas, empresas, Gobierno), con el fin de respaldar la toma de

decisiones, gestionar los recursos y llevar a cabo la planeación de un futuro prometedor para el campo y su habitantes, contemplando la utilidad de su desarrollo en la región y la contribución en futuras investigaciones relacionadas al tema.

2. Objetivos

2.1 Objetivo General

Aplicación de modelos de predicción basados en Maquinas de Aprendizaje Automático, para el pronóstico de precios de productos agrícolas en Santander.

2.2 Objetivos Específicos

- Realizar una revisión de literatura sobre la aplicación de modelos basados en máquinas de aprendizaje automático para la predicción de precios.
- Seleccionar diversos modelos de predicción basados en máquinas de aprendizaje automático para cada uno de los productos seleccionados.
- Identificar el modelo más preciso para la predicción de los precios, mediante un conjunto de métricas del Benchmarking para los modelos desarrollados.
- Generar un artículo de carácter publicable sobre los resultados obtenidos a través de la investigación.

3. Revisión de Literatura

Para el desarrollo de la revisión de literatura se plantea la siguiente ecuación de búsqueda:

(("price predict") OR ("price forecast*") OR ("future price") OR ("predict* methods") OR ("time serie* predict*") OR ("future price*") OR ("modelling price")) AND ("price") AND ("SVR" OR "hybrid method*" OR "decision trees" OR "neural network*")) AND ("Agr* products" OR "agr*" OR "egg*" OR "Farm products" OR "Soft Commodit*" OR "maize" OR "hog" OR "banana" OR "mijo" OR "vegetables" OR "fruits" OR "tubers")) AND (LIMIT-TO (DOCTYPE, "ar ")*

Figura 1: Ecuación de búsqueda

Esta ecuación es ingresada en la base de datos Scopus (base de datos de resúmenes y citas de Elsevier, febrero de 2018) encontrando de esta forma un volumen total de 817 artículos a partir de tres tópicos de búsqueda (predicción de precios, modelos basados en técnicas de aprendizaje automático y productos agrícolas, agropecuarios o commodities), de los que se obtienen 40 artículos relacionados con la investigación luego de realizada la depuración expuesta a través de la Figura 2:

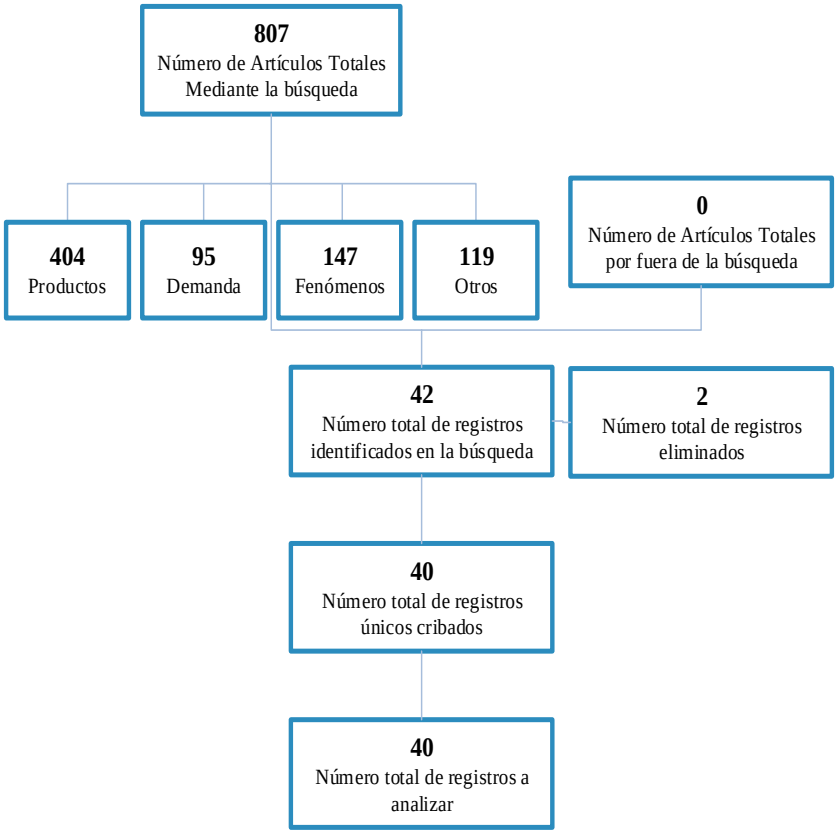

Figura 2: Artículos seleccionados para la revisión de literatura. Adaptado de Scopus (2018)

Los resultados del análisis de los 40 artículos seleccionados se exponen a continuación.

3.1 Análisis de precios

El análisis de los precios futuros por medio de modelos predictivos se presenta en diversos campos, como estudios financieros, trabajos académicos y proyectos de investigación, donde a partir de un pronóstico con bajo margen de error es posible tomar decisiones, siendo de esta forma, la previsión de precios **"**una parte integral del comercio de productos básicos y del análisis de precios" (Girish K Jha & Sinha, 2013, p. 1), además de ser "beneficiosa para orientar correctamente la circulación de los productos agrícolas, la producción agrícola y realizar el equilibrio entre la oferta y la demanda del área agrícola" (Zhanga et al., 2014, p. 1).

De esta manera, el análisis del comportamiento de los precios (entendiendo comportamiento como las diversas características de una serie: tendencia, estacionalidad, homocedasticidad, linealidad entre otras), es un tema que ha sido debatido y estudiado varios años a través del uso de modelos matemáticos como el análisis de regresión lineal, el cual es "un método conveniente para pronosticar una respuesta o una variable dependiente cuando las variables explicativas o independientes, con una asociación razonable, están fácilmente disponibles" (Ahmad & Mariano, 2006), y a pesar de que "durante las últimas décadas se han dedicado muchos esfuerzos para desarrollar y mejorar varios modelos de previsión de series de tiempo" (Mitra & Paul, 2017, p. 1) otorgando mejoras en la previsión y oportunidades para los involucrados, se hace indispensable el examen de técnicas alternativas que generen una mejor predicción con un mayor reconocimiento de patrones dentro del mismo conjunto de datos (Ahmad & Mariano, 2006, p. 2).

Gracias al avance en tecnología y en las ciencias de la computación, se ha dado paso a varias herramientas para mejorar el tema de los pronósticos dentro de las que se destaca el Aprendizaje

Automático, el cual, gracias a su enfoque aplicativo, practico y de bajo costo, permiten contribuir en la actualidad a este campo.

3.2 Redes Neuronales Artificiales (RNA)

Para el análisis de los precios se encuentran diversas técnicas, de las cuales, las más utilizadas en el aprendizaje automático supervisado (según las investigaciones consideradas para el proyecto) son los modelos basados en RNA, SVR, Box Jenkins (autorregresivo integrados de medía móvil - ARIMA) y modelos híbridos generados a partir de estos últimos, en los cuales se hace uso de diferentes herramientas para la limpieza, normalización y el pre procesamiento de los datos. Las RNA son la técnica en el aprendizaje supervisado que mayor aplicación ha tenido en el estudio de la predicción de los precios y series de tiempo, encontrando múltiples autores que han optado por esta y sus diferentes hibridaciones; debido principalmente a las características que posee en comparación con los modelos econométricos tradicionales, la cual al momento del análisis de datos lineales como no lineales, según Zankova (2016) no necesita el uso de suposiciones previas sobre los datos para obtener aproximaciones universales de funciones, el problema de la falta de especificación del modelo les afecta en menor grado, son tolerantes a la presencia de errores en los datos y presentan un costo computacional relativo al tipo de modelamiento a realizar.

3.2.1 Modelos basados únicamente en RNA. Con el uso únicamente de RNA para la predicción de precios en el sector agrícola, destacan los trabajos que se mencionan a continuación; Zhanga et al. (2014) desarrollan un modelo de predicción en torno al precio del tomate en el lapso de un año (iniciando 1 de Enero hasta el 30 de diciembre del 2013).

Usando para ello la inclusión de la transformada *Wavelet* para la limpieza de datos, obteniendo que el porcentaje de error de predicción es menor que 0.01 y la correlación R^2 entre el valor de

predicción y el valor real es 0.908 con el uso del *Wavelet*, lo que demuestra su utilidad en el proceso de predicción; Ahmad & Mariano (2006) luego de observar el complejo trabajo que representan las predicciones de los precios de los huevos y considerando la recopilación durante muchos años de datos que consideran confiables y costosos, optan por el uso de tres RNA (*ward, backpropagation, y RNA general regretion*) para la predicción de los mismos y la comparación con el análisis de regresión, el cual solo explica el 37% de la variación de los precios usando tres predictores (número de gallinas ponedoras, exportación de huevos de EE.UU. y número de pollitos de huevos incubados en las incubadoras comerciales), sin embargo, aunque las RNA explican el mismo porcentaje, lo hacen solo con las series de tiempos de los precios históricos, con lo que consiguen minimizar esfuerzos en controlar las variables exógenas, logrando optimizar la velocidad de aprendizaje y resolución del modelo. Sobresalen también los trabajos realizados por Girish K Jha & Sinha (2013) y Anjoy & Kumar (2017), quienes aplican las RNA de Retardo de Tiempo (TDNN), la cual se implementa para la construcción de la memoria de corto plazo en la estructura de una RNA utilizando Anjoy y Kumar dicho modelo en conjunto con un *Wavelet* para mejorar el rendimiento de la red en el pronóstico del precio de la cebolla.

También se encuentra que los autores Misra & Goswami (2015), Shih, Huang, Chiu, Chiu, & Hu (2009) y Ahmad, Dozier, & Roland Sr. (2001), mediante el uso de las RNA buscan la predicción de los precios de unos determinados productos (productos agrícolas, azúcar, pollo y huevos respectivamente) de forma más precisa y confiable con el fin de minimizar los impactos de estos en los mercados locales, usando para ello diferentes RNA. Los autores Misra y Goswami utilizan el Perceptron Multicapa para el desarrollo del modelo junto con el modelo *Star Garch*, con el objetivo de mejorar el manejo de la volatilidad; los autores Shih, Huang, Chiu, Chiu y Hu utilizan un enfoque ponderado de razonamiento basado en casos (CBR) en conjunto con un

perceptrón multicapa para el pronóstico del pollo, el cual depende de la región; y por último los autores Ahmad, Dozier, & Roland Sr hacen uso de dos modelos de RNA en su estudio, aplicando primero una Red Neuronal basada en regresión general y luego una Red Neuronal con aprendizaje de propagación hacia atrás (retropropagación).

3.2.2 Modelos híbridos basados en RNA y modelos Autorregresivos. Estos son los tipos de modelos más encontrados dentro de la revisión desarrollada en Scopus (2018), donde autores como Mitra & Paul (2017), Xiong, Li, & Bao (2017), G.K. Jha & Sinha (2014), Mishra & Singh (2013), Fahimifard, Salarpour, Sabouhi, & Shirzady (2009) y Kohzadi, Boyd, Kermanshahi, & Kaastra (1996), en sus respectivos trabajos tratan el desarrollo de diversos modelos basados en RNA y ARIMA. En todos estos casos de estudio, los autores indican que los resultados obtenidos por los modelos evaluados de RNA son mejores y más precisos que los obtenidos por los modelos ARIMA, concluyendo que los modelos ARIMA poseen una desventaja respecto al procesamiento de datos no lineales con comportamientos caóticos y volátiles como los que presentan las series.

De estos artículos destacan: los autores Li et al. (2013) quienes establecen el uso de RNA caóticas para la predicción de precios de huevos en China en el periodo de enero de 2008 a diciembre del 2012, la cual es comparada con un modelo ARIMA determinando que las RNA caóticas poseen una capacidad de ajuste no lineal y una mayor precisión en la predicción del precio minorista semanal de los huevos; resultado que también indica que las RNA caóticas pueden ser ampliamente utilizadas en el campo de la predicción a corto plazo de precios de productos agrícolas. Por su parte, los autores Zhu, Pan, Yin, & Li (2013) proponen un enfoque de pretratamiento de datos a través de la normalización de estos y la normalización del orden de magnitud que tiene efecto sobre la previsión de precios, con el objetivo de mejorar los resultados

de la red neuronal antes de operar los datos, encontrando que los resultados arrojados por el SVR son satisfactorios con un MAE de 99.67% normalizado y las RNA con un 98.67% sin normalizar y un 99.36% normalizado, afectando la normalización en menor medida en el caso de los modelos basados en las RNA.

3.2.3 Modelos híbridos basados en Redes Neuronales, SVR y otras técnicas de predicción.

Los artículos encontrados dentro de este apartado en la base de datos relacionados con modelos híbridos entre RNA y SVR son muy pocos, sin embargo, existen unos determinados autores que trataron a estos de forma independiente como son el caso de Xiong, Li, Bao, Hu, & Zhang (2015), estos con base en previas investigaciones de literatura encuentran que la combinación de series de tiempo lineales y no lineales ofrecen mejores resultados que de forma individual, por lo tanto, hacen uso de un vector modelo de corrección de error (VECM) en conjunto con una SVR multisalida; esto con el fin de capturar los patrones lineales y no lineales exhibidos en los precios futuros de productos básicos agrícolas, encontrando que el método VECM-SVR es una alternativa prometedora para intervalos de valores en futuros de precios agrícolas; los autores Mustaffa & Sulaiman (2015) llevan a cabo una comparación entre modelos híbridos y evolutivos, donde enfocan su investigación en un modelo híbrido entre SVM y el *grey wolf optimizer* (GWO). El rendimiento de GWO-LSSVM se desarrolla para el análisis predictivo del precio del oro midiendo los resultados con base en los índices del MAPE, obteniendo el mejor puntaje con 5.45% y del RMSE con 0.0678%, logrando resultados satisfactorios frente a otros modelos basados en GWO, ABC-LSSVM y en computación evolutiva como el GA-LSSVM.

Además, existen trabajos que igualmente combinan a las dos técnicas RNA y SVR, los cuales tienen como objetivos demostrar la precisión de un modelo con respecto a otro mediante sus

desempeños de predicción, usando para ello diferentes índices de evaluación como lo son: el error absoluto medio (MAE), el error porcentual absoluto medio (MAPE), el error porcentual absoluto medio simétrico (SMAPE) y el error cuadrático medio (MSE), etc. En estos trabajos el objetivo principal es encontrar la mejor técnica de predicción haciendo comparaciones entre modelos de RNA, ARIMA y SVR.

3.2.4 Modelos híbridos basados en RNA y diferentes formas de procesamiento de datos. Dentro de los resultados obtenidos en la búsqueda se identifican igualmente artículos enfocados en el desarrollo de modelos híbridos para la predicción del precio, no obstante, estos basan su hibridación en diferentes formas de procesar la información, la limpieza de los datos, el preprocesamiento y la activación de las redes.

De este tipo de modelos, se encuentran que los autores Zhao, Zhang, & Xu (2012) desarrollan un enfoque de RNA híbrida que puede probar el cambio estructural de los enlaces sin un conocimiento previo de la distribución de los datos a introducir, donde luego estudian los vínculos de precios aplicando RNA híbrida a dos precios principales de cultivos bioenergéticos, incluyendo el maíz de Argentina y la soja de Brasil; descubriendo que los puntos de salto en series temporales tienen un efecto negativo en el entrenamiento de la RNA híbrida en el proceso experimental. Cuanto mayor sea el número de puntos de salto en la serie de datos, más tiempo entrenará la red antes de lograr la estabilidad; para eliminar o reducir la influencia de los puntos de salto, los autores proponen el uso de *wavelet* con el modelo híbrido para realizar un filtrado de datos. Los autores Ribeiro & Oliveira (2011) desarrollan su proyecto entorno al uso de RNA para el preprocesamiento de los datos en la previsión de los precios combinándolo con el modelo estocástico *Kalman,* que es capaz de dar cuenta de variables que no pueden ser observadas en el proceso de previsión. Con

su trabajo proponen un modelo híbrido para pronosticar los precios de los productos agrícolas basado en dicho enfoque, el cual se aplica para pronosticar el precio del azúcar. El filtro *Kalman* considera la estructura del proceso estocástico que describe la evolución de los precios y las RNA permiten variables que pueden impactar en los precios de los activos de forma indirecta y no lineal, lo que no se puede incorporar fácilmente en los modelos econométricos tradicionales. Los autores Ayankoya, Calitz, & Greyling (2016)**,** combinan el Big Data (es definida como la capacidad de la sociedad de aprovechar la información de formas novedosas, para obtener percepciones útiles o bienes y servicios de valor significativo" (Mayer-Schonberger & Cukier, 2013)) con las RNA, utilizando el primero en la recolección e integración de datos de varias fuentes para descubrir patrones útiles y extraer información procesable, y las RNA para modelar patrones lineales y no lineales que podrían existir en los conjuntos de datos con el fin de realizar las predicciones; sus resultados indican que al aplicar esta técnica en el manejo de grandes volúmenes de datos hace que sea posible, adquirir e integrar los datos en tiempo real mejorando el rendimiento de las RNA.

De igual manera, Pinheiro & de Senna (2017) utilizan en su investigación la metodología propuesta que combina el modelo RNA en previsión de precios de los productos agrícolas (azúcar, algodón, maíz, café y soja) con el análisis de espectro singular multivariado (MSSA); este último modelo captura la estructura de la serie de tiempo de estudio basado en el conjunto de otras series de tiempo mejorando el rendimiento de las RNA. De Mattos Neto, Cavalcanti, & Madeiro (2017), plantean la aplicación de un sistema hibrido de combinación no lineal (NoLiC) compuesto por dos fases, entrenando en la primera fase dos modelos (uno para la serie temporal y el otro para sus residuales) para en la segunda fase ser combinados empleando un perceptron multicapa (MLP), dando resultados con mayor precisión en comparación con los modelos modernos y los modelos basados en híbridos con combinación lineal; de igual manera, Tao Xiong, Li, & Bao (2017)**,**

generan un modelo hibrido, donde optan por combinar el procedimiento de descomposición de tendencia estacional basado en loess (STL) y las máquinas de aprendizaje extremo (ELM) para pronosticar a corto, mediano y largo plazo, los precios estacionales de los vegetales, usando este modelo para predecir la tendencia y los componentes de residuo de manera independiente.

En la Figura 3 se puede observar la cantidad de modelos encontrados en la revisión de literatura por cada una de las técnicas planteadas. La técnica de mayor uso según la revisión son las redes neuronales, seguidas de los modelos ARIMA, y en última instancia los modelos híbridos y SVR.

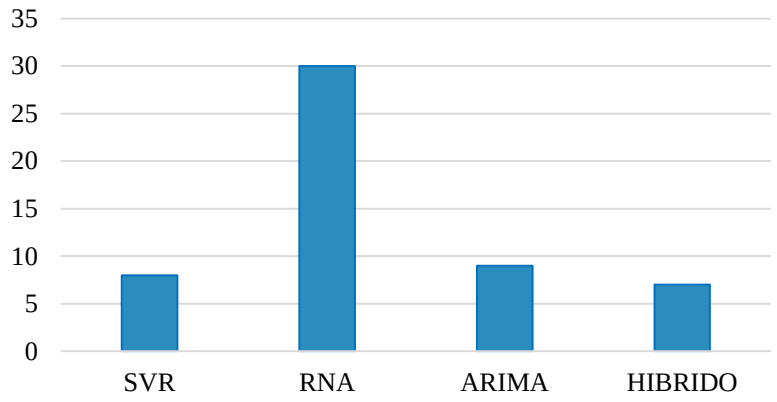

Figura 3: Modelos desarrollados en los artículos seleccionados para la revisión de literatura

Los resultados expuestos en los artículos mencionados anteriormente están sujetos al tipo de producto analizado y pueden variar dependiendo de los periodos de tiempo tomados para el estudio, como también el hecho de usar variables exógenas o series temporales en la predicción, hasta la consideración de las características de las técnicas en el análisis de un determinado producto.

A partir de la información expuesta en la revisión de literatura se plantea el desarrollo de tres modelos basados en redes neuronales artificiales, debido a la precisión al momento de realizar el

pronóstico en comparación con las otras técnicas; los tres modelos propuestos para el proyecto son el modelo de red neuronal con aprendizaje *backpropagation,* el modelo neurodifuso ANFIS y un modelo hibrido entre ARIMA y RNA *feed forward*.

4. Marco de Antecedentes

El estudio del comportamiento del sector agropecuario ha sido un tema de gran interés en los últimos años en el mundo, lo cual ha generado un determinado número de investigaciones y publicaciones respecto al tema (M. Reina et al., 2011); algunos de estos estudios se enfocan en la previsión de los precios agrícolas mediante el uso de diferentes técnicas de predicción, como los análisis con modelos econométricos y el uso de técnicas de aprendizaje automático. A continuación, se presentan tres proyectos de autores que han incursionado en el uso de estas técnicas, buscando obtener un poder predictivo mayor en el desarrollo del mismo proyecto, aplicándolo a diferentes campos académicos, científicos, económicos, etc., utilizando para ellos las diferentes técnicas de Aprendizaje Automático que subsisten para tal fin.

En su trabajo Jabez Harris (2017), "Una aproximación con Máquinas de Aprendizaje Automático al pronóstico de precio de consumo de alimentos" busca la mejor forma de predecir los precios de 8 categorías de alimentos (Carnes, Comida de Mar, Lácteos, Panadería, Fruta, Vegetales, Otros y Restaurantes) del canadiense promedio, donde compara el Índice de Precios al Consumidor (IPC) de Canadá junto con la evaluación de tres diferentes modelos de datos de pronóstico, Holt-Winters, informe de precios de alimentos y mercados financieros basados en futuros, ejecutados en tres diferentes técnicas de aprendizaje automático (redes neuronales artificiales: perceptron multicapa, máquinas de soporte vectorial: SVMOreg y árbol de decisión:

árbol M5P), los cuales son evaluados en la investigación por medio del error porcentual absoluto medio (MAPE), obteniendo un error del 9.78% para la regresión lineal, 4.412% para las redes neuronales, 6.808% para la máquina de soporte vectorial y 4.417% para el árbol de decisión; cabe destacar que aunque los resultados de la redes neuronales y el árbol de decisión son similares, las redes neuronales obtuvieron los mejores resultados en 5 de las 8 categorías de productos evaluadas, carnes, lácteos, panadería, restaurantes y otros, afirmando su capacidad de asimilación en el comportamiento aleatorio que poseen los precios; sin embargo, el mejor comportamiento a nivel de productos como la fruta y los vegetales se obtuvieron en las máquinas de soporte vectorial y el árbol de decisión respectivamente, existiendo una diferencia significativa de estos últimos resultados en comparación con la de los obtenidos en los demás modelos, encontrándose también que el mejor modelo de datos de pronóstico lo otorgó el mercado financiero basado en futuros.

Zeleckis (2015), en su trabajo de maestría "Biocombustible Solido – Modelación de precios de las virutas de madera en Lituania usando RNA y modelos ARIMA" el cual tiene como objetivo "probar la viabilidad de un modelo híbrido para el precio de las virutas de madera en la previsión del mercado de Lituania", con el fin de usarlas como alternativas de fuentes de energía, analiza los resultados a partir de seis métricas, la primera el *Akaike Information Criterion* (AIC) el cual es un estimador de la calidad relativa de los modelos estadísticos para un conjunto dado de datos, el segundo es el *Bayesian Information Criterion* (BIC) que es un criterio que parte de la selección de modelos a partir de un conjunto de modelos finitos utilizando para ello el resultado con valor mínimo, el R^2 o coeficiente de determinación, Error Cuadrado Medio, Error Medio Absoluto y Error Porcentual Absoluto Medio, encontrando que, según los criterios de decisión el peor modelo para la previsión es la caminata aleatoria, teniendo como principal inconveniente el no ajustarse bien a los datos, luego se encuentra el modelo ARIMA debido a la imposibilidad de acomodarse a

datos con un comportamiento variable y aleatorio (no lineal), ya que el ARIMA es un modelo de carácter lineal; y el mejor resultado de los tres lo obtienen las redes neuronales logrando un mejor rendimiento en la evaluación de errores (siendo significativa dicha diferencia) y el R^2, sólo mostrando resultados poco satisfactorios en el AIC y el BIC, lo cual ejemplifica la complejidad de este tipo de modelo; sin embargo, el modelo híbrido creado a partir de la combinación del modelo ARIMA y las RNA en el cual se centra la investigación da como resultado el mejor desempeño con base en las medidas de errores y R^2, ya que al ser un modelo combinado aumenta tanto el AIC como el BIC, siendo el mejor de todos los modelos evaluados, superando incluso al modelo de las redes neuronales. Dado lo anterior, el autor enuncia que, aunque las redes neuronales poseen comportamientos excelentes y prometedores y los modelos ARIMA son buenos, la generación de un híbrido a partir de éstos mejora sustancialmente los resultados de la previsión de un producto sin dejar de lado las grandes características que posee cada técnica o método.

El proyecto "Predicción de series temporales financieras de alta frecuencia: enfoque de aprendizaje automático" (Zankova, 2016), tiene como propósito aplicar técnicas de aprendizaje automático para predecir series de tiempo financieras de alta frecuencia, tomando como fuente de información principal los datos financieros de alta frecuencia de la bolsa de valores de Oslo; este conjunto de datos cubre un intervalo de tres años (2006-2008) de la actividad de negociación de tres compañías diferentes: *Birdstep Technology* (BIRD), *REC Silicon* (REC) y *Statoil* (STL). Su objetivo principal es "desarrollar una solución que prediga las fluctuaciones de precios utilizando las implementaciones de regresores existentes", luego "evaluar el rendimiento de los regresores y revelar combinaciones de sus parámetros que proporcionan los mejores resultados de acuerdo con los criterios de rendimiento predeterminados", los cuales serán evaluados con respecto a la calidad de predicción y el costo de calidad, para finalmente "revelar el regresor más adecuado para el

análisis de datos de alta frecuencia". Para ello se realizaron experimentos utilizando cuatro regresores: árbol de decisión regresor, regresor de los vecinos más cercanos, regresor de perceptrón multicapa (MLP) y regresor de soporte vectorial (SVR) con ayuda del lenguaje de programación Python, elegido por su simplicidad, multifuncionalidad y aplicabilidad en todo lo relacionado con la inteligencia artificial. Los resultados arrojaron que la predicción a corto plazo (aproximadamente 1 minuto más adelante) es más favorable para los datos financieros de alta frecuencia dados, es decir, la predicción a corto plazo utilizando los regresores propuestos puede ser más precisa. Así mismo, la validación cruzada demostró que los cuatro regresores pueden alcanzar mejores resultados con parámetros ya determinados a través de una mejor elección del conjunto de entrenamiento. "Las combinaciones de parámetros que producen los mejores resultados para cada regresor se enumeran a continuación: árbol de decisión: profundidad del árbol d = 4; perceptrón multicapa: una capa oculta con 100 neuronas, función de activación logística, algoritmos de optimización de peso SGD y ADAM, velocidad de aprendizaje $\mu = 0.001$ k; Vecino más cercano: función de peso uniforme, 50 vecinos más cercanos y regresor de soporte vectorial: kernel lineal, parámetro de penalización C = 1, parámetro de tubo $\varepsilon = 0.1$". El análisis comparativo de regresores examinados ayudó a identificar el mejor regresor con un conjunto de parámetros correspondientes, el regresor MLP demostró los mejores resultados en la predicción a corto plazo con respecto a gastos de tiempo y precisión de predicción.

Los tres proyectos que se analizaron anteriormente tienen una relación directa con el proyecto en cuestión, debido a que presentan objetivos de desarrollo en común, comparan modelos para la obtención de un óptimo, hacen uso de técnicas de Aprendizaje Automático y un análisis de datos de precios agrícolas en series temporales con su respectiva metodología, lo cual sirve de base para el desarrollo de las etapas del proyecto y la finalización del mismo.

5. Marco Teórico

5.1 Pronóstico

"Un pronóstico busca predecir el futuro lo más preciso que sea posible, otorgando toda la información que se encuentre disponible, incluyendo datos históricos y conocimientos sobre cualquier evento futuro que pueda afectar la previsión" (R. J. Hyndman & Athanasopoulos, 2014). Un pronóstico se centra en la reducción de la incertidumbre que se tiene respecto a un tema en específico teniendo como base cualquier tipo de información que sea relevante, recurriendo al pronóstico en cualquier escenario donde sea necesario su uso, por ejemplo, en situaciones de tipo industrial y que impliquen la inversión de una gran suma de capital humano, monetario, tecnológico, etc., radicando su importancia en la confianza con la cual se soporta el sujeto para la toma de decisiones.

5.2 Métodos de Predicción

Los métodos de predicción son técnicas para realizar pronósticos de las cuales se sirve la ciencia, la tecnología, la industria y diversas ramas académicas en la investigación y previsión de un hecho en particular. González Casimiro (2009) enuncia en su trabajo que los métodos de predicción se pueden dividir en dos grupos principales: métodos de predicción cualitativos tecnológicos o subjetivos y métodos cuantitativos; utilizando estos últimos cuando se cuentan con datos históricos, y los primeros cuando existen datos escasos o no disponibles.

5.2.1 Métodos de Predicción Cualitativos. Son métodos que se basan en el criterio de sujetos expertos o con determinado tipo de conocimiento sobre los temas de estudio, otorgando una opinión sobre el tema que acontece; este tipo de métodos se utilizan según González Casimiro (2009) en casos donde el pasado no proporciona información directa del fenómeno estudiado, por lo que también son denominados métodos sin historia; un ejemplo, es el caso de la aparición de nuevos productos en un mercado determinado, ya que no se puede acudir a información de ventas del pasado, pues no existían todavía. De acuerdo con González Casimiro, estos métodos también pueden aplicarse a estudios o investigaciones sobre posibles cambios en los patrones sociales históricos.

5.2.2 Métodos de Predicción Cuantitativos. En los métodos o técnicas de predicción cuantitativas "se parte del supuesto de que se dispone de información sobre el pasado del fenómeno que se quiere estudiar. Generalmente la información sobre el pasado aparece en forma de series temporales" (González Casimiro, 2009) o datos de entrada bien definidos. González Casimiro afirma que se debe analizar los datos del pasado y enfocar los resultados de predicción en este análisis utilizando la información, en primer lugar, para analizar e identificar el patrón a usar para describirlos, luego extrapolar este patrón de comportamiento al futuro para analizar la predicción. Esta suposición se basa en que el comportamiento de los resultados pasados continuara presentándose en el futuro.

5.2.2.1 Métodos Causales. El método basado en análisis causales o "explicativos tratan de identificar las relaciones que produjeron (o causaron) los resultados observados en el pasado. Después, se puede predecir aplicando estas relaciones al futuro" (González Casimiro, 2009), es decir, se hace uso de distintas variables que explican el comportamiento de una variable regresora. En este método se encuentra la regresión multivariable, modelos econométricos, técnicas de aprendizaje automático (*machine learning*), etc.

5.2.2.2 Series Temporales. En el método basado en series temporales, el cual "trata de hacer predicciones de valores futuros de una o varias variables, utilizando como información únicamente la contenida en los valores pasados de la serie temporal que mide la evolución de las variables objeto de estudio" (González Casimiro, 2009) se pueden encontrar el promedio simple, los promedios móviles, los promedios móviles dobles, la suavización exponencial, los modelos autoregresivos, las técnicas de aprendizaje automático (*machine learning*), entre otros.

La introducción del *machine learning* ha planteado nuevas formas de estudio para la predicción de los precios de los productos agrícolas; estas son abordadas para el desarrollo del proyecto y son explicadas con mayor profundidad en el siguiente apartado.

5.3 Máquinas de Aprendizaje Automático (Machine Learning)

El aprendizaje automático o *machine learning* es una disciplina que por medio de patrones en un conjunto de datos, aprende y define comportamientos obteniendo resultados, es decir, *machine learning* según González Casimiro (2009) es una disciplina de la inteligencia artificial que crea sistemas que aprenden automáticamente identificando patrones en millones de datos para obtener un resultado objetivo. Este aprendizaje lo hace un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros, además de mejorarse autónomamente. En la Figura 4, se pueden

observar los principales métodos dentro de cada una de las clasificaciones del aprendizaje automático.

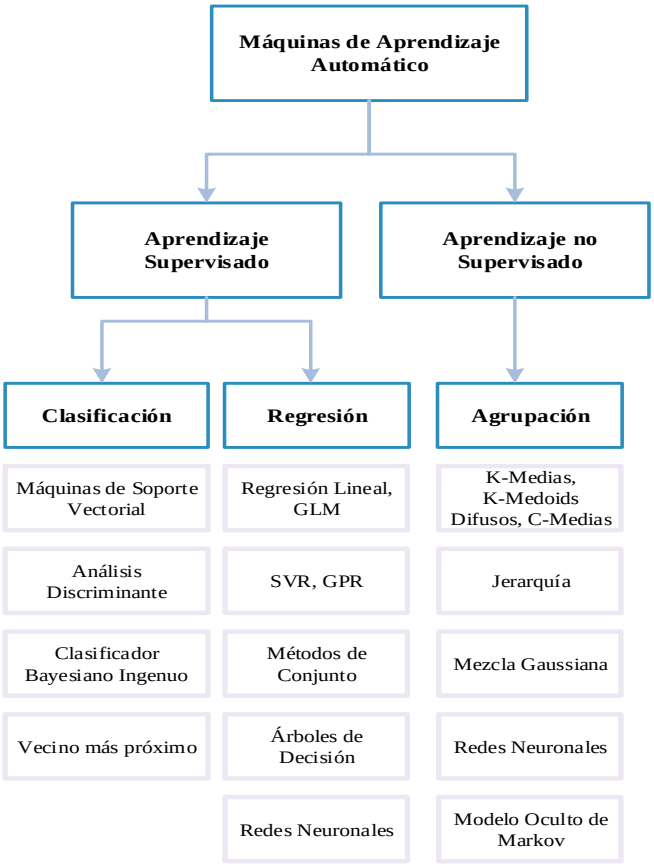

Figura 4: Métodos de aprendizaje automático. Adaptado de Math Works (2016). https://www.mathworks.com/tagteam/89703_92991v00_machine_learning_section1_ebook_v12.pdf

Las técnicas de aprendizaje automático pueden dividirse en aprendizaje automático supervisado o no supervisado; siendo el aprendizaje automático supervisado el proceso de aprendizaje de un conjunto de reglas a partir de unas instancias (ejemplos en un conjunto de entrenamiento), o más en general, para crear una herramienta que pueda ser utilizada para generalizar datos a partir de nuevas instancias (Kotsiantis, Zaharakis, & Pintelas, 2007).

El algoritmo propuesto por Kotsiantis et al (2007) en la Figura 5 presenta la aplicación del aprendizaje automático supervisado a un problema clasificación del mundo real el cual puede ser extrapolado a un problema de regresión:

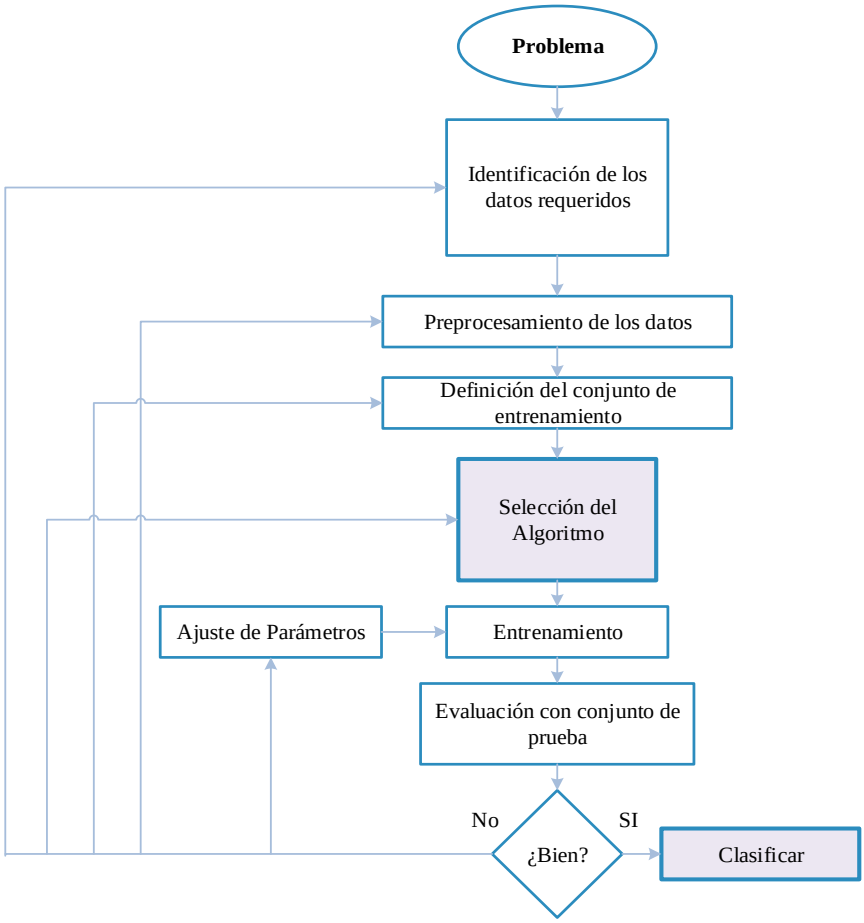

Figura 5: El proceso del aprendizaje automático. Adaptado de Kotsiantis, Zaharakis, & Pintelas (2007). https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

En la anterior figura se define la secuencia de cada paso con el fin de entender un suceso de la vida real para luego ser procesado por medio de la técnica dividendo el gráfico en dos partes; la primera parte es la recolección de los datos de entrada y la segunda es la preparación de los datos y el preprocesamiento de los mismos; donde dependiendo de la situación los investigadores poseen métodos que pueden utilizar para manejar los datos faltantes.

Dentro de los métodos de aprendizaje automático supervisado se puede encontrar los métodos de clasificación, que consisten en que la variable explicativa o de salida es una categoría como un tipo de color, enfermedad (etiquetas); por otro lado, los métodos de regresión entregan por medio del procesamiento de unas entradas, unos resultados de tipo numérico o de valor; encontrándose dentro de este ultimo la técnica a utilizar en el proyecto (RNA).

5.4 Conceptos generales de las Redes Neuronales Artificiales (RNA)

El neurofisiólogo Warren McCulloch, y el matemático Walter Pitts fueron los primeros en 1943 en lanzar una teoría acerca de la forma de trabajar de las neuronas, modelando una red neuronal simple mediante circuitos eléctricos (Damián, 2001). Las redes neuronales son una aproximación del comportamiento que sucede con el cerebro humano ofreciendo un enfoque fundamentalmente diferente, es decir, según Yasrebi & Emami (2008), las RNA son una simplificación del cerebro humano, la cual se compone de unidades de procesamiento individuales denominadas neuronas; esta es capaz de aprender y generalizar a partir de datos experimentales incluso si poseen ruido o son imperfectos, esta habilidad le permite a este sistema computacional aprender relaciones constitutivas a partir del resultado de la experimentación con datos. Este modelo se diferencia de los convencionales al no depender de conocimiento previo, o constantes y/o suposiciones para sus resultados.

Las redes neuronales pueden clasificarse según su arquitectura y su aprendizaje. La arquitectura de una red incluye la determinación de sus diferentes componentes como el número de capas y neuronas, las interconexiones entre ellos, los procedimientos y los parámetros de predicción; aunque toda la arquitectura se determine mediante la experimentación de los datos dados, ya que no existe una base teórica para determinar estos parámetros, es posible tener un conocimiento de

los diferentes tipos de red que existen. Según Galán & Martínez (2006), las redes neuronales se clasifican de acuerdo a su arquitectura de la siguiente manera:

5.4.1 Redes Monocapa. Cuentan con solo una capa de neuronas, las cuales intercambian señales con el exterior constituyendo la entrada y la salida del sistema a un solo tiempo; son usualmente usadas en circuitos eléctricos. Una de las redes más representativas de este modelo es la red de Hopfield.

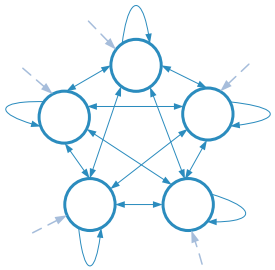

Figura 6: Arquitectura de una red de Hopfield. Adaptado de Galán & Martínez (2006) http://www.it.uc3m.es/jvillena/irc/practicas/10-11/06mem.pdf

5.4.2 Redes Multicapa. Están constituidas por más de dos capas neuronales conectadas entre sí; son una de las redes neuronales más utilizadas para la predicción de series de tiempo no lineales (J. D. Velásquez, Zambrano, & Vélez Laura, 2011), lo que se debe según Llano, Hoyos, Arias, & Velásquez (2007) a su simplicidad y buenos resultados. Una típica red neuronal multicapa de acuerdo a Sánchez Anzola (2016), consiste en una serie de elementos de procesamiento (PE) o nodos que se encuentran dispuestos en una capa de entrada (es la capa que recibe toda la información directamente de fuentes externas), una capa de salida (son las neuronas que transfieren la información que la red proceso al exterior) y otras capas ocultas (son internas y sin contacto al exterior, pueden ser más de una, y las neuronas pueden interconectarse entre ellas mismas).

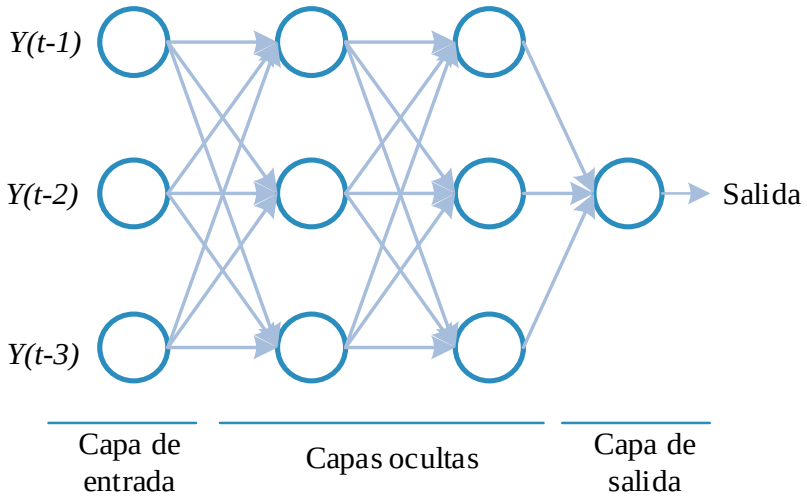

Figura 7: Modelo de red neuronal múlticapa. Adaptado de Sánchez Anzola (2016) http://revistas.uexternado.edu.co/index.php/odeon/article/view/4414/5256

Así mismo, atendiendo al flujo de datos en la red neuronal es posible hacer otra subdivisión; de acuerdo con Galán & Martínez (2006) existen dos clasificaciones:

5.4.3 Redes con conexiones hacia delante o *Feed forward*. Este tipo de redes contienen solo conexiones entre capas hacia delante, es decir, la información circula en un único sentido, desde las neuronas de entrada hacia las de salida. Lo cual implica que una capa no puede conectarse a una que reciba la señal antes que ella en la dinámica de la computación.

5.4.4 Redes con conexiones hacia atrás o *Feed back*. En este tipo de redes la información puede circular entre las capas en cualquier sentido, de esta manera pueden existir conexiones de capas hacia atrás y por tanto la información puede regresar a capas anteriores en la dinámica de la red.

Además del tipo de arquitectura de la red, otro concepto que caracteriza un modelo neuronal es el tipo aprendizaje, el cual es de suma importancia para el entrenamiento de la red; este proceso consiste en que la red pueda procesar los patrones de forma iterativa hasta encontrar respuestas

satisfactorias. Galán & Martínez (2006) las categorizan de la siguiente forma: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje híbrido (no se proporcionan los patrones objetivo, sino que solo se dice si la respuesta acierta o falla ante un patrón de entrada).

Actualmente, las redes neuronales debido a sus propiedades y características se destacan por encima de las técnicas estadísticas. De acuerdo con Pitarque et al. (2000), las redes neuronales han mostrado su superioridad respecto a otras técnicas estadísticas con su capacidad clasificatoria, ya que pueden ser utilizadas independientemente del cumplimiento de supuestos teóricos relativos a estas técnicas, de esta manera se habla de ellas como técnicas de distribución libre o no paramétricas. Todo esto ha llamado la atención de estadísticos, ya que exhiben un buen rendimiento para tratar problemas no lineales y con mucho ruido, lo que ha llevado a su uso en problemas de clasificación y predicción.

En la actualidad, existen diversas aplicaciones de las RNA al pronóstico, su naturaleza no paramétrica impulsada por datos y propiedades autoadaptativas, además de las propiedades dichas anteriormente, han hecho que sea un método más fácilmente accesible en comparación con otras alternativas no lineales basadas en modelos. La aplicación potencial de las redes neuronales se puede encontrar en campos como biología, ingeniería, economía, etc. (Anjoy & Kumar, 2017).

Para el uso y desarrollo de la técnica RNA es necesario conocer los pasos básicos para construir el modelo. El primer paso es la recolección de datos, y algo a lo que se enfrentan la mayoría de investigadores y tomadores de decisiones en esta etapa es a la presencia de datos faltantes; sin embargo, la aplicación de métodos de imputación puede resolver este problema, pero si no hace de manera correcta pueden producir resultados inadecuados en etapas posteriores. Debido a esto, en el análisis de los datos se han diseñados diversas formas de imputación con el fin de reparar la

ausencia de los mismos y se garantice un estudio más preciso; estas formas van desde los métodos tradicionales conocidos como la eliminación de datos *(listwise)*, la cual es una metodología de limpieza que implica eliminar una categoría si la cantidad de datos faltantes es significativa, el pareo de observaciones o eliminación por pares (pairwise), sustitución por el promedio y métodos de máxima verosimilitud (MV), hasta algoritmos de simulación como la imputación múltiple (IM), la cual sustenta que cada dato faltante debe ser reemplazado a partir de m>1 simulaciones del proceso (Medina & Galván, 2007a); sin embargo, los primeros carecen de veracidad al eliminar datos que pueden ser importantes y el último modelo es utilizado para encontrar los datos faltantes de una variable a partir de la existencia de otras diferentes, es decir, para procesos multivariantes y no para univariantes, lo que dificulta la imputación de los mismos; no obstante, se pueden aplicar métodos basados en series temporales multivariantes y extrapolar el proceso a una serie univariante.

Antes de usar los datos necesarios para el estudio de predicción que se va a ejecutar se hace un preprocesamiento de los datos de entrada con el fin de limpiarnos y disminuir la cantidad de ruido presente en ellos. Según García, Ramirez, Luengo, & Herrera (2016) el preprocesamiento de datos es esencial para el descubrimiento de la información o KDD (*Knowledge Discovery In Databases*). Esta parte del proceso limpia, integra, transforma y reduce los datos, para mejorar su calidad, facilitar los siguientes pasos y obtener mejores resultados. Otra parte importante en el desarrollo de las redes neuronales para obtener una buena solucion, es elegir adecuadamente el tipo de arquitectura de la red y el algoritmo de aprendizaje, pues de eso depende el desempeño del entrenamiento de la red. A continuación, se expondrán la teoria de los tres modelos que seran desarrollados en el proyecto como se menciona en el análisis preliminar de literatura (modelo híbrido ARIMA-RNA *feedforward*, modelo RNA *bp* y modelo ANFIS).

5.5 Modelo Hibrido ARIMA-RNA *feed forward*

La metodología ARIMA es una de las más utilizadas por su sencillez y su capacidad de representar procesos estacionarios y estocásticos; pero presenta una desventaja, ya que asume como una estructura lineal el proceso de series temporales cuando esto no se cumple en todos los casos. Debido a lo anterior, se plantea su hibridación con el modelo de RNA *ff*, dado que este permite suplir el problema de no linealidad del modelo ARIMA.

La Figura 8, presenta de forma general el proceso de la hibridación del modelo, del cual se hablará detalladamente en el espacio dedicado al desarrollo de este.

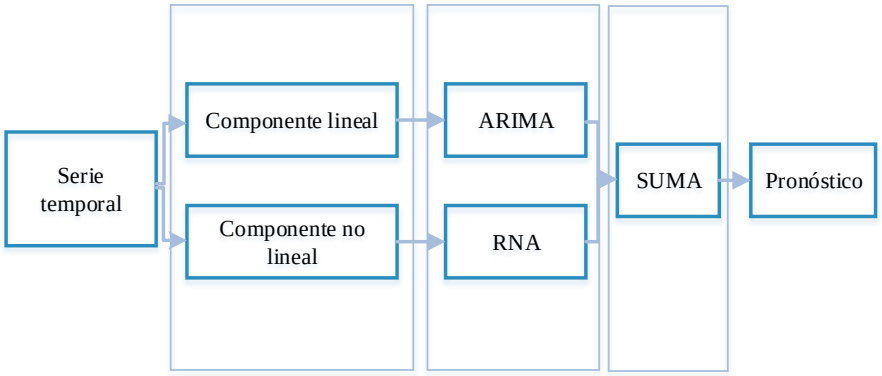

Figura 8: Metodología modelo ARIMA-RNA *ff*. Adaptado de Mitra & Paul, (2017) https://content.iospress.com/articles/model-assisted-statistics-and-applications/mas400

Para entender el modelo de hibridación también es necesario definir el modelo ARIMA. El modelo ARIMA es un modelo univariante (entendiendo univariante como el análisis de solo una característica o una única cualidad). De Arce & Mahía (2003, pp. 1–3) definen un modelo como autorregresivo si el periodo t de una variable endógena se explica así misma de acuerdo a períodos anteriores añadiendo un término de error. Como su nombre lo indica, este modelo es una combinación de AR (1) (modelo autoregresivo), I (integrado) y MA (modelos de media móvil) (2); la parte AR del modelo explica el comportamiento de una variable en función de los valores que

tomó en el pasado (rezagos) y la parte MA explica ese comportamiento a través de los errores al estimar el valor de la variable en los períodos anteriores. Ambas partes indican la estructura y el orden de la ecuación de regresión con los valores p (número de rezagos entre la variable y el valor de ella misma en el pasado) y q (número de rezagos entre el error de la variable y el error de ella misma en el pasado) respectivamente; por otro lado, el valor de integración *I* representa el número d de veces necesarias para diferenciar y de esta forma convertir la serie en estacionaria; formando de esa manera un ARIMA de orden p, d, q o ARIMA (p, d, q) (4) .

Siendo la expresión AR:

$$y_t = \varphi_0 + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} +_{\dots} + \varphi_p y_{t-p} + a_t \quad (1)$$

La expresión MA:

$$y_t = \mu + a_t + \theta_1 a_{t-1} +_{\dots} + \theta_q a_{t-q} \quad (2)$$

La expresión ARMA:

$$y_t = \mu + \varphi_1 y_{t-1} +_{\dots} + \varphi_p y_{t-p} + a_t + \theta_1 a_{t-1} +_{\dots} + \theta_q a_{t-q} \quad (3)$$

Y, un modelo ARMA integrado:

$$\Delta^d y_t = \mu + \varphi_1 \Delta^d y_{t-1} +_{\dots} + \varphi_p \Delta^d y_{t-p} + a_t + \theta_1 a_{t-1} +_{\dots} + \theta_q a_{t-q} \quad (4)$$

Para entender los modelos ARIMA a continuación se mencionarán algunos conceptos claves para su uso.

5.5.1 Proceso estocástico y estacionario. "Un proceso estocástico, es una sucesión de variables aleatorias $\boldsymbol{Y_t}$ ordenadas, pudiendo tomar t cualquier valor entre $\boldsymbol{Y_{-t}}$ y $\boldsymbol{Y_t}$. […]"(de Arce & Mahía, 2003, pp. 1–3); este tipo de proceso es considerado estacionario si con respecto al desplazamiento en el tiempo las funciones de distribución conjuntas son invariantes (variación de t).

5.5.2 Ruido Blanco. "Es una sucesión de variables aleatorias (proceso estocástico) con esperanza (medía) cero, varianza constante igual a 1 e independientes para distintos valores de $\boldsymbol{t}$ (covarianza/correlación nula)" (de Arce & Mahía, 2003, p. 2).

5.5.3 Autocorrelación. Es una característica que indica relación lineal entre los errores de una sucesión de variables aleatorias $\boldsymbol{Y_t}$ ordenadas, por ejemplo, en ocasiones en series temporales algunos valores dependen en el periodo de estudio de los valores anteriores en la línea de tiempo; a continuación, se mencionan dos formas de medir dicha dependencia.

5.5.3.1 Autocorrelación simple. De acuerdo con Villavicencio (2011), la autocorrelación mide la correlación entre dos variables separadas por $\boldsymbol{k}$ periodos (5)**.**

$$\rho_j = corr(X_j X_{j-k}) = \frac{cov(X_j X_{j-k})}{\sqrt{V(X_j)}\sqrt{V(X_{j-k})}} \quad (5)$$

Además, Villavicencio muestra que la función de autocorrelación simple tiene algunas características:

- $\rho_0 = 1$
- $-1 \leq \rho_j \leq 1$
- Simetría $\rho_j = \rho_{-j}$

5.5.3.2 Autocorrelación parcial. También se encarga de medir la correlación entre dos variables separada por ***k*** periodos, sin embargo este tipo de autocorrelación no tiene en cuenta el posible efecto de los residuos intermedios entre los periodos observados (Villavicencio, 2011).

$$\pi_j = corr\left(X_j X_{j-k} / X_{j-1} X_{j-2} \dots X_{j-k+1}\right) \tag{6}$$

$$\pi_j = \frac{cov(X_j - X, X_{j-k} - X_{j-k})}{\sqrt{V(X_j - \hat{X}_j)}\sqrt{V(X_{j-k} - \hat{X}_{j-k})}} \tag{7}$$

En relación a los estudios necesarios para el análisis y procesamiento de los datos a través del modelo ARIMA, se presentan las siguientes pruebas que permiten obtener una percepción de su comportamiento.

5.5.4 Ljung-Box. "Esta prueba permite probar en forma conjunta que todos los coeficientes de autocorrelación son simultáneamente iguales a cero, es decir que son independientes" (Villavicencio, 2011) o aleatorios; la prueba está definida como:

$$LB = n(n+2)\sum_{k=1}^{m}\left(\frac{\hat{\rho}_k^2}{n-k}\right) \sim X_{(m)}^2 \tag{8}$$

Donde n es el tamaño de la muestra y m la longitud del rezago.

Ho: *La autocorrelación es independiente*

Ha: *La autocorrelación es no independiente*

En una aplicación si el Q calculado excede el valor Q crítico de la tabla ji cuadrada al nivel de significancia seleccionado, no se acepta la hipótesis nula de que todos los coeficientes de autocorrelación son iguales a cero; por lo menos algunos de ellos deben ser diferentes de cero.

5.5.5 Prueba Durbin-Watson. Según Gujarati & Porter (2010) la prueba Durbin-Watson permite comprobar si existe correlación serial entre los errores de los rezagos de la serie temporal, obteniendo el estadístico d de Durbin Watson a partir de la siguiente ecuación:

$$d = \frac{\sum_{t=2}^{t=n}(u_t - u_{t-1})^2}{\sum_{t=1}^{t=n} u_t^2} \quad (9)$$

Donde d es aproximadamente igual a:

$$d \approx 2(1 - \rho) \quad (10)$$

Y se encuentra definido entre $0 < d < 4$; siendo ρ el coeficiente de autocorrelación muestral de primer orden el cual se encuentra en $-1 < \rho < 1$. La prueba contrasta la hipótesis nula (Ho) de inexistencia de autocorrelación contra la hipótesis alternativa (Ha) de existencia de autocorrelación de primer orden, obteniendo los siguientes resultados con los valores de d:

Si d = 2 (ρ = 0) no existe correlación serial en los residuos

Si d = 0 (ρ = 1) existe correlación serial positiva perfecta

Si d = 4 (ρ = -1) existe correlación serial negativa perfecta

Sin embargo, la prueba solo puede ser aceptada cuando el estadístico no cae en la zona de indecisión. En la Tabla 2 se presentan los escenarios para los cuales la prueba puede ser o no aceptada:

Tabla 2
Niveles de aceptación o rechazo de la prueba d

Hipótesis nula	Decisión	Si
No hay autocorrelación positiva	Rechazar	$0 < d < dL$
No hay autocorrelación positiva	Sin decisión	$dL \leq d \leq dU$
No hay correlación negativa	Rechazar	$4\text{-}dL < d < 4$
No hay correlación negativa	Sin decisión	$4\text{-}dU \leq d \leq 4\text{-}dL$
No hay autocorrelación, positiva o negativa	No rechazar	$dU < d < 4\text{-}dU$

dL y dU: valores críticos inferior y superior para comparar con d
Nota. Recuperado de Gujarati, D. N., & Porter, D. C. (2010). Econometria. McGraw-Hill. p 436.

5.5.6 Prueba de raíz unitaria. La prueba estadística de raíz unitaria evalúa la hipótesis nula de aleatoriedad, la cual consiste en determinar si un proceso presenta un comportamiento como el que se observa en las siguientes ecuaciones donde $\boldsymbol{\delta}$ se determina a partir de:

$$\delta = (\rho - 1) \tag{11}$$

5.5.6.1 Caminata aleatoria.

$$\Delta Y_t = \delta Y_{t-1} + u_t \tag{12}$$

Donde Y_t es la variable predictora, Y_{t-1} es el valor rezagado de la serie, u_t es el error de ruido blanco, δ la raíz del proceso y ρ el valor de autocorrelación entero.

5.5.6.2 Caminata aleatoria con deriva.

$$\Delta Y_t = \beta_1 + \delta Y_{t-1} + u_t \tag{13}$$

Donde β_1 es el intercepto o deriva del proceso.

5.5.6.3 Caminata aleatoria con deriva y tendencia determinística.

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + u_t \tag{14}$$

Donde *t* es el tiempo del proceso y β_2 es la tendencia en el tiempo.

Según Gujarati & Porter (2010), la hipótesis nula plantea que si ρ es igual a 1, la serie presenta un proceso de raíz unitaria, de lo contrario se define al proceso como estacionario.

$$Ho: \delta = 0$$

$$Ha: \delta < 1$$

Si la hipótesis nula es rechazada significa que Yt es estacionaria con media igual a cero (primera ecuación), Yt es estacionaria con media diferente de cero (segunda ecuación) o Yt es estacionaria con media diferente de cero y existencia de tendencia (tercera ecuación).

5.5.7 Prueba de Bartels. Es una prueba de rango para la aleatoriedad de los datos; esta prueba es la versión de rango de la prueba de *randomness* de Von Neumann (RVN) (15), la cual consiste en una transformación lineal del coeficiente de correlación serial de rango y sólo se puede utilizar con residuos independientes y homoscedasticos. El coeficiente de correlación serial de rango fue introducido en el año 1943 por Wald y Wolfowitz, quienes demostraron su normalidad asintótica (Bartels, 1982).

La estadística de prueba RVN es:

$$RVN = \frac{\sum_{i=1}^{N-1}(R_i - R_{i+1})^2}{\sum_{i=1}^{n}(R_i - \frac{n+1}{2})^2} \quad (15)$$

Donde $Ri = rango\ (Xi)$, con i = 1, …n. Se debe saber que $(RVN - 2) / \sigma$ es asintóticamente estándar normal, donde $\sigma 2 = [4(n-2)\ (5n2 - 2n - 9)]/[5n(n+1)\ (n-1)2]$.

La hipótesis de la prueba plantea:

Ho: *Aleatoriedad*

Ha: *No aleatoriedad*

5.6 Modelo de Red Neuronal *backpropagation*

El modelo de RNA de propagación hacia atrás consiste en utilizar para el entrenamiento de la red neuronal el algoritmo de aprendizaje *bp*. Según Llano et al. (2007), este algoritmo fue creado como generalización del algoritmo para perceptrón continuo cuando se tienen múltiples capas o (redes neuronales perceptrón multicapa) y es popularmente utilizado para la solución de problemas de clasificación y pronóstico.

El nombre de *backpropagation* es producto de la forma en que el error es propagado hacia atrás a través de la red neuronal desde la capa de salida. Según Damián (2001), "esto permite que los pesos sobre las conexiones de las neuronas ubicadas en las capas ocultas cambien durante el entrenamiento"; este cambio en el peso de las conexiones influye en la entrada global, en la activación y consecuentemente en la salida de la neurona, lo que lleva a tener en cuenta las variaciones de la función de activación. Damián llama a esto sensibilidad de la función de activación. En la Figura 9 se muestra el modelo de la Red Neuronal con retropropagación.

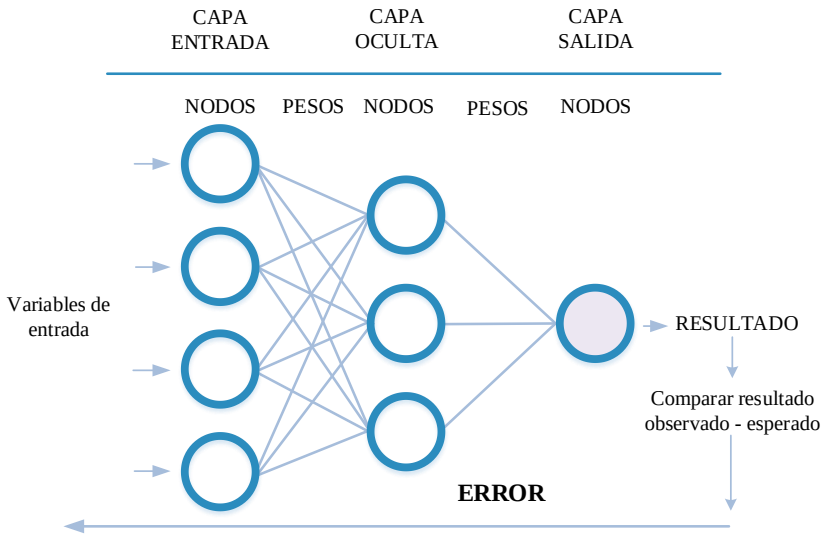

Figura 9: Modelo de Red Neuronal Artificial con retropropagación. Adaptado de Trujillano Cabello et al., (2005) http://dx.doi.org/10.1016/S0210-5691(05)74198-X

El funcionamiento de la RNA *bp* consiste en el aprendizaje de un conjunto predefinido de pares de entradas y salidas. Valencia, Yáñez, & Sánchez (2006, p. 5) explican que primero se hace un estímulo a la primera capa mediante un patrón de entrada y este se va propagando a través de las capas superiores hasta generar una salida, para luego comparar los resultados en las neuronas de salida con los resultados deseados, donde se calcula un error para cada neurona de salida; después dichos errores se propagan hacia atrás desde la capa de salida hacia las neuronas intermedias que estén directamente relacionas con ella en la proporción que estas aportaron a esta salida. Por otra parte, para el desarrollo del modelo de RNA con *bp* se requiere definir algunos parámetros fundamentales para su ejecución, los cuales se mencionarán a continuación:

5.6.1 Validación cruzada. Es una herramienta para dividir el conjunto de datos en $\boldsymbol{k}$ subconjuntos, con el objetivo de validar los resultados generados. Uno de los subconjuntos se utiliza como datos de validación para estimar el error de generalización y el resto $(\boldsymbol{k}-\mathbf{1})$ como datos de entrenamiento para determinar los parámetros de la red neuronal (J. Velásquez, Fonnegra, & Villa, 2013). De acuerdo con Velásquez, Fonnegra, & Villa, "El proceso de validación cruzada es repetido durante $\boldsymbol{k}$ iteraciones, con cada uno de los posibles subconjuntos de datos de validación. Finalmente, se selecciona el que mayor capacidad de generalización posea", controlando de esta manera el problema de sobreajuste. Es importante resaltar que el conjunto de validación no contenga datos usados en el entrenamiento de la red; en el caso de una red neuronal regresora se utiliza el 80% de los datos introducidos de forma cronológica para el entrenamiento y el 20% restante (o últimos) para la prueba o pronostico (*leave all out*).

5.6.2 Número de capas ocultas. Las capas ocultas son las capas de la red que no poseen conexión directa con el entorno, éstas permiten que la red neuronal pueda representar de mejor manera determinadas características del entorno que trata de modelar, mediante la proporción de grados de libertad a la red (Larrañaga, Inza, & Moujahid, 1997). El número de capas deben ser mayores si el problema es no lineal y complejo, pero en general un problema podrá representarse bastante bien con una o dos capas ocultas (Llano et al., 2007). El número de neuronas por capa normalmente se determina por ensayo y error, aunque existe el criterio de considerar al promedio entre el número de entradas y salidas para determinar número de neuronas en estas capas.

5.6.3 Pesos o sinapsis. Siendo el conjunto de entradas $\boldsymbol{X_j(t)}$, los pesos sinápticos representan la intensidad de interacción entre cada neurona presináptica j y la neurona postsinaptica i a través de una magnitud w_{ij} (Martín del Brio & Sanz Molina, 2007), con lo que se determina si se excita o inhibe la neurona. Este valor del peso afecta las entradas, pues de este valor depende la importancia de una entrada con respecto a la otra (Torres Soler, 2010).

5.6.4 Regla de propagación. Proporciona el valor del potencial postsinaptico $\boldsymbol{h_i(t) = \sigma\left(w_{ij}, x_j(\mathbf{t})\right) = \sum w_{ij}\, x_j}$ de una neurona $\boldsymbol{i}$ en función de sus pesos y entradas, mediante la suma de las entradas ponderadas con sus pesos correspondientes; esta es la regla de propagación más simple y utilizada (Martín del Brio & Sanz Molina, 2007). Otra regla de propagación comúnmente usada es la distancia euclidea, la cual es el tipo de regla que tienen redes como los mapas auto-organizativos SOM (*Self Organizing Map*) o las redes de función de base radial RBF.

5.6.5 Función de activación. Representa simultáneamente la salida de la neurona y su estado de activación $\boldsymbol{y_i(t) = f_i(h_i(t))}$. En relación a los tipos de funciones de activación, algunos algoritmos de aprendizaje requieren que la función cumpla la condición de ser derivable como es el caso del *bp.* Las más empleadas en este sentido son las funciones de tipo *sigmoide*, una función continua y diferenciable en cierto intervalo, por ejemplo, en el [0, +1] o en el [-1, +1], lo que depende de la función elegida, la logística sigmoide o la tangente hiperbólica respectivamente (Martín del Brio & Sanz Molina, 2007). Según Llano et al. (2007) la función logística es la más comúnmente usada por sus características de derivación y se recomienda para problemas de predicción, aunque en el algoritmo de retropropagación la función más comúnmente utilizada es la logística, este puede trabajar con cualquier otra función de activación que sea diferenciable; la función tangente hiperbólica es similar a la logística y se utiliza con frecuencia en redes multicapa.

5.6.6 Umbral. Es un parámetro $\boldsymbol{\theta i}$ que se resta del potencial postsinaptico. Este parámetro representara el umbral de disparo de la neurona, es decir, el nivel mínimo que debe alcanzar el potencial postsinaptico para que la neurona se active (Martín del Brio & Sanz Molina, 2007). Por lo que la función de activación queda de la siguiente forma:

$$\sum_j wij\, x_j - \theta i \tag{16}$$

5.7 Modelo ANFIS (*Adaptive Network- based in Fuzzy Inference Systems).* El modelo ANFIS fue desarrollado por J.R. Jang en 1993; el modelo es una red adaptativa que combina el conocimiento experto plasmado en reglas difusas con la capacidad de aprendizaje que poseen las redes neuronales. Según Mora, Pérez, & Pérez (2006), "el modelo ANFIS consta de un sistema híbrido neurodifuso, el mismo que es funcionalmente equivalente al mecanismo de inferencia Takagi-Sugeno (T-S), para un sistema de inferencia T-S de primer orden".

Las redes ANFIS por su capacidad adaptativa son aplicables a un gran número de áreas como el control adaptativo, procesamiento y filtrado de señales, series de tiempo, clasificación de datos y extracción de características a partir de ejemplos. Una propiedad interesante del modelo es que el conjunto de parámetros puede descomponerse para utilizar una regla de aprendizaje (Mora et al., 2006).

Para crear modelos ANFIS, se pueden utilizar dos metodologías que generan reglas difusas, una genera reglas de tipo Takagi-Sugeno de manera automática, la otra construye reglas de tipo Mandani mediante el algoritmo Wang – Mendel, sin embargo, este último genera un error alto en el pronóstico, por lo cual se recomienda el uso de la primera regla (Mora et al., 2006). Las redes ANFIS tienen la ventaja de construir modelos a partir de datos de entrada y salida reduciendo el tiempo de modelamiento y el conocimiento de expertos para ser aplicada (Correa & Montoya, 2013).

5.7.1 Sistema de inferencia difuso (*Fuzzy Inference Systems* – *FIS*). Los sistemas de inferencia difusa permiten tratar información que no es precisa como estatura medía, temperatura baja o mucha fuerza. De acuerdo con Sanjay Krishnankutty, Torres Blanc, & Sánchez Torrubia (2015), los FIS son aquellos que transforman mediante lógica difusa un espacio de entrada en un espacio de salida. Estos tratan de formalizar razonamientos del lenguaje humano, por ejemplo, "si el servicio es muy bueno, aunque la comida no sea excelente, le daré una buena propina" utilizando la lógica difusa, construyendo reglas borrosas; una regla típica es la de tipo si entonces *(if-then)* que en su forma más simple seria de la siguiente forma "Si $\boldsymbol{x}$ es $\boldsymbol{A}$, entonces $\boldsymbol{y}$ es $\boldsymbol{B}$".

La Figura 10 representa la estructura de un sistema de interferencia difusa donde los valores numéricos de entrada ingresan al módulo de fusificación, ahí las entradas se convierten en conjuntos difusos aplicando una función de fusificación, luego son transferidos al motor de interferencia que simula el razonamiento humano a través de las entradas y las reglas obtenidas por los expertos, las cuales se almacenan en el módulo de la base del conocimiento, para finalmente convertir el conjunto difuso en un valor numérico mediante el proceso de defusificación.

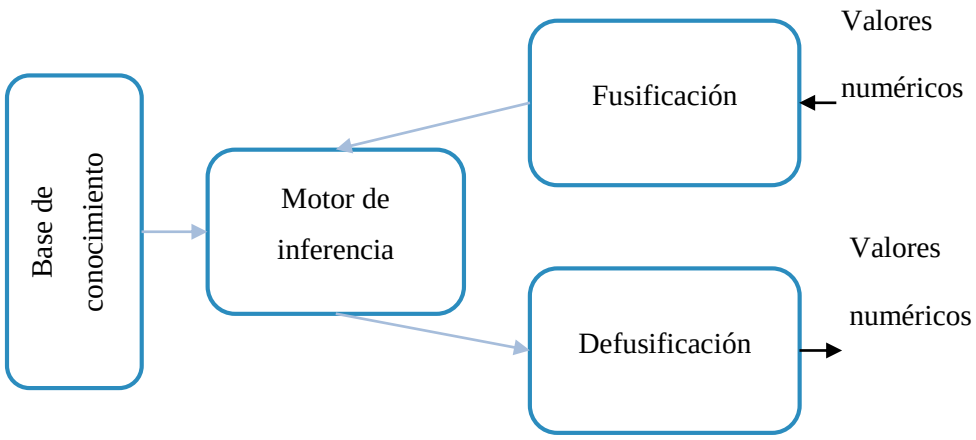

Figura 10: Estructura de un sistema de interferencia difusa. Adaptado de, Sanjay Krishnankutty et al. (2015): http://www.dma.fi.upm.es/recursos/aplicaciones/logica_borrosa/web/fuzzy_inferencia/introfis.htm

5.7.1.1 Fusificación. Consiste en convertir los valores de entrada en valores difusos para que puedan procesarse por el motor de interferencia, ya que en muchos casos las entradas son valores numéricos que proceden generalmente de sensores y son no difusos. El sistema difuso más simple sustituye el valor numérico por un conjunto difuso de tipo *singleton* o también llamado puerto difuso (Ojeda Magaña, 2010).

5.7.1.2 Base de reglas. Según Ojeda Magaña (2010), las reglas difusas pueden ser de tipo: SISO (*Single – Inputs, Single – Outputs*) que tienen solo una entrada y una salida, las MIMO (*Multiple-Inputs, Multiple-Outputs*) con varias entradas y varias salidas; y las MISO (*Multi-Input, Single-Output*), las cuales poseen varias entradas y una salida. Dentro de estas se encuentran los FLU (*fuzzy logic unit*) que tienen solamente dos variables de entrada y una de salida.

5.7.1.3 Motor de inferencia difusa. De acuerdo Ojeda Magaña (2010), este componente se encarga de que las salidas de un sistema difuso estén en función de sus entradas y sus reglas, puede dar dos tipos de salida, valores numéricos o conjuntos difusos.

5.7.1.4 Defusificación. Se encarga de convertir las salidas de tipo difuso en valores numéricos. Los defusificadores comúnmente utilizados son: el método del centroide, el de la máxima pertenencia y el método del centro de gravedad (Ojeda Magaña, 2010).

5.7.1.5 Base de conocimiento. Este elemento es el que se encarga de almacenar todas las relaciones entre las entradas y salidas del sistema, y de acuerdo a este conocimiento se obtienen las salidas asociadas con las entradas; está formado por dos partes, la base de datos y la base de reglas, la primera contiene las funciones de pertenencia que definen las etiquetas lingüísticas, y la segunda guarda la colección de reglas lingüísticas del sistema, la cual se representa como una lista de reglas, una tabla o matriz de decisión. Para obtener dichas reglas se puede basar en el conocimiento de un experto a través de cuestionarios que permitan generar las reglas o basados en aprendizajes, mediante algoritmos automáticos (Ojeda Magaña, 2010).

5.7.2 Conjuntos difusos. A diferencia de los conjuntos clásicos, donde algo está incluido completamente en él o no lo está en absoluto, asignando el valor de 1 a todos los elementos que están incluidos y 0 a aquellos que no lo están, mediante una función de inclusión o pertenencia, los conjuntos difusos permiten describir el grado de pertenencia del valor de una variable u objeto asignando no solo 0 o 1, sino también valores entre ellos, es decir, para los conjuntos difusos una persona no es solo alta o baja, sino que también podrá ser medio baja o muy alta (Martín del Brio & Sanz Molina, 2007).

5.7.3 Funciones de inclusión o pertenencia de conjuntos difusos. Estas funciones permiten representar gráficamente un conjunto difuso. De acuerdo con Martín del Brio & Sanz Molina (2007), "la función de inclusión o pertenencia de un conjunto borroso está conformada por pares ordenados $\boldsymbol{F} = \{ \left(\boldsymbol{u}, \boldsymbol{\mu_F}(\boldsymbol{u})\right) / \boldsymbol{u \epsilon U}\}$ si la variable es discreta o una función continua si no lo es." $\boldsymbol{\mu_F}(\boldsymbol{u})$ indica el grado de pertenencia del valor $\boldsymbol{u}$ en la variable $\boldsymbol{U}$ de un conjunto $\boldsymbol{F}$.

5.7.3.1 Función de tipo campana.

$$\mu_F(u) = \frac{1}{1 + \left(\frac{u - m}{a}\right)^{2b}} \tag{17}$$

La función de tipo campana generalizada se caracteriza mediante tres parámetros, a, b y c, donde c y a definen el centro y ancho de la función respectivamente y el parámetro b controla las pendientes en los puntos de cruce.

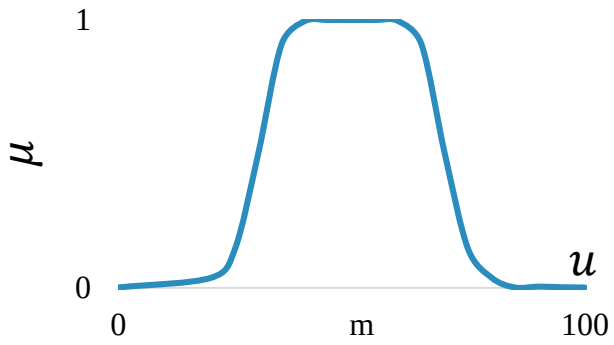

Figura 11: Función de pertenencia tipo Campana. Adaptado de Ojeda Magaña (2010): http://oa.upm.es/4838/

5.7.3.2 Función de pertenencia tipo Gaussiana.

$$A = e^{-k\,(u-m)^2} \tag{18}$$

Está definida por un valor medio m y una desviación estándar $k > 0$, con forma de campana, la cual se hace más estrecha entre menor sea k.

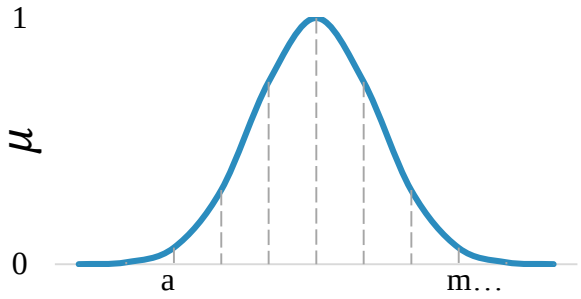

Figura 12: Función de pertenencia tipo Gaussiana. Adaptado de Andrade Revelo (2006): http://catarina.udlap.mx/u_dl_a/tales/documentos/meie/revelo_a_s/capitulo4.pdf

5.7.4 Tamaño del paso. Hace referencia al error que comete la red en función del conjunto de pesos sinápticos utilizados. El objetivo del aprendizaje es definir una configuración de pesos que corresponda al mínimo global de la función de error, aunque en muchos casos es suficiente determinar el mínimo local que sea bueno (Bertona, 2005).

5.7.5 Intersección difusa o T norma. "Es una función que toma como su argumento el par que consiste en grado de pertenencia en los conjuntos A y B, y retorna el grado de pertenencia de sus elementos que constituyen la intersección de A y B" (D. Reina, 2008). La función retorna el grado en el que el dato de entrada pertenece a la intersección de los conjuntos difusos.

5.7.6 Unión difusa o S norma. "El argumento a esta función es el par que consiste en el grado de pertenencia de algunos elementos x en el conjunto difuso A y el grado de pertenencia de ese mismo elemento en el conjunto difuso B. La función devuelve el grado de pertenencia para los elementos en el conjunto difuso $\boldsymbol{A} \cup \boldsymbol{B}$" (D. Reina, 2008). La función retorna el grado en el que el dato de entrada pertenece a la unión de los conjuntos difusos.

5.7.7 Función de implicación. Es la función que permite generar las reglas difusas a partir del cumplimiento de unas determinadas proposiciones. Es decir "si $\boldsymbol{A}$ es $\boldsymbol{a}$ y $\boldsymbol{B}$ es $\boldsymbol{b}$" (antecedente de la regla) y "entonces $\boldsymbol{f}$ es" (consecuente de la regla); por contraparte, existe el operador negación que invierte el sentido de la proposición.

5.7.8 Métodos de inferencia difusa. Los métodos de interferencia difusa se clasifican en métodos indirectos y métodos directos, estos últimos a su vez se dividen en 3 métodos diferentes, el método directo de Mamdani, el método simplificado y el Modelado de Takagi-Sugeno, siendo este último uno de los más utilizados y el cual se utiliza para el proyecto.

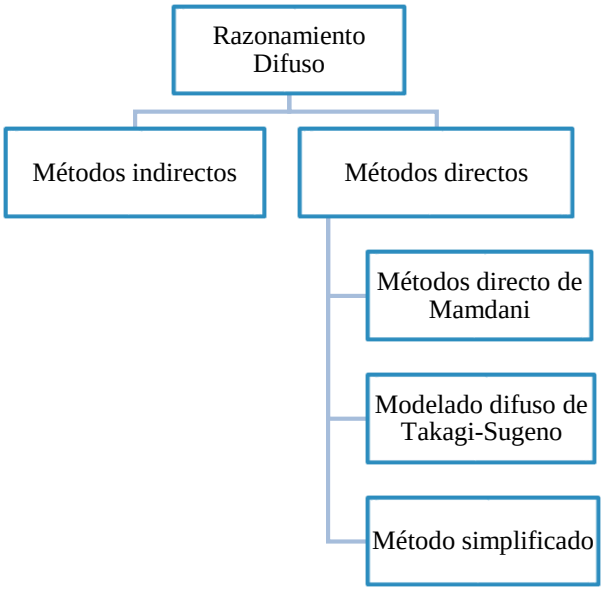

Figura 13: Métodos de interferencia difusa. Adaptado de Sanjay Krishnankutty et al. (2015): http://www.dma.fi.upm.es/recursos/aplicaciones/logica_borrosa/web/fuzzy_inferencia/introfis.htm

5.7.9 Modelado Takagi-Sugeno. Es una metodología que genera reglas difusas del tipo Takagi-Sugeno, este modelo en lugar de trabajar con variables lingüísticas usa reglas cuyas partes consecuentes están representadas por una función de las variables de entrada. Según Ojeda Magaña (2010), una regla típica tiene la siguiente forma:

$$R_1: Si\ x_1\ es\ A_1^j\ y \ldots y\ x_n\ es\ A_n^j \text{ entonces}$$

$$y = f_j(x_1, x_2, \ldots, x_n)\ j = 1, 2 \ldots M$$

Siendo $x_i, i = 1 \dots n,$ las entradas del sistema, e y es la salida inferida por la regla j, los elementos A_n^j son los conjuntos difusos del antecedente de la regla j definidos en los universos de discurso de sus entradas asociadas, y la función $f_j(x)$, que calcula la consecuencia inferida.

La función consecuente más utilizada es una combinación lineal (polinomios de grado 0 y de grado 1) de las variables de entrada, ya que la utilización de polinomios de ordenes mayores o funciones no polinómicas hace más complejo el sistema y su interpretación. Cuando la función es de orden 0, las reglas son de la siguiente forma:

$$R_1: Si\ x_1\ es\ A_1^j\ y \dots y\ x_n\ es\ A_n^j \text{ entonces } y = w_0^j,\ j = 1.2 \dots M$$

Pero si la función es lineal de orden 1, las reglas se emplean de esta manera:

$$R_1: Si\ x_1\ es\ A_1^j\ y\ x_2\ es\ A_2^j\ y\ \dots y\ x_n\ es\ A_n^j \qquad \text{entonces}$$

$$y = w_0^j + w_1^j x_1 + w_2^j x_2 + \cdots + w_n^j x_n .\ j = 1.2 \dots M.$$

El modelo TSK se ha aplicado con éxito a una gran variedad de problemas, ya que con un consecuente polinómico permite aproximar funciones con gran precisión y con un número menor de reglas que los sistemas de tipo Mamdani. No obstante, su principal inconveniente es que las reglas obtenidas no son tan fáciles de interpretar (Ojeda Magaña, 2010).

Así mismo Ojeda Magaña (2010) recomienda que para emplear este tipo de modelos difusos es necesario hacer uso de sistemas automáticos de aprendizaje, como los algoritmos basados en redes neuronales, ya que los consecuentes funcionales utilizados en los sistemas difusos de tipo TSK son difíciles de definir por un experto. Para el modelo ANFIS basado en un sistema Takagi-Sugeno, Guarnizo Lemus (2011) enuncia que no se requiere realizar una defusificacion.

A continuación, se expone una estructura general de una red neuronal difusa Takagi-Sugeno (TSK) con dos variables de entrada.

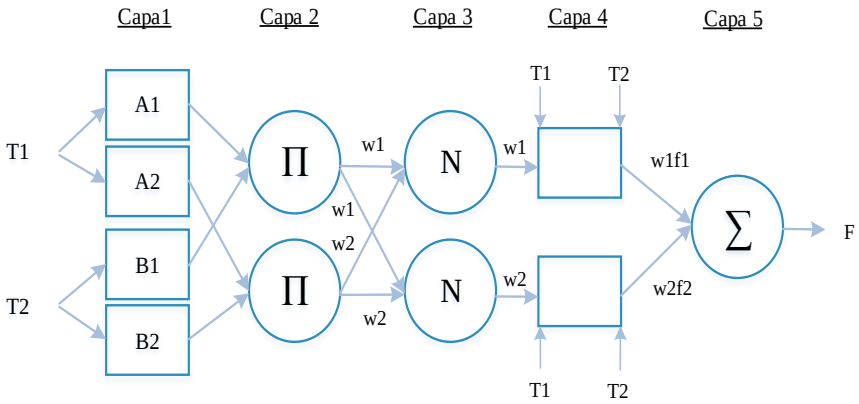

Figura 14: Estructura TSK. Adaptado de Abdulshahed, Longstaff, & Fletcher (2015)

En la Figura 14 se observa que la arquitectura del ANFIS con Takagi-Sugeno está constituida por cinco capas, donde cada capa está formada por diferentes nodos, unos son los nodos adaptativos, los cuales se representan por cuadrados e indican los conjuntos de parámetros que son ajustables en estos nodos; por otro lado, están los nodos denotados con círculos que presentan los conjuntos de los parámetros que son fijos en el modelo. El ejemplo mostrado en la figura anterior consta de dos variables de entrada $T1$ y $T2$ y una variable de salida F.

Capa 1: en la primera capa se realiza el proceso de fusificación mencionado anteriormente, mediante funciones de pertenencia; esta capa contiene nodos adaptativos con funciones de nodo descritas de la siguiente forma:

$$O_1 = \mu_{A_i}(T_1), \qquad para\ i = 1{,}2 \tag{19}$$

$$O_1 = \mu_{B_{i-2}}(T_2), \qquad para\ i = 3{,}4 \tag{20}$$

Donde A y B son las etiquetas lingüísticas asociadas a cada entrada. Existen muchas funciones de pertenencia, pero de acuerdo con Abdulshahed et al. (2015), una función gaussiana con máximo 1 y minino 0 generalmente se adapta a este modelo.

Capa 2: esta capa contiene nodos fijos denotados por Π, donde la función del nodo se multiplica por las señales de entrada las cuales servirán luego como señales de salida.

$$O_{2,i} = w_i = \mu_{A_i}(T_1) \,.\, \mu_{B_{i-2}}(T_2), \;\; para\; i = 1{,}2 \tag{21}$$

Siendo $O_{2,i}$ la salida de la capa 2 y w_i indica la señal de salida que representa la fuerza de disparo de la regla.

Capa 3: los nodos de esta capa también son nodos fijos etiquetados por la letra N, con la función de normalizar la fuerza de disparo de la regla mediante el cálculo de la relación entre la fuerza de disparo de la regla del nodo i con la suma de las fuerzas de disparo de todas las reglas.

$$O_{3,i} = \overline{w} = \frac{w_i}{w_1 + w_2}\,, \qquad para\; i = 1{,}2 \tag{22}$$

En la cual $O_{3,i}$ indica la salida de la capa 3 y $\overline{w}$ presenta la fuerza de disparo normalizada.

Capa 4: en esta capa los nodos son adaptativos y la función del nodo es:

$$O_{4,i} = \overline{w_i} \,.\, f_i\,, \qquad para\; i = 1{,}2 \tag{23}$$

Donde f_1 y f_2 son las reglas *the fuzzy if–then:*

Regla 1: si T_1 es A_1 y T_2 es B_1, entonces $f_1 = p_1T_1 + q_1T_2 + r_1$

Regla 2: si T_1 es A_2 y T_2 es B_2, entonces $f_2 = p_2T_1 + q_2T_2 + r_2$

Siendo $p_i, q_i \; y \; r_i$ los parámetros establecidos como parámetros consecuentes.

Capa 5: en esta capa cada nodo es fijo y se denota por $\sum$, con la función de nodo para calcular la salida definida como:

$$O_{5,i} = \sum_i \overline{w_i} \, . \, f_i = \frac{\sum_i w_i f_i}{w_i} = f_{out} = salida \; global \tag{24}$$

5.8 Indicadores de precisión

La evaluación y validación de los modelos es otra fase trascendental para determinar el rendimiento y poder de previsibilidad de valores futuros de cada modelo. Recientemente, se utilizan medidas de error para probar su desempeño, dentro de las cuales las más usadas son:

Error porcentual absoluto medio (*Mean Absolute Percentage Error - MAPE*): mide la exactitud de un modelo al calcular el tamaño del error en términos porcentuales y comparar cada valor observado con cada valor predicho mediante su diferencia.

$$\frac{1}{n} . \sum_{i=1}^{n} \frac{|ref_i - X_i|}{ref_i} \; x100 \tag{25}$$

Raíz del error cuadrático medio (*Root-Mean-Square Deviation - RMSE*): prueba el desempeño de forma cuantitativa calculando la raíz del error cuadrático medio a través de la fracción generada por la diferencia de los datos al cuadrado y el número de datos. El RMSE amplifica y penaliza con mayor fuerza los errores de mayor magnitud.

$$\sqrt{\frac{1}{n}.\sum_{i=1}^{n}(ref_i - X_i)^2} \quad (26)$$

5.9 Precios de venta de productos agrícolas

Antes de hablar acerca de los precios de venta de productos agrícolas es necesario definir qué se entiende por precio: precio es la suma de valores o el dinero que los consumidores dan a cambio de los beneficios que obtienen de utilizar el producto o servicio. El precio es el factor que más influye en las decisiones de los clientes, la participación del mercado y la rentabilidad de una empresa (Kotler & Armstrong, 2013).

La función del precio es revelar cuando es escaso o abundante un bien o servicio, permite direccionar o indica donde se demandan bienes, servicios o factores de producción, para facilitar a los productores tomar decisiones acerca de si se mantienen o cambian su estructura de producción o si deben salir del mercado (Acosta Leal, 2014).

Definir los precios agrícolas seria condicionar lo que representan, ya que es un concepto bastante amplio y que depende de muchos aspectos, pues los precios agrícolas se analizan desde varios escenarios, por ejemplo, en la finca, al por mayor urbano y rural, y al consumidor; también en las temporadas de escasez, los periodos de cosechas, en fronteras, según la calidad del producto, para importaciones y exportaciones, etc. (FAO., 2004).

Según la FAO los precios pueden analizarse también desde otra perspectiva, que son los precios relativos o reales; “los precios agrícolas reales se calculan dividiendo los precios agrícolas nominales, o brutos, por otros precios: los de otros sectores o los de la economía en su conjunto” (FAO, 2004) para poder analizar las variaciones de los precios en el sector agrícola y sus políticas,

los índices de los precios reales y nominales se hacen muy útiles, ya que respaldan la toma de decisiones con la información que generan acerca del desempeño y tendencias de los precios.

De acuerdo con Acosta Leal (2014)**,** los precios de los productos agrícolas dependen mayormente del mercado en la relación demanda-oferta y las necesidades del agricultor, ya que su producción es estacional, lo que hace que en épocas de cosecha los precios bajen y aumente la oferta, y suban cuando no hay cosecha, y aunque el agricultor prefiera almacenarlo para lograr un mejor precio tiene el riesgo de la alta perfectibilidad de los productos y la necesidad económica.

Para el desarrollo del presente proyecto se plantea el uso de precios de productos agrícolas de la central de abastos de Bucaramanga obtenidos del Sistema de Información de Precios y Abastecimiento del Sector Agropecuario (SIPSA).

Los productos que son seleccionados para el desarrollo del presente proyecto se encuentran adjuntados en el Apéndice A.

6. Metodología

Para el desarrollo de la investigación se plantea una metodología que consta de las fases que se mencionan a continuación con sus respectivos pasos; el diagrama presente en la Figura 15 contempla cada uno de las fases que son desarrolladas de forma general. Las fases metodológicas del proyecto son desarrolladas en el software R versión 3.4.2 (2018).

6.1 Primera fase: Análisis de los datos, limpieza e imputación

- Selección de la fuente de información que permita obtener los datos necesarios para la construcción y análisis de los modelos.

- Selección de los productos que serán utilizados para el desarrollo del proyecto.
- Aplicación del método de limpieza de datos.
- Selección de la técnica de imputación de los datos faltantes. La imputación se desarrolla mediante el uso de 5 técnicas de imputación de las que se escoge la de mejor ajuste; las técnicas utilizadas son la imputación con base en la media, mediana y moda desarrolladas a través de la función *na.mean,* imputación con base en un promedio móvil ponderado a través de la función *na.ma* y la imputación con base en una interpolación lineal a través de la función *na.interpolation* del paquete *imputeTS* desarrollado por Moritz & Bartz-Beielstein (2017).
- Realizar una caracterización de los datos con el fin de identificar propiedades y comportamientos de los mismos. Se realiza un diagrama de caja y bigotes a través de la función *boxplot* del paquete *ggplot2* desarrollado por Wickham (2016), un diagrama de correlación mediante la función *corrplot* del paquete *corrplot* desarrollado por Wei & Simko (2017) y una serie de graficas temporales a través de la función *plotly* del paquete *plotly* desarrollado por Sievert (2018).

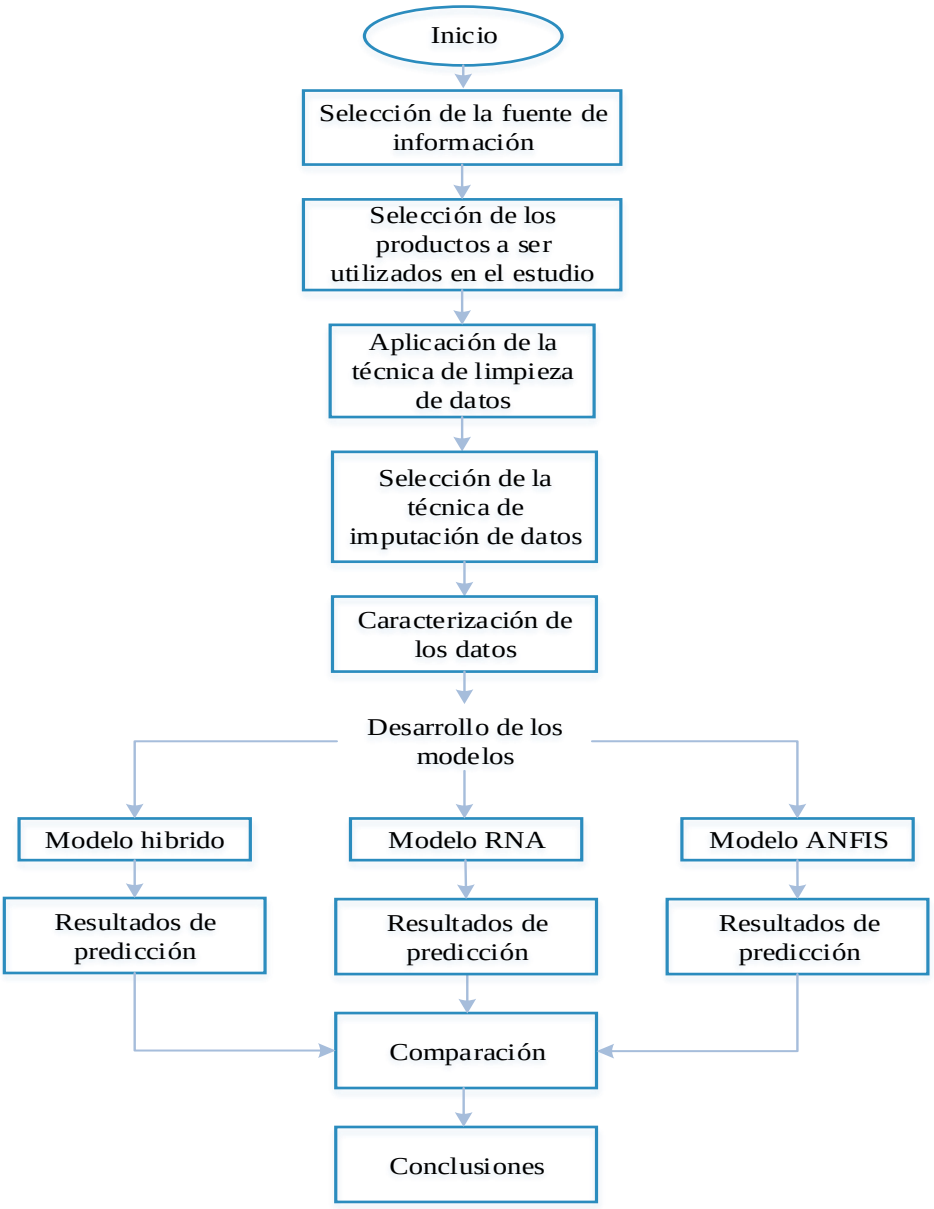

Figura 15: Metodología general de la investigación.

6.2 Segunda fase: Modelo ARIMA-RNA *ff* (*híbrido).*

Desarrollo del primer modelo de pronóstico híbrido ARIMA-RNA *ff*. El diagrama que se expone en la Figura 16 plantea de forma específica los pasos que se deben desarrollar para la generación de los modelos.

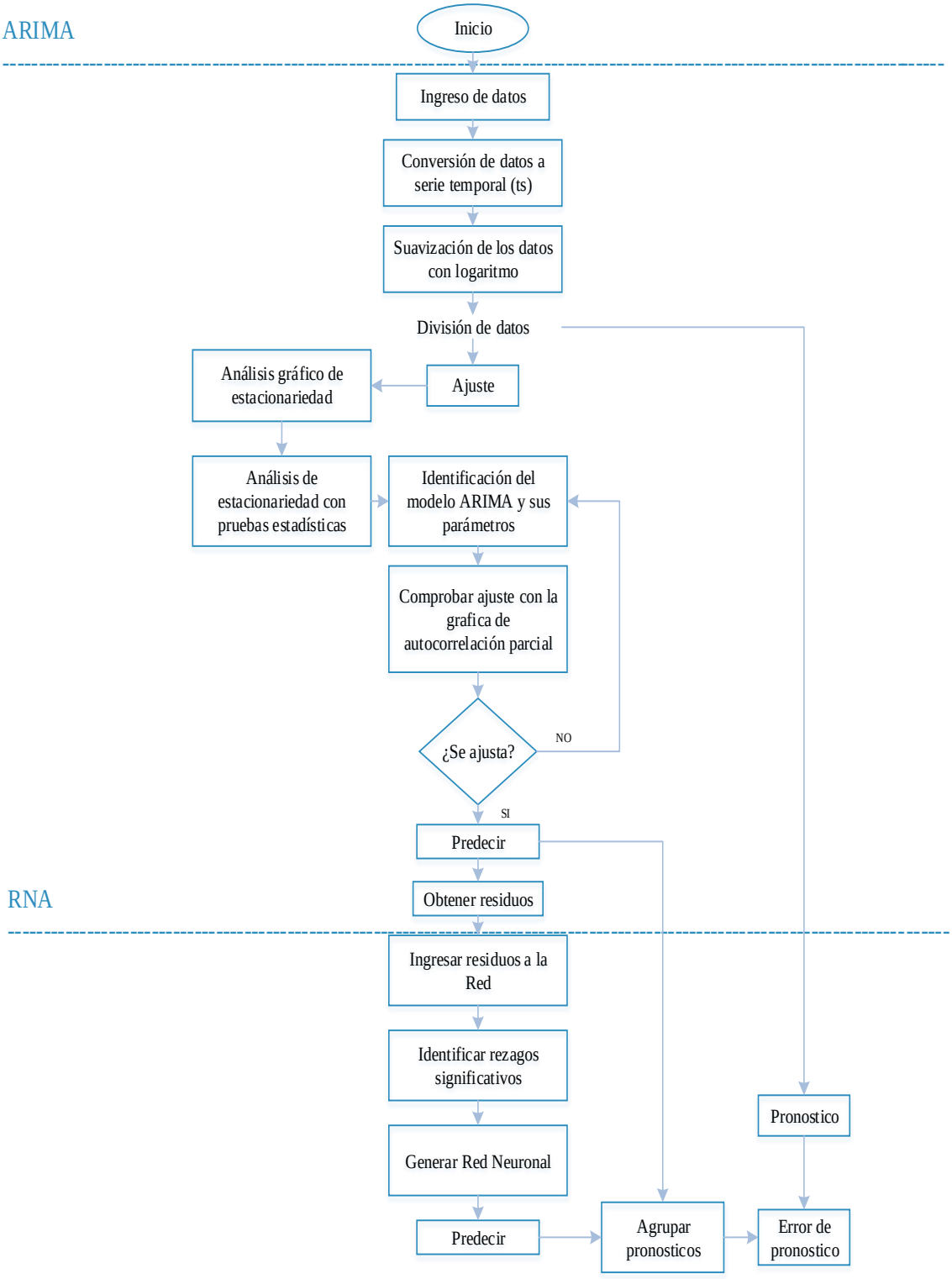

Figura 16: Metodología modelo hibrido.

- Realizar un tratamiento al conjunto de datos con el objetivo de ajustar los datos según los requerimientos de procesamiento exigidos por la función a utilizar; se realiza una transformación y suavización de los datos a través de la función *ts* del paquete *stats* desarrollado por R Core Team (2018) y la función *log* del paquete *base* desarrollado por R Core Team (2018) .
- Análisis gráfico de estacionariedad. Con el objetivo de definir si los errores de los datos se encuentran relacionados y por lo tanto poseen un comportamiento no aleatorio, se producen las gráficas de autocorrelación simple. La gráfica de autocorrelación se obtiene mediante la aplicación de la función *acf* del paquete *stats*.
- Pruebas de hipótesis para comprobar estacionariedad o aleatoriedad en los datos. Las pruebas que se desarrollan son la prueba de Durbin-Watson mediante el uso de la función *dwtest* del paquete *lmtest* desarrollado por Zeileis & Hothorn (2002), la prueba de raíz unitaria a través de la funcion *ur.df* del paquete *urca* desarrollado por Pfaff (2008) y la prueba de Bartels haciendo uso de la función *bartels.rank.test* del paquete *randtest* desarrollado por Caeiro & Mateus (2015).
- División de los datos. Los datos son divididos en datos de procesamiento y datos de pronóstico.
- Desarrollo del modelo. Los modelos se desarrollan utilizando la metodología Box-Jenkins; esta metodología consta de 4 pasos divididos en identificación, estimación, examen de diagnóstico y pronóstico como se muestra en la Figura 17.

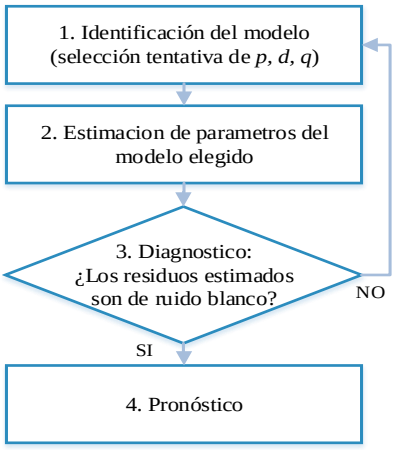

Figura 17: Metodología Box Jenkins. Adaptado de Gujarati & Porter (2010).

La parte ARIMA del modelo se desarrolla mediante la función *auto.arima* (identificación) del paquete *forescast* desarrollado por Hyndman R Athanasopoulos G (2013), y la función *arima* (estimación) del paquete *stats* y se ajusta a través de la función de autocorrelación parcial *pacf* del paquete *stats* con el uso del valor AIC y la parte neuronal del modelo se desarrolla mediante la función *nnetar*, y la predicción para ambas partes se desarrolla a partir de la función *forecast* del paquete *forecast* desarrollado por Hyndman R Athanasopoulos G (2013).

- Validación. La validación de la parte ARIMA del modelo se lleva a cabo mediante el uso la gráfica de autocorrelación parcial a través de la función *pacf* y la prueba Ljung Box mediante la función *Box.test*, ambas del paquete *stats*. La validación del modelo RNA *ff* se lleva a cabo mediante la validación cruzada de los datos.
- Predicción. En este paso se desarrolla el pronóstico de los datos 15 pasos adelante.

6.3 Tercera fase: Modelo RNA *bp*

Desarrollo del modelo de pronóstico RNA *bp*. El modelo RNA *bp* se desarrolla a partir de la metodología que se expone en la Figura 18. Esta metodología permite observar los pasos planteados para el desarrollo del modelo neuronal; la metodología puede ser sintetizada en los pasos que se mencionan a continuación.

- Identificar los rezagos significativos para ser utilizados como nodos/variables de entrada. Para determinar los rezagos que son utilizados como entradas del modelo se utiliza la gráfica de autocorrelación parcial a través de la función *pacf* del paquete *stats*.
- Preprocesamiento. Se aplica una normalización de los datos con el fin de poder utilizarlos en la función *neuralnet*.
- División de los datos. Los datos se dividen en dos conjuntos, un conjunto de entrenamiento y un conjunto de prueba, donde este último posee los datos que son utilizados para la comparación del pronóstico.
- Desarrollo de los modelos neuronales. Para desarrollar los modelos de Red Neuronal Artificial se utiliza la función *neuralnet* del paquete *neuralnet* desarrollado por Fritsch & Guenther (2016).
- Validación. Se realiza a través de la validación cruzada de los datos y con el uso de las métricas de error.
- Predicción. En este paso se pronostican los datos mediante el uso de la función *compute* del paquete *neuralnet*.

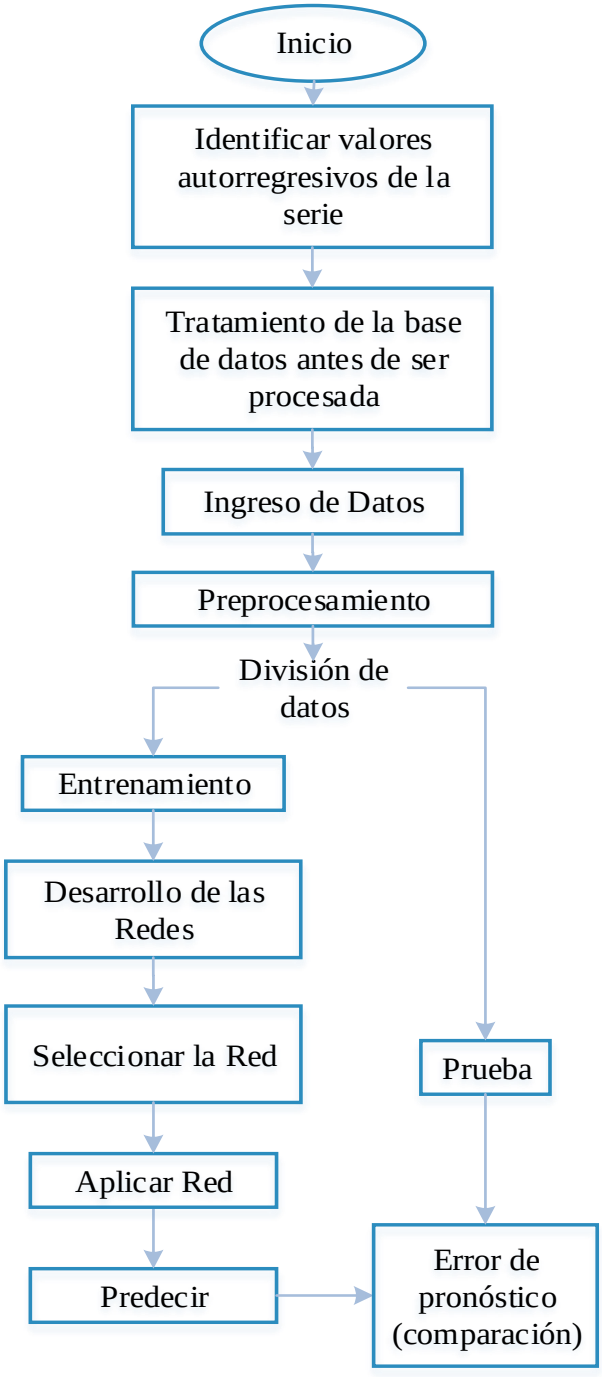

Figura 18: Metodología modelo de la Red Neuronal Artificial *backpropagation*.

6.4 Cuarta fase: Modelo ANFIS

Desarrollo del modelo de pronóstico ANFIS. Para el desarrollo del tercer modelo se plantea la metodología que se observa en la Figura 19 los cuales son descritos a continuación.

- Identificar los rezagos significativos para ser utilizados como variables de entrada. En este paso se utilizan las mismas entradas definidas para el modelo RNA *bp*, ya que estas proceden del análisis de correlación de los rezagos. Para determinar los rezagos que son utilizados como entradas del modelo se utiliza la gráfica de autocorrelación parcial a través de la función *pacf* del paquete *stats*.
- División de los datos. Los datos se dividen en dos conjuntos, un conjunto de entrenamiento y un conjunto de prueba, donde este último posee los datos que son utilizados para la comparación del pronóstico.
- Desarrollo de los modelos ANFIS. Para desarrollar los modelos ANFIS se utiliza la función *object.reg* del paquete *frbs* desarrollado por Riza et al. (2015).
- Verificación del modelo. Se realiza a través de la validación cruzada de los datos y a través de las métricas de error.
- Predicción. En este paso se pronostican los datos mediante la función *predict* del paquete *frbs* desarrollado por Riza et al. (2015).

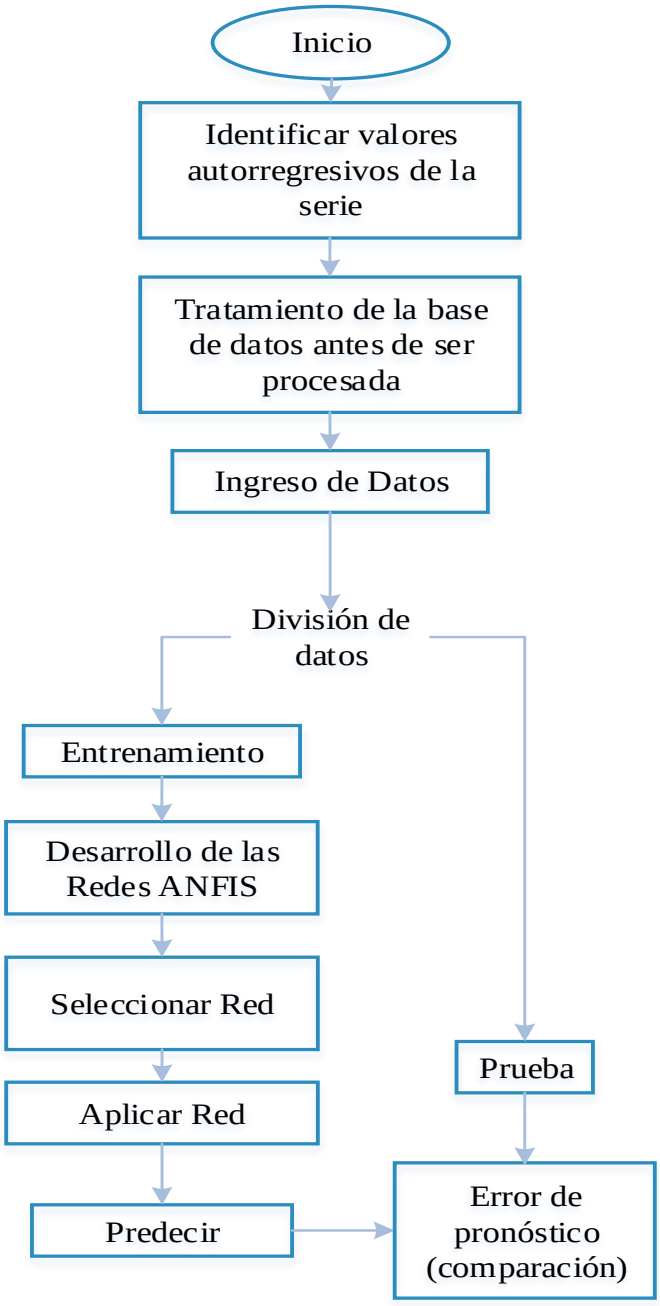

Figura 19: Metodología modelo ANFIS.

6.5 Quinta fase: Comparación

Se realiza la comparación de los modelos mediante las métricas de error RMSE y MAPE, tiempo de procesamiento y un análisis de varianzas (ANOVA) según el modelo, el día y el producto.

7. Caso de estudio en el proyecto

7.1 Análisis de datos

7.1.1 Selección de la fuente de información y generación de la base de datos. La fuente de información seleccionada para la obtención de la base de datos fue el Sistema de información de precios y abastecimiento del sector agropecuario (SIPSA) (DANE, 2018), ya que posee los precios de venta de los productos en la central de abastos de Bucaramanga. De los datos obtenidos a partir de la página del DANE en la sección del SIPSA se encuentra que los fines de semana no se producen boletines, por tanto, los días sábados y domingos (fines de semana) no se tienen en cuenta para el presente estudio. Se genera una base de datos con los precios de los productos en la herramienta Excel 2016.

7.1.2 Selección de los productos agrícolas. La selección de los productos agrícolas se realiza teniendo como criterio el potencial productivo que presenta el departamento de Santander, el cual, según el último censo realizado por el DANE, posee aproximadamente 507.000 hectáreas para el cultivo agrícola, las cuales representan el 26.1% del territorio del departamento; de este porcentaje el 92% se encuentra destinado a tierra cultivada, 7.1% en tierra de descanso y 0.9% en tierra de barbecho. Esas hectáreas se encuentran divididas en las 6 provincias de Santander como se expone en la Figura 20.

Se puede observar que a partir de la Figura 20, la provincia de mares presenta la mayor área de siembra de productos agrícolas en el departamento con 148.201 hectáreas, seguido por la provincia de Vélez con 86.401 hectáreas, Soto con 85.918, Guanentá con 78.457 hectáreas, y las provincias Comunera y García Rovira con 69.669 y 39.703 hectáreas respectivamente.

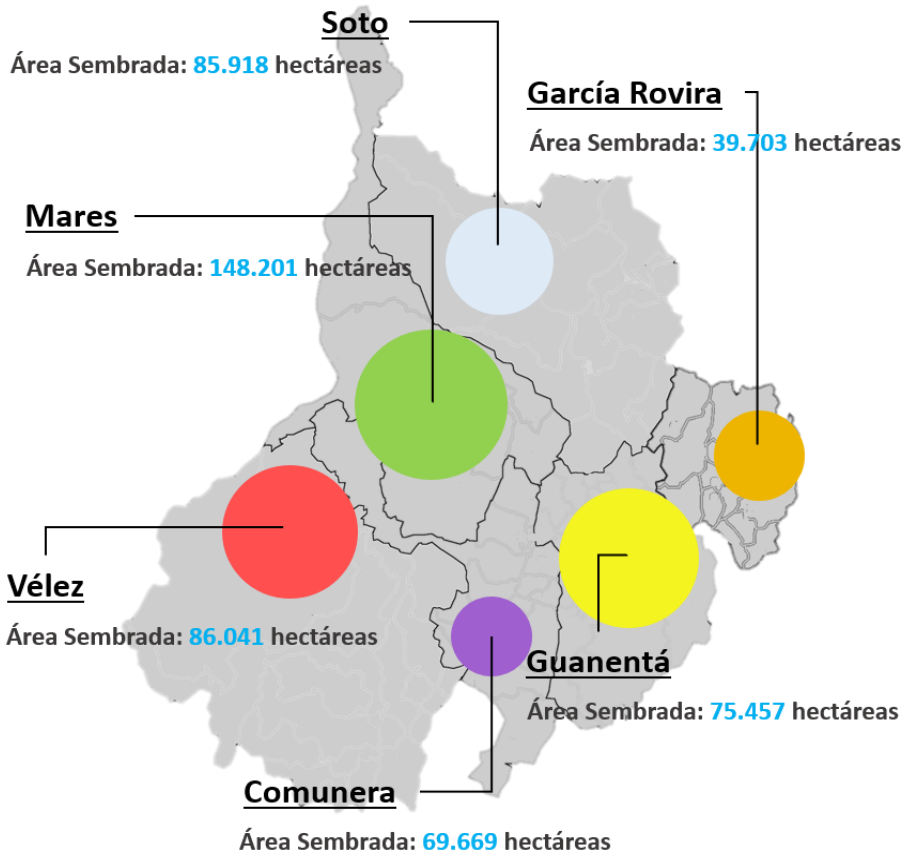

Figura 20: Área sembrada por provincia en Santander. Adaptado de Cámara de Comercio de Bucaramanga (2018). https://www.camaradirecta.com/temas/documentos%20pdf/informes%20actualidad%20provincias/agricola%20provincias%20santander%20c2014.pdf

Según estudios de la Cámara de Comercio de Bucaramanga los principales cultivos producidos en cada provincia son:

Tabla 3
Porcentaje de área sembrada por grupo de cultivos

Provincia	Agroindustriales	Fruta	Tubérculos	Cereales	Flores
Soto	39.20%	42.80%	10.40%	7.50%	0.10%
García Rovira	71.60%	7.30%	11.90%	9.10%	0.10%
Mares	73.80%	15.70%	6.90%	3.60%	0.01%
Vélez	60.70%	17.80%	13.40%	8.10%	0.04%
Guanentá	74.50%	7.70%	7%	10.70%	0.10%
Comunera	71.60%	7.30%	11.90%	9.10%	0.10%

Nota. Adaptado de Cámara de Comercio de Bucaramanga (2018). https://www.camaradirecta.com/temas/documentos%20pdf/informes%20actualidad%20provincias/agricola%20provincias%20santander%20c2014.pdf

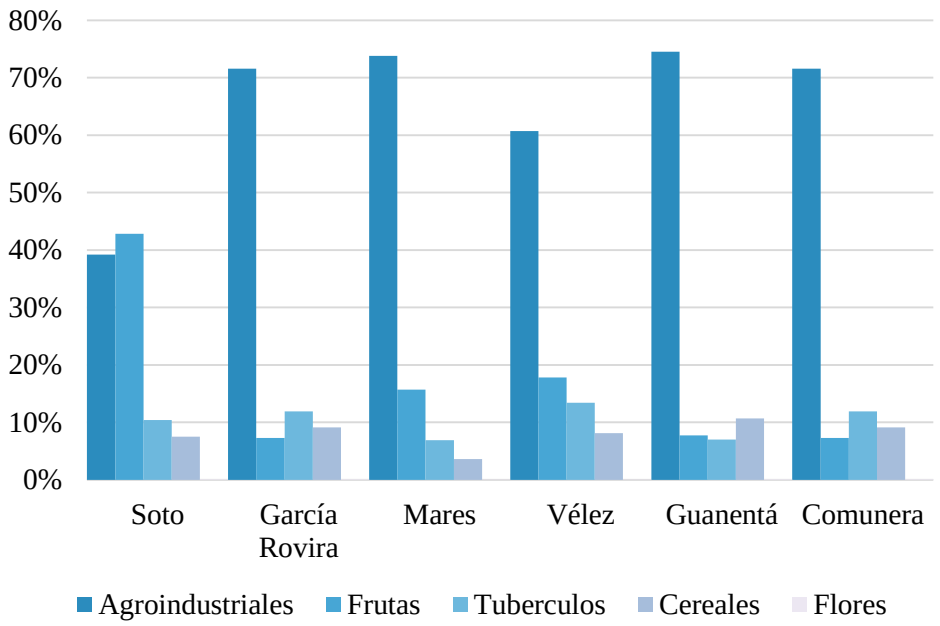

Figura 21: Área sembrada por grupo de cultivos. Adaptado de Cámara de Comercio de Bucaramanga (2018). https://www.camaradirecta.com/temas/documentos%20pdf/informes%20actualidad%20provincias/agricola%20provincias%20santander%20c2014.pdf

En la figura anterior se observa que los cultivos predominantes cambian dependiendo de la provincia; sin embargo, se puede establecer que los cultivos agroindustriales mantienen su importancia a nivel de hectáreas cultivadas en la mayoría de las provincias, solo siendo superada por los cultivos frutales en la provincia de Soto; no obstante, para el estudio que se plantea desarrollar los productos agrícolas obtenidos por medio de los cultivos agroindustriales como lo son el cacao, el café, la caña panelera y la palma de aceite, aunque son los cultivos con mayor presencia en las provincias y en general en el departamento (Gobernación de Santander, 2017), no se tendrán en cuenta debido a que no son comercializados en la central de abastos de Bucaramanga y sus precios no se encuentran registrados en la página del Sistema de Información de Precios y

Abastecimiento del Sector Agropecuario (SIPSA), lo que dificulta su recolección y posterior estudio, imposibilitando el desarrollo del proyecto. Teniendo en cuenta lo enunciado anteriormente se propone el uso de productos agrícolas no agroindustriales tanto permanentes como transitorios para la investigación, los cuales representan al año 2016 el 40% del área total sembrada en el departamento (Agronet, 2016).

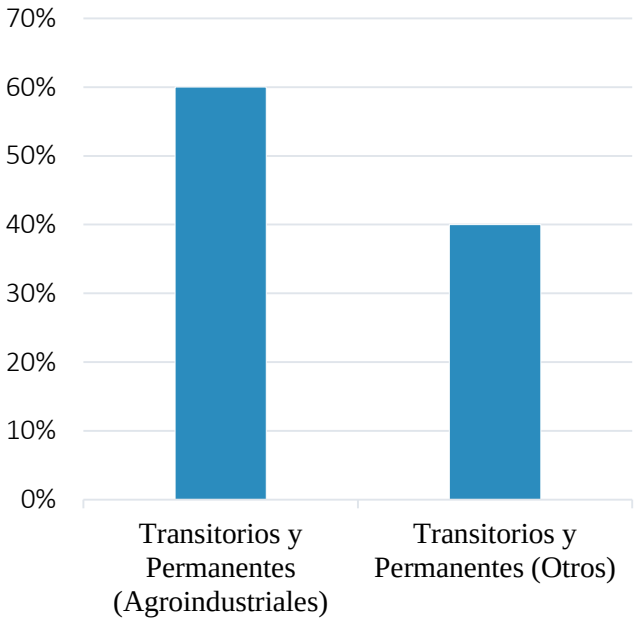

Figura 22: Participación de los cultivos en Santander. Adaptado de Agronet (2016). http://www.agronet.gov.co/Documents/Santander2016.pdf

Dentro de este tipo de cultivos se preseleccionan los más relevantes a nivel de área por cultivo para la región y el departamento según los últimos informes presentados del Ministerio de Agricultura por medio de la red de información y comunicación estratégica del sector agropecuario (Agronet, 2016). En la siguiente tabla se presentan los productos preseleccionados para la investigación:

Tabla 4
Superficie por tipo de cultivo sembrado en Santander

Tipo de cultivo	Producto	Superficie sembrada en hectáreas
Transitorio	Yuca	9.248
Transitorio	Maíz	12.364
Transitorio	Frijol	9.138
Transitorio	Papa	5.782
Transitorio	Tomate	1.963
Permanente	Plátano	17.538
Permanente	Banano	3.335
Permanente	Cítricos	16.946
Permanente	Piña	11.517
Otros		43.524

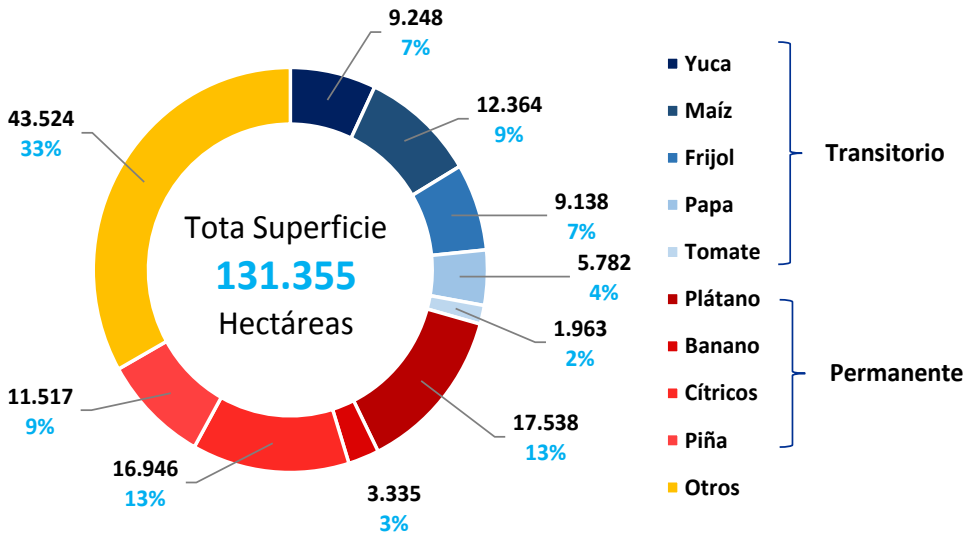

Figura 23: Hectáreas por cultivo en Santander.

7.1.3 Limpieza de la base de datos. En esta sección se exponen los datos faltantes que se tienen por día para cada uno de los productos mediante la base de datos generada y su respectiva limpieza. En la Tabla 5 se puede observar la cantidad de datos faltantes de los recolectados a través del SIPSA.

Tabla 5
Datos Faltantes

Días	Datos	Yuca	Plátano	Papa	Piña	Naranja	Mandarina	Banano	Tomate
Lunes	Faltantes	301	301	301	301	301	301	301	301
	Porcentaje	98.05%	98.05%	98.05%	98.05%	98.05%	98.05%	98.05%	98.05%
Martes	Faltantes	6	6	6	6	7	24	6	6
	Porcentaje	1.95%	1.95%	1.95%	1.95%	2.28%	7.82%	1.95%	1.95%
Miércoles	Faltantes	288	288	288	288	288	292	289	288
	Porcentaje	93.81%	93.81%	93.81%	93.81%	93.81%	95.11%	94.14%	93.81%
Jueves	Faltantes	20	20	21	20	21	35	20	21
	Porcentaje	6.51%	6.51%	6.84%	6.51%	6.84%	11.40%	6.51%	6.84%
Viernes	Faltantes	19	21	16	19	26	56	18	20
	Porcentaje	6.19%	6.84%	5.21%	6.19%	8.47%	18.24%	5.54%	6.51%

Del total de los datos que se presentan en la Tabla 5 se observa que del máximo número posible de datos por día (307), en los días lunes y miércoles el porcentaje de datos faltantes (datos que no existen porque el SIPSA no público el boletín o no salió el informe de la ciudad dentro del boletín) o perdidos (datos de precios que no salieron dentro del informe de la ciudad) son aproximadamente 98.05% y 93.81% respectivamente, por lo que se decide eliminar estos días mediante la metodología *list wise* (método de limpieza que elimina una categoría si la cantidad de datos faltantes es significativa), ya que utilizarlos implicaría una imputación de datos de aproximadamente el 58%, lo que según Medina & Galván (2007b) no se recomienda, debido a que la cantidad de datos a imputar supera el 20% de la totalidad, lo que disminuye la precisión en el pronóstico. De esta forma, se mantienen para el estudio los datos distribuidos en los días martes, jueves y viernes (921 datos entre datos existentes, perdidos y faltantes) solo necesitando imputar de esta manera una totalidad de alrededor de 47 datos por producto (a excepción de la mandarina), lo que reduce el porcentaje de datos a determinar a un 5% aproximadamente, procurando evitar así incurrir en errores de precisión.

Teniendo en cuenta las depuraciones mencionadas anteriormente, la tabla de datos faltantes o datos a imputar se reduce a lo expuesto en la Tabla 6, obteniendo de esta forma el número de datos total a imputar.

Tabla 6
Datos perdidos de los boletines

Días	Datos	Yuca	Plátano	Papa	Piña	Naranja	Mandarina	Banano	Tomate	Total día
Martes	Faltantes	6	6	6	6	7	24	6	6	67
	Porcentaje	1.95%	1.95%	1.95%	1.95%	2.28%	7.82%	1.95%	1.95%	0.9%
Jueves	Faltantes	20	20	21	20	21	35	20	21	178
	Porcentaje	6.51%	6.51%	6.84%	6.51%	6.84%	11.40%	6.51%	6.84%	2.4%
Viernes	Faltantes	19	21	16	19	26	56	18	20	195
	Porcentaje	6.19%	6.84%	5.21%	6.19%	8.47%	18.24%	5.54%	6.51%	2.6%
Total producto	Faltantes	45	47	43	45	54	115	44	47	440
	Porcentaje	4.9%	5.1%	4.7%	4.9%	5.9%	12.5%	4.7%	5.1%	

7.1.4 Imputación de datos faltantes. Los métodos de imputación son aplicados a tres productos con comportamiento diferente con el propósito de analizar su ajuste en cada caso y escoger el mejor; esta imputación se realiza sin discriminar el día (martes, jueves y viernes), ya que los precios pueden presentar mayor variabilidad en la imputación si sólo se tienen en cuenta por día, debido que al encontrarse más alejados en el periodo de tiempo puede afectar el resultado de la misma. Las gráficas iniciales sin datos ajustados (datos faltantes) se presentan en la Figura 24 junto con las gráficas de los métodos utilizados en los productos seleccionados:

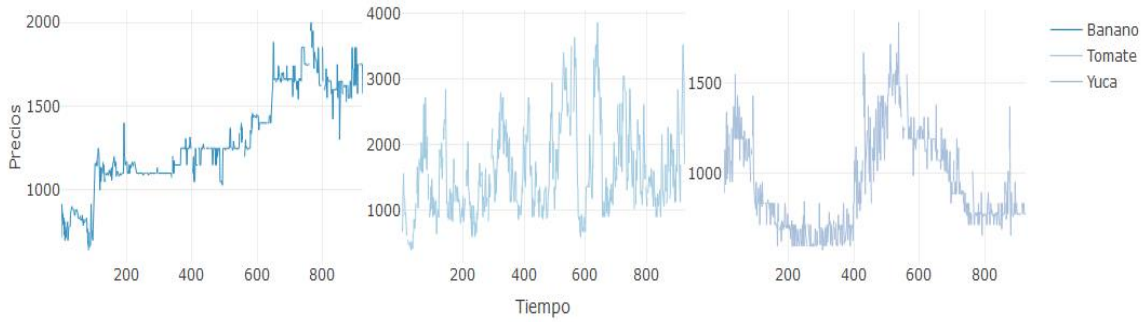

Figura 24: Representación de los datos faltantes en el día martes para los productos banano, tomate y yuca. Adaptado de R versión 3.4.2 (2018)

7.1.4.1 Imputación con base en la medía, mediana y moda. Este tipo de imputación utiliza estas medidas de tendencia central con base en los datos existentes para reemplazar a los datos faltantes.

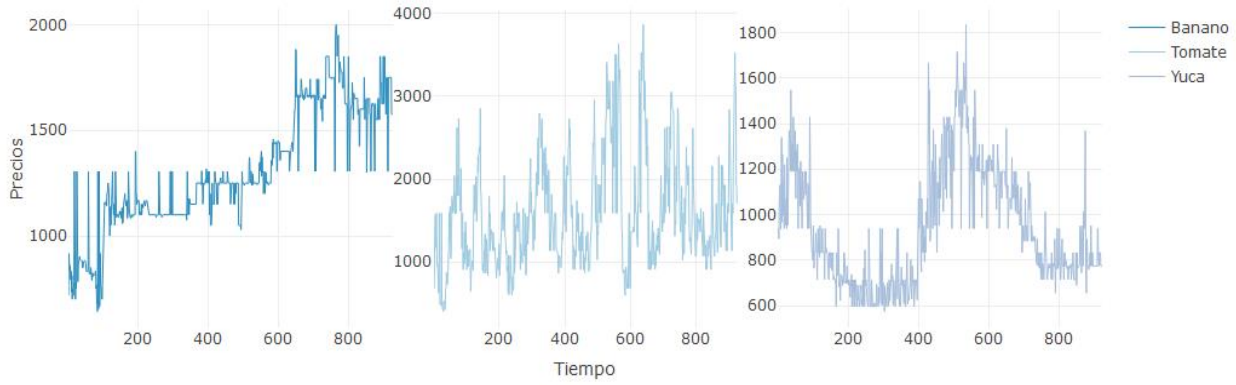

Figura 25: Imputación con base en la medía. Adaptado de R versión 3.4.2 (2018)

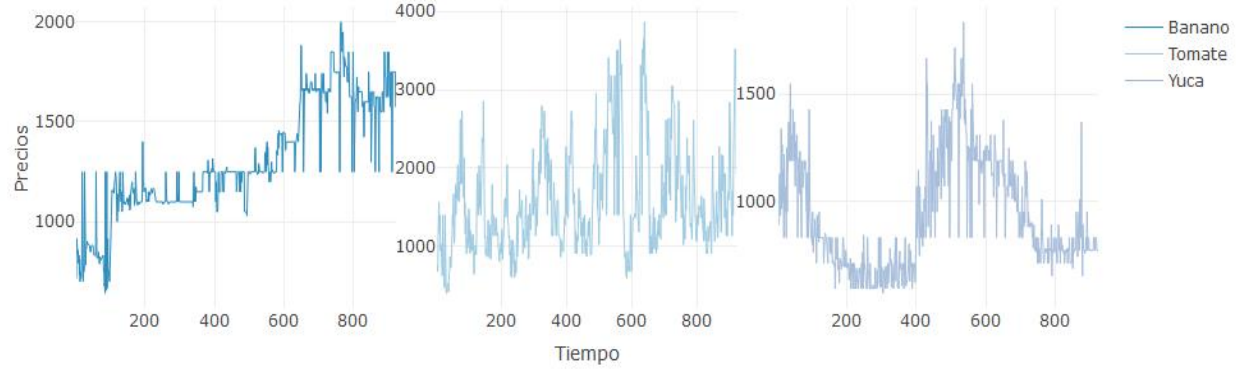

Figura 26: Imputación con base en la mediana. Adaptado de R versión 3.4.2 (2018)

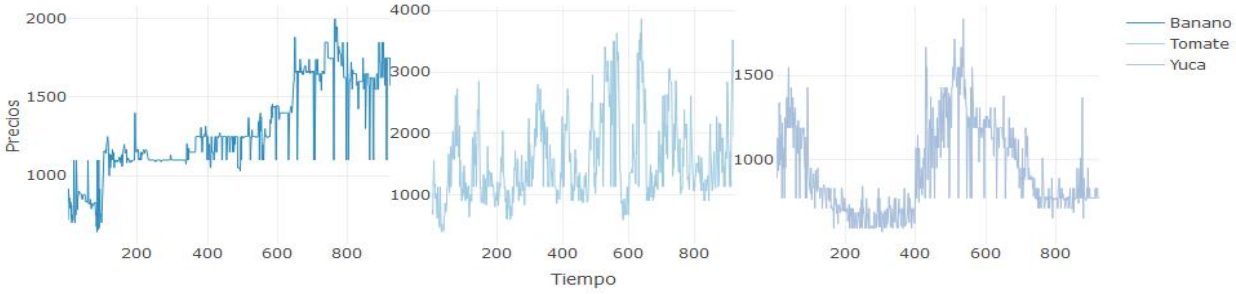

Figura 27: Imputación con base en la moda. Adaptado de R versión 3.4.2 (2018).

7.1.4.2 Imputación con promedio móvil ponderado

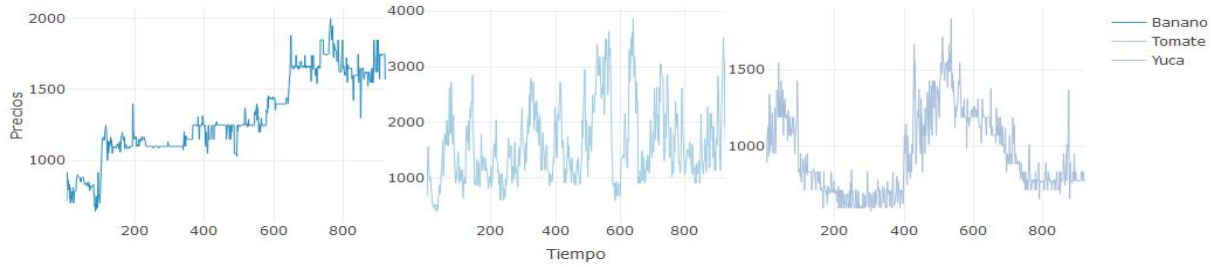

Figura 28: Imputación con base en medías móviles. Adaptado de R versión 3.4.2 (2018).

En la imputación de medías móviles los datos faltantes se determinan mediante el promedio móvil ponderado lineal con base en el número de vecinos (k) más próximos indicados a través de una ventana móvil semiadaptativa que garantice el reemplazo de todos los datos faltantes; en la imputación se trabaja una ventana $k = 5$ que tiene en cuenta 10 observaciones cinco anteriores y cinco posteriores asignando pesos ponderados mayores a los datos más cercanos al dato faltante.

7.1.4.3 Imputación por interpolación lineal. Los datos faltantes son obtenidos mediante la estimación de una interpolación generada a través de una regresión lineal, es decir, el dato faltante se cambia por el dato generado con la predicción dada por la ecuación de regresión estimada.

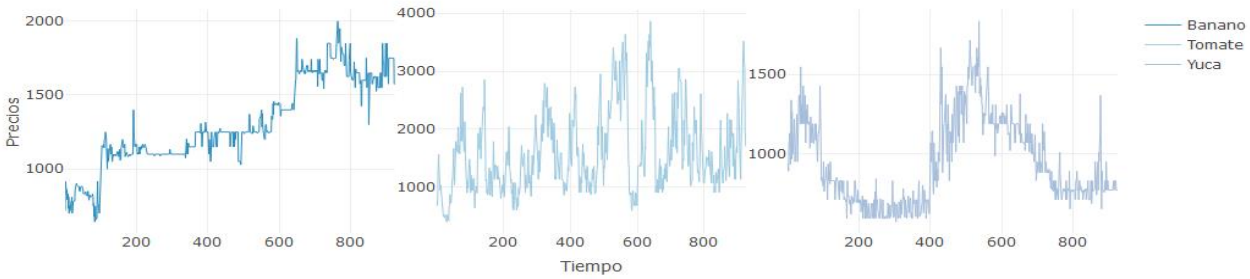

Figura 29: Imputación con base en regresión lineal. Adaptado de R versión 3.4.2 (2018).

Después de observar los tipos de imputación que se realizaron y con base en el análisis gráfico se determina que la imputación mediante el promedio móvil ponderado es la que demuestra mayor

semejanza al comportamiento que presentan los datos; esto se puede presentar debido que al poseer una ventana móvil semiadaptativa constituida por los 5 valores posteriores y los 5 anteriores intenta mantener la variabilidad, característica que poseen los datos en este periodo de tiempo.

Las imputaciones mediante medía, mediana y moda, aunque son una "estrategia sencilla y puede resultar intuitivamente satisfactoria, presenta un importante defecto y es que […] tiende a subestimar la variabilidad real de la muestra al sustituir los faltantes por valores centrales de la distribución" (Gómez, Palarea, & Martín, 2006), y la imputación a través de regresión simple (Medina & Galván, 2007b) sugieren no aplicarla "cuando el análisis secundario de datos involucra técnicas de análisis de covarianza o de correlación, ya que sobreestima la asociación entre variables".

Para estudiar y seleccionar el mejor método de imputación solo se tuvieron en cuenta los tres productos expuestos debido a que el uso de estos permite realizar un análisis gráfico en una serie temporal con tendencias y diferentes variabilidades.

7.1.5 Características de los datos. Se hace la mención que los resultados gráficos (gráficos lineales) que se exponen en las siguientes secciones están asociados únicamente al banano, ya que presentar todos los gráficos no representaría ningún agregado a lo que se desea mencionar, además de incurrir en uso excesivo del espacio en el documento; sin embargo, las demás etapas metodológicas abarcan a todos los productos del estudio.

En la Tabla 7 se pueden observar determinadas medidas de tendencia central, las cuales permiten dar un entendimiento del comportamiento de cada uno los precios sujetos al estudio, pudiendo así identificar los productos que presentan mayor y menor variabilidad como también los más estables.

Tabla 7
Características de los precios

Producto	Precio Max	Precio Min	Rango	Medía	Mediana
Banano	2000	642	1358	1304	1250
Mandarina	3000	565	2435	1350	1246
Naranja	1170	340	830	606.7	590
Papa	2200	383	1817	904.3	820
Piña	883	426	457	659.5	667
Plátano	2580	577	2003	1405	1320
Tomate	3864	402	3462	1592	1402
Yuca	1835	500	1335	940.3	833

A partir de los resultados de la anterior tabla se puede deducir que el producto con menor variación en sus precios es la piña con 457 pesos de rango, seguido de la naranja con 830 pesos, como también los productos con mayor variación son el tomate con 3.462 pesos de rango seguido de la mandarina con 2.435 pesos. Se obtiene como el producto más estable a nivel de cambio de precio a la piña, ya que fue el producto que mantuvo el valor medio más cercano a sus límites; sin embargo, el producto cuyo precio permaneció en mayor medida y durante más tiempo (moda) fue el banano con dos precios particularmente 1.250 y 1.100.

Por otro lado, se desarrolla un diagrama de bloques con el objetivo de observar si los datos poseen alta variabilidad. Con base en este diagrama se puede enunciar que en la mayoría de los productos la mediana casi siempre se mantuvo en el centro de los limites otorgados por el primer y tercer cuartil (50% de los datos), lo que permite suponer que este porcentaje de datos no tuvo gran variación con respecto a ésta, ya que aunque crecieran y disminuyeran no se alejaban, lo cual proporciona una visión general de la simetría de la distribución de los datos; sin embrago, se contempla que en algunas observaciones existen datos atípicos, siendo estos valores muy lejanos con respecto al comportamiento normal del precio.

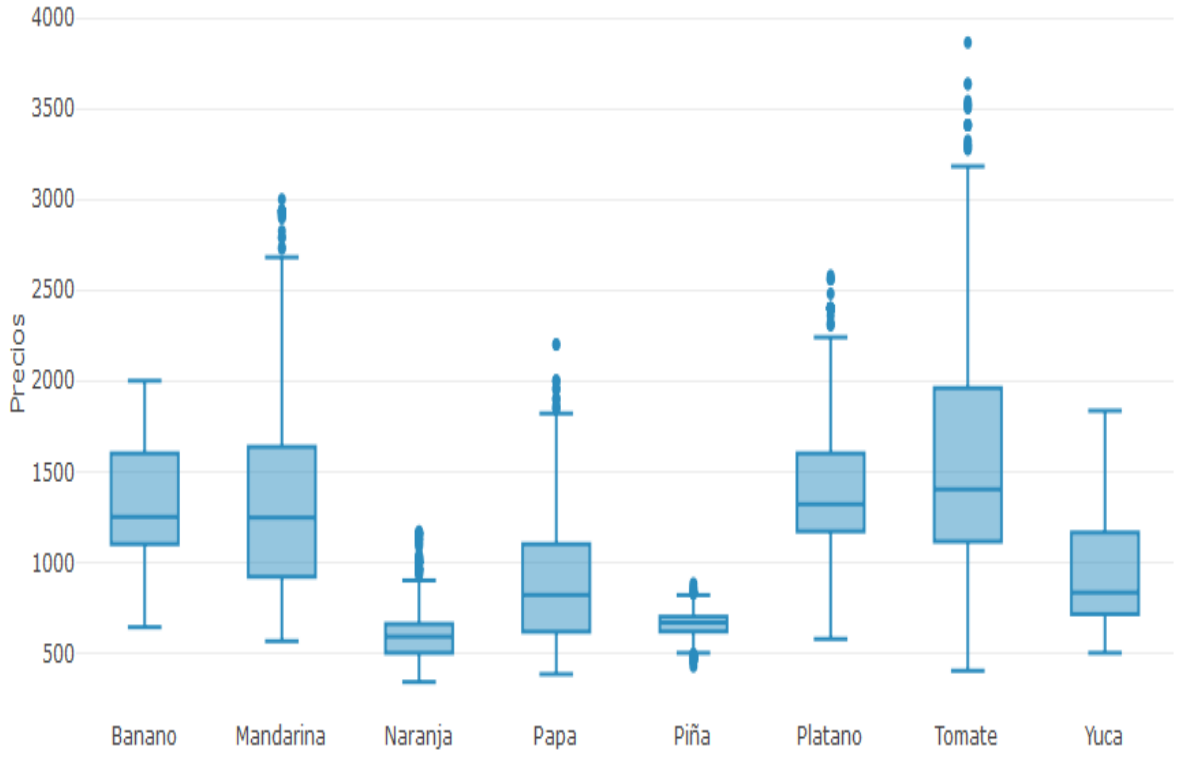

Figura 30: Diagrama de bloques de los precios de los productos en estudio. Adaptado de R versión 3.4.2 (2018)

7.1.5.1 Correlación entre productos agrícolas. Se realiza una correlación entre los productos con el propósito de observar si existe una relación entre el cambio de precio de estos. En la siguiente figura se ilustra el coeficiente de correlación entre par de los productos; la figura se puede interpretar que entre mayor sea el radio del circulo y más intenso sea su color se presenta una correlación alta entre los productos, y baja de lo contrario; siendo el azul el color (color cercano a 1) que representa una correlación serial positiva o directa y el rojo una correlación serial negativa o inversa (color cercano a -1).

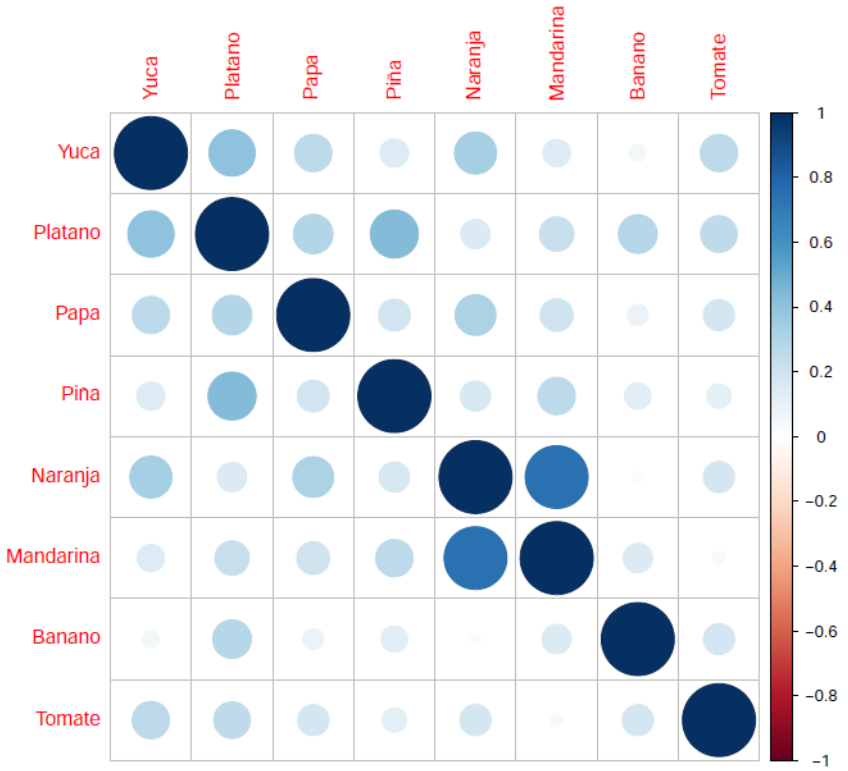

Figura 31: Mapa de calor sobre correlaciones entre productos. Adaptado de R versión 3.4.2 (2018).

Después de realizar el gráfico de correlación de los precios de los productos con el objetivo de determinar si los comportamientos de un par de productos se ven afectados simultáneamente, se encuentra que sólo la mandarina y la naranja (cítricos) poseen correlación positiva fuerte; no obstante, esto no quiere decir que un producto afecte al otro en determinada magnitud. Para determinar el grado de correlación el cual presentan la combinación de productos, se utiliza la siguiente escala como referencia, la cual divide la correlación en 9 diferentes categorías.

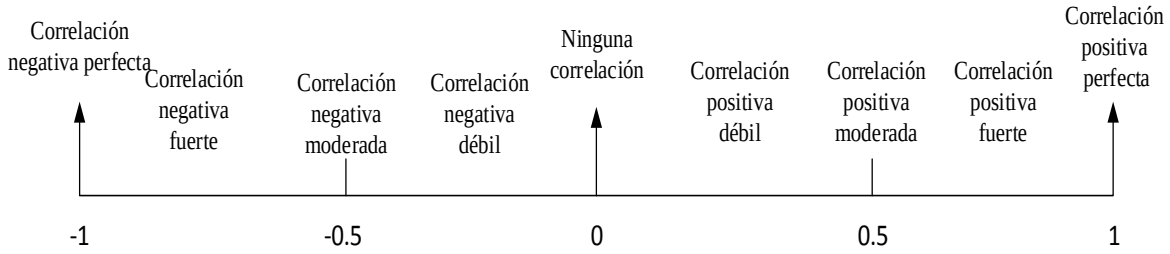

Figura 32: Análisis del coeficiente de correlación entre dos variables. Adaptado de Vila & López: https://www.uoc.edu/in3/emath/docs/RegresionLineal.pdf

Tabla 8
Correlación y categoría entre los productos

Producto 1	Producto 2	Correlación	Categoría	Producto 1	Producto 2	Correlación	Categoría
Yuca	Plátano	0.38	+ débil	Papa	Naranja	0.31	+ débil
Yuca	Papa	0.25	+ débil	Papa	Mandarina	0.21	+ débil
Yuca	Piña	0.15	+ débil	Papa	Banano	0.08	ninguna
Yuca	Naranja	0.26	+ débil	Papa	Tomate	0.17	+ débil
Yuca	Mandarina	0.09	+ débil	Piña	Naranja	0.13	+ débil
Yuca	Banano	-0.08	- débil	Piña	Mandarina	0.20	+ débil
Yuca	Tomate	0.21	+ débil	Piña	Banano	-0.15	- débil
Plátano	Papa	0.30	+ débil	Piña	Tomate	0.09	+ débil
Plátano	Piña	0.40	+ débil	Naranja	Mandarina	0.77	+ fuerte
Plátano	Naranja	0.18	+ débil	Naranja	Banano	0.15	+ débil
Plátano	Mandarina	0.21	+ débil	Naranja	Tomate	0.20	+ débil
Plátano	Banano	0.27	+ débil	Mandarina	Banano	0.24	+ débil
Plátano	Tomate	0.28	+ débil	Mandarina	Tomate	0.05	ninguna
Papa	Piña	0.17	+ débil	Banano	Tomate	0.20	+ débil

De la

Tabla *8* se puede observar que la correlación existente entre los precios de los productos se encuentra alrededor del 0.25 positiva y directa, convirtiéndolos en productos cuyo cambio de precio en el mercado no se ve afectados por los otros según la escala de Vila & López; no obstante, permite percibir una determinada relación, así sea mínima; también se observan varias correlaciones que son bajas, muy cercanas a cero en comparación con las demás, como el hecho de presentarse dependencias inversas o correlaciones negativas mínimas entre productos como el banano con la piña y la yuca.

7.1.5.2 Gráficas promedio varianza. El análisis del promedio y varianza para el precio del banano se presenta a continuación, con el fin de identificar la relación existente entre los cambios de la varianza, el promedio de los precios y observar a través de un análisis grafico cómo se comporta la variabilidad de los datos.

El eje x de las gráficas representa los años y el eje y los datos de precios, promedio y varianza normalizados; la normalización se desarrolla, debido a que los precios presentan valores

positivamente grandes, lo cual no permitiría con el uso del valor base una presentación clara de la variabilidad en el gráfico.

Figura 33: Comportamiento promedio y varianza del banano. Adaptado de R versión 3.4.2 (2018)

En la Figura 33 se puede observar que el promedio crece a través del tiempo, esto se debe a que el precio puede alcanzar valores desde 600 hasta 2000 pesos; sin embargo, la varianza se mantiene aparentemente estable hasta mediados del año 2013, donde cambia debido a la subida del precio de venta; luego mantiene una relativa estabilidad desde finales del 2013 hasta mediados del año 2014, lo cual corresponde con un periodo de precios que poseen un rango (mínimo de 574 y máximo de 845) de cambio pequeño, después se presentan cambios no muy altos hasta mediados del año 2016, año en el cual se presenta el precio más alto del banano haciendo que el promedio se aproxime a cero, indicando que los precios presentan poca variación con respecto a su promedio; luego a partir de mediados del año 2017 el precio del banano cae notablemente hasta inicios del

año 2018, donde comienza a crecer nuevamente, pero manteniendo una varianza aparentemente estable.

7.1.5.3 Gráficas por día (todos los años). Se presenta un desglose de los precios de los productos por día (martes, jueves y viernes) de tal forma que se pueda apreciar con mayor claridad los cambios que se presentan según el día, con el fin de determinar patrones y características en estos.

En la Figura 34 se puede observar que en términos generales los precios mantienen un comportamiento que posee poca variación con respecto al día a través del tiempo, exhibiendo diferencias en determinados periodos, pudiéndose inferir que los precios no se ven afectados por el día en el cual se generó el dato; sin embargo, en el proyecto se plantea el desarrollo de los modelos por día para cada producto con el objetivo de no generar un conflicto de precisión en la capacidad de predicción del modelo que se desarrolle, dado que al tener en cuenta todos los días (martes, jueves y viernes) como una serie temporal, se descarta el posible impacto que puede presentar el inicio y cierre de la semana en los precios diarios de los productos y que estos al no ser considerados en el estudio generarían resultados menos precisos.

Figura 34: Precios por día del banano. Adaptado de R versión 3.4.2 (2018)

7.1.5.4 Gráficas por año (todos los días). A continuación, se realiza un análisis del comportamiento de los precios del banano durante cada año haciendo una comparación entre cada uno de ellos con el fin de identificar si existen similitudes o diferencias de uno a otro. El eje x de las gráficas representa los días durante todo un año del periodo de estudio, teniendo en cuenta que solo están los datos de los días martes, jueves y viernes; en promedio, cada 13 días representan un mes.

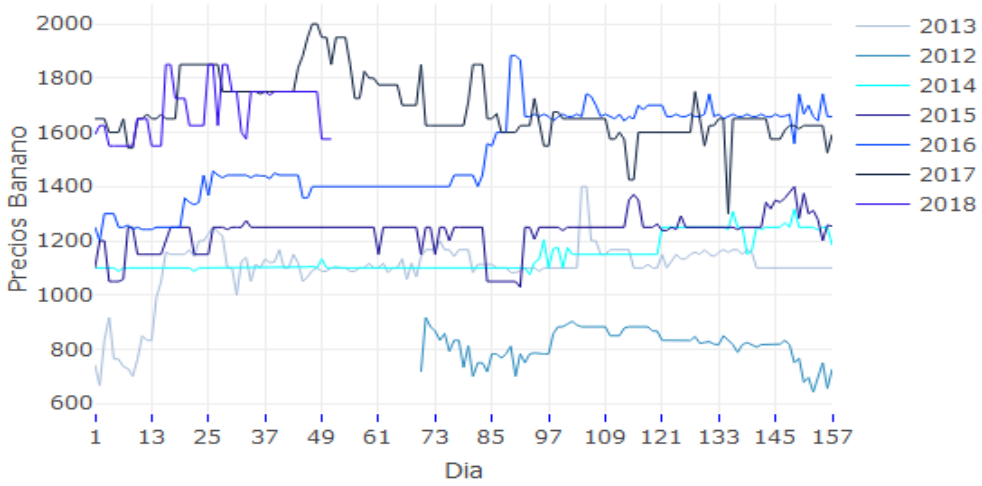

Figura 35: Comportamiento del precio del banano por año. Adaptado de R versión 3.4.2 (2018)

En la Figura 35 se observa que los precios más altos se alcanzaron durante el año 2017; también se resalta la poca variabilidad de los precios en el año 2014, los cuales muestran un comportamiento estable sólo con algunas variaciones, especialmente después del mes de agosto, además, el año 2012 presenta los precios más bajos de todos los años en su segundo semestre, este mismo comportamiento se refleja en otros productos; todos los resultados anteriormente expuestos más que pretender definir una forma diferente de abarcar el problema de los pronósticos a partir de los precios, permiten visualizar una aproximación a las características generales que poseen

estos según el producto y los periodos temporales. A continuación, en los siguientes tres incisos se presenta cada uno de los modelos propuestos con los respectivos resultados en cada una de las etapas de su desarrollo.

Se hace la observación que los modelos que se desarrollan en este apartado son de pronósticos semanales, es decir, se desarrollan modelos de pronostico independientes para los días martes, jueves y viernes, de tal forma que los datos del día martes se usen para predecir a los precios de este día y de la misma forma con el día jueves y viernes dando como resultado un total de 307 datos por día para ser usados en los modelos.

7.2 Modelo Hibrido ARIMA-RNA *ff*

El modelo ARIMA-RNA *ff* fue propuesto por Zhang (2003), el cual expone que las series temporales pueden ser pronosticadas a través del uso de un modelo lineal (ARIMA) y un modelo no lineal (RNA *ff*). Esta metodología busca ajustar un modelo ARIMA (p, d, q) a los datos con el objetivo de asimilar la parte lineal y obtener la parte no lineal mediante los residuos del mismo modelo; estos datos no lineales obtenidos de los residuos son procesados por el modelo RNA *ff* para la predicción. Una vez obtenido el modelo ARIMA y generado la red neuronal se dispone a realizar la predicción de forma independiente para luego integrar los resultados del pronóstico. A continuación, se exponen los resultados de los pasos mencionados en la segunda fase de la metodología.

7.2.1 Tratamiento de datos. Esta fase se encuentra constituida por tres partes; en la primera se ingresan los datos sobre nivel obtenidos de la base de datos del SIPSA, en la segunda se realiza una transformación de los datos a serie temporal para ser procesados dentro de las funciones del modelo, y en la tercera se realiza una suavización de los datos con el objetivo de intentar eliminar la tendencia que pueda estar presente en estos.

7.2.2 División de datos. Una vez realizados los tratamientos pertinentes los datos iniciales son divididos en dos conjuntos, el primer conjunto constituido por 292 datos es procesado por el modelo hibrido y los 15 datos (pronóstico de 3 meses y 3 semanas adelante) restantes se excluyen del modelo para ser utilizados en la comparación de la precisión del pronóstico generado por el modelo desarrollado.

7.2.3 Análisis gráfico de estacionariedad de los precios. En la Figura 36, Figura 37 y Figura 38, se observa que para el banano en cada uno de los días la correlación entre los errores disminuye o cae lentamente, pero en todos los casos se supera el límite de significancia, lo que se puede interpretar como que el resultado de la variable regresora en el presente depende o es afectada positivamente por sus rezagos en el tiempo; esto se puede presentar debido que al ser recopilados a través del tiempo en periodos muy cercanos se relacionan entre sí, afectándose mutuamente.

Se aclara que en las gráficas de autocorrelación que se muestran a continuación inician en el *lag* (valor desfasado de la variable respuesta en el tiempo) o rezago 0 (cero) por lo que se puede apreciar que en este instante la autocorrelación (ACF) existente es de 1 (dado que es la correlación consigo mismo).

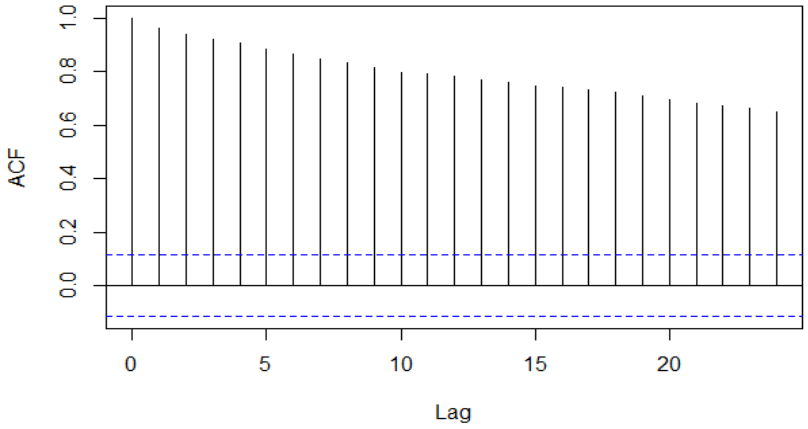

Figura 36: Autocorrelación banano del día martes. Adaptado de R versión 3.4.2 (2018)

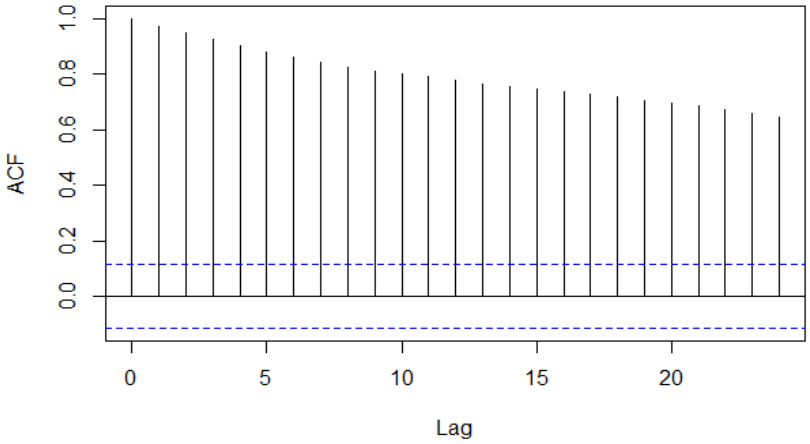

Figura 37: Autocorrelación banano día jueves. Adaptado de R versión 3.4.2 (2018

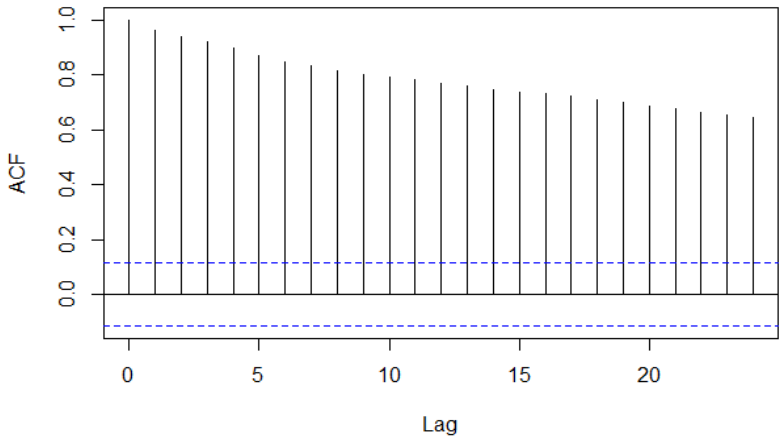

Figura 38: Autocorrelación banano día viernes. Adaptado de R versión 3.4.2 (2018

Para determinar si los errores se encuentran relacionados y de esta forma ser no aleatorios se aplica la prueba estadística d de Durbin-Watson para comprobar los resultados arrojados por las gráficas de autocorrelación simple.

7.2.4 Prueba Durbin-Watson. La prueba d se utiliza para determinar si existe autocorrelación en los residuos o errores de predicción en un análisis de regresión, este contraste permite verificar la hipótesis de no autocorrelación frente a la alternativa de autocorrelación de primer orden. Al aplicar la prueba de d a los precios de todos los productos se obtienen los siguientes resultados:

Tabla *9*
Resumen del estadístico

Producto	Martes	Jueves	Viernes
Banano	0.38	0.33	0.36
Mandarina	0.15	0.16	0.13
Naranja	0.20	0.20	0.19
Papa	0.15	0.11	0.11
Piña	0.75	0.58	0.64
Plátano	0.26	0.30	0.27
Tomate	0.44	0.38	0.41
Yuca	0.20	0.17	0.18

La diferencia entre los límites en los cuales la hipótesis no tiene respuesta o es indecisa disminuye a medida que aumenta el número datos en la muestra, siendo que para 200 datos los límites dL y dU (límite superior (dU) e inferior (dL) en los que se acepta, rechaza o no tiene respuesta la hipótesis) son 1.758 y 1.778 respectivamente; es decir, para una totalidad de 307 datos la diferencia entre los limites es menor, debido a esto y en contraste con los valores de los estadísticos obtenidos, se encuentra que estos no están dentro de dicha diferencia lo que permite concluir que los errores poseen un comportamiento de correlación serial positiva dada la cercanía del estadístico a cero, implicando que los residuos en periodos cercanos se encuentran relacionados.

7.2.5 Prueba de Raíz Unitaria. Esta prueba se evalúa con el objetivo de encontrar aleatoriedad en el proceso en la que se busca que el valor de $\boldsymbol{\delta}$ (pendiente de la primera variable) sea 1; si lo anterior se cumple se dice que existe un problema de raíz unitaria y aleatorio, y no aleatorio de lo contrario. Mediante la prueba se obtienen los siguientes resultados en los cuales por medio del valor $\boldsymbol{p}$ se observa que la hipótesis nula de raíz unitaria es rechazada en la mayoría de los casos.

Tabla 10
Resultados prueba unitaria según los p-valores

Producto	Día	t y d		d		n	
	Martes	0.000	R	0.000	R	0.970	A
Banano	Jueves	0.001	R	0.000	R	0.948	A
	Viernes	0.001	R	0.000	R	0.946	A
	Martes	0.000	R	0.000	R	0.236	A
Mandarina	Jueves	0.000	R	0.000	R	0.202	A
	Viernes	0.000	R	0.000	R	0.234	A
	Martes	0.009	R	0.001	R	0.487	A
Naranja	Jueves	0.021	R	0.003	R	0.605	A
	Viernes	0.023	R	0.004	R	0.584	A
	Martes	0.000	R	0.000	R	0.138	A
Papa	Jueves	0.000	R	0.000	R	0.465	A
	Viernes	0.000	R	0.000	R	0.377	A
	Martes	0.001	R	0.000	R	0.227	A
Piña	Jueves	0.001	R	0.000	R	0.247	A
	Viernes	0.001	R	0.000	R	0.228	A
	Martes	0.019	R	0.001	R	0.237	A
Plátano	Jueves	0.015	R	0.000	R	0.194	A
	Viernes	0.032	R	0.001	R	0.175	A
	Martes	0.000	R	0.000	R	0.000	R
Tomate	Jueves	0.000	R	0.000	R	0.000	R
	Viernes	0.000	R	0.000	R	0.000	R
	Martes	0.000	R	0.000	R	0.087	A
Yuca	Jueves	0.000	R	0.000	R	0.104	A
	Viernes	0.000	R	0.000	R	0.148	A

t y d: Proceso con tendencia y con deriva *R: Rechaza la hipótesis nula*
d: Proceso sin tendencia y con deriva *A: Acepta la hipótesis nula*
n: proceso sin tendencia y sin deriva

Luego de aplicada la prueba de raíz unitaria se encuentra que los precios están relacionados y por lo tanto poseen un comportamiento no aleatorio en las series de todos los productos en las raíces con tendencia y deriva y con solo deriva, siendo aceptada la hipótesis en todos los productos a excepción del banano en el análisis de la raíz sin deriva y sin tendencia.

7.2.6 Prueba de Bartels. La prueba de rangos de Bartels se usa para determinar la aleatoriedad de una muestra de datos. Bajo la hipótesis nula de aleatoriedad cualquier disposición de rango de todas las $\boldsymbol{n}$ posibilidades deben ser equiprobables. Los resultados de la prueba de todos los productos son los que se pueden observar a continuación:

Tabla 11
Resultados de la prueba de Bartels para todos los productos

Producto	Día	p valor
Banano	martes	0.000
	jueves	0.000
	viernes	0.000
Mandarina	martes	0.000
	jueves	0.000
	viernes	0.000
	martes	0.000
Naranja	jueves	0.000
	viernes	0.000
	martes	0.000
Papa	jueves	0.000
	viernes	0.000
	martes	0.000
Piña	jueves	0.000
	viernes	0.000
	martes	0.000
Plátano	jueves	0.000
	viernes	0.000
	martes	0.000
Tomate	jueves	0.000
	viernes	0.000
	martes	0.000
Yuca	jueves	0.000
	viernes	0.000

En la Tabla 11 se puede observar que en todos los casos la hipótesis nula de aleatoriedad es rechazada, debido a que el p-valor es menor a 0.05, teniendo en cuenta que la prueba trabaja con un intervalo de confianza del 95% se puede concluir que los residuos de los datos se encuentran correlacionados.

7.2.7 Modelo y validación. Una vez procesadas todas las series temporales de los productos en las pruebas, se obtiene que prácticamente en todos los casos los productos poseen un comportamiento autocorrelacionado o por defecto son no aleatorios. Dado que se comprueba que los datos se encuentran relacionados por medio del error se emplea la metodología Box-Jenkins con el objetivo de encontrar los modelos ARIMA.

En la identificación (primer paso), se busca determinar el modelo $ARIMA\ (p, d, q)$ que mejor ajuste presente en los datos, siendo p el número de rezagos, d el número de veces que se debe diferenciar la serie para convertirla en estacionaria y q el número de errores rezagados de los que depende. El modelo autorregresivo que se obtiene para el precio del banano en el día martes es un $ARIMA\ (3, 1, 1)$; este modelo permite generar un buen ajuste a los datos y no es necesario modificarlo. De ser necesario modificar los modelos, se hace uso de la gráfica de autocorrelación parcial para identificar el número de rezagos significativos.

Una vez que se define el modelo ARIMA preliminar para cada día se continúa con la estimación de los parámetros (segundo paso) del modelo seleccionado; el modelo que se presenta a continuación esta diferenciado en primera diferencia.

Tabla 12
Modelo ARIMA para el día martes

Ar 1	Ar 2	Ar 3	Ma 1
-0.876	-0.390	-0.327	0.651

$$\Delta^1 y_t = -0.876\,\Delta^1 y_{t-1} \; - 0.390\Delta^1 y_{t-2} \; - 0.327 y_{t-3} \; + \; 0.651 a_{t-1} \; + \; e_t \qquad (27)$$

Para verificar/examinar (tercer paso) si el modelo se ajusta a los datos de manera apropiada se genera nuevamente la gráfica de autocorrelación parcial, ya que según Gujarati & Porter (2010),

si los errores no superan los límites de significancia y si la hipótesis nula de aleatoriedad de Ljung-Box es aceptada, se puede determinar al modelo como preciso y los errores de este serian de ruido blanco o aleatorio; de lo contrario, se repite el proceso planteado en la metodología Box Jenkins; se aclara que la determinación de un modelo ARIMA según Gujarati & Porter es un "arte más que una ciencia" y debido a esto el proceso de selección del modelo debe ser preciso.

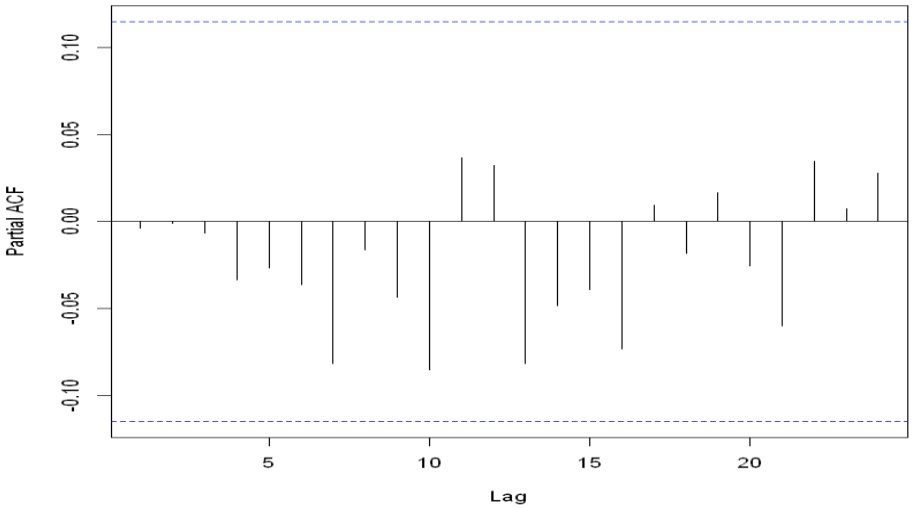

Figura 39: Comportamiento de los residuos del modelo ARIMA ajustado. Adaptado de R versión 3.4.2 (2018

Tabla 13
Prueba Ljung-Box

Producto	p-valor
Banano martes	0.999

A partir de la Figura 39 (gráfica de autocorrelación parcial) y los resultados de la prueba Ljung-Box en la Tabla 13 se observa que el modelo generado para el día martes responde a las características necesarias para establecerse como un modelo preciso, ya que los residuos del modelo se comportan como ruido blanco o independientes en la gráfica y se supera el nivel de significancia de 0.05 en la prueba. Este modelo se aplica a los datos y se obtienen los residuales que serán ingresados al modelo RNA *ff*.

No obstante, este resultado (cumplimiento de las características de un modelo ARIMA preciso) no se presenta en todos los casos y debido a esto se realizan ajustes al modelo ARIMA propuesto por la función teniendo en cuenta los rezagos significativos identificados en las gráficas de autocorrelación parcial; estos rezagos son utilizados con el fin de modificar el valor AR y MA, obteniendo un nuevo modelo que es seleccionado a partir del menor error AIC y que cumpla con las condiciones Box Jenkins (rezagos aleatorios y no significativos). En la Tabla 14 se exponen los modelos propuestos por la función y los ajustados manualmente mediante el uso de los rezagos.

Tabla 14
Modelos ARIMA propuestos y ajustados

Producto	Día	Modelo Propuesto	Modelo Ajustado
Banano	martes	3, 1, 1	3, 1, 1
	jueves	3, 1, 2	3, 1, 1
	viernes	1, 1, 1	1, 1, 6
Mandarina	martes	3, 0, 4	22, 0, 4
	jueves	3, 0, 3	3, 0, 8
	viernes	3, 0, 2	20, 0, 2
Naranja	martes	0, 1, 1	22, 1, 1
	jueves	0, 1, 2	0, 1, 23
	viernes	0, 1, 2	0, 1, 23
Papa	martes	1, 1, 2	14, 1, 2
	jueves	0, 1, 1	14, 1, 1
	viernes	0, 1, 1	10, 1, 1
Piña	martes	0, 1, 3	0, 1, 24
	jueves	2, 1, 3	14, 1, 3
	viernes	2, 1, 3	2, 1, 16
Plátano	martes	0, 1, 1	7, 1, 1
	jueves	0, 1, 1	2, 1, 1
	viernes	0, 1, 1	0, 1, 21
Tomate	martes	1, 1, 0	1, 1, 16
	jueves	3, 1, 3	3, 1, 24
	viernes	1, 1, 1	16, 1, 1
Yuca	martes	3, 1, 1	3, 1, 14
	jueves	1, 1, 2	1, 1, 15
	viernes	2, 1, 3	2, 1, 10

Como se mencionó anteriormente el modelo ajustado se desarrolla a partir del ajuste de los valores AR y MA en el modelo, escogiendo el que mejor AIC otorgue cumpliendo las características de un modelo ARIMA.

En la Tabla 15 se presentan nuevamente los resultados de la prueba Ljung-Box para la validación de los modelos de predicción seleccionados en cada producto. Esta prueba se desarrolla teniendo en cuenta los últimos 25 rezagos de los datos, esto debido a que en ningún modelo se utiliza un rezago superior o igual a 25. Con esta prueba se puede definir a los modelos ARIMA generados como precisos.

Tabla 15
Resultados prueba Ljung-Box

Producto	Día	p-valor
	martes	0.999
Banano	jueves	0.997
	viernes	0.967
	martes	0.999
Mandarina	jueves	0.898
	viernes	0.999
	martes	1
Naranja	jueves	0.999
	viernes	0.993
	martes	0.999
Papa	jueves	0.773
	viernes	0.954
	martes	1
Piña	jueves	0.999
	viernes	0.999
	martes	0.992
Plátano	jueves	0.891
	viernes	1
	martes	0.931
Tomate	jueves	0.999
	viernes	0.999
	martes	1
Yuca	jueves	0.999
	viernes	0.999

p- valor: corresponde al p-valor con el que se rechaza o acepta la hipótesis nula

Una vez definido que los modelos ARIMA seleccionados son precisos se desarrolla la segunda parte del modelo; esta es una Red Neuronal Artificial *feedforward* o de propagación hacia adelante que se genera a partir de los residuos de los modelos ARIMA. En este parte del modelo se procesan los datos y se determina el número de rezagos significativos en la serie según el criterio AIC, y con base en esta red se realiza el pronóstico pertinente.

En la Tabla 16 se pueden observar los errores de pronósticos RMSE y MAPE generados mediante el pronóstico de los dos modelos de predicción (primero se predice con el modelo ARIMA, luego con el modelo RNA *ff* y posteriormente se integran los resultados de estos); también se exhiben los resultados obtenidos mediante el uso de un modelo compuesto por solo la RNA *ff* y el uso de solo el modelo ARIMA.

Tabla 16
Errores de pronostico primer modelo y sus derivaciones

Productos	Día	ARIMA-RNA *ff*		RNA *ff*		ARIMA	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Banano	martes	171.12	8.19	132.98	6.45	176.02	8.43
	jueves	153.01	7.08	111.43	5.29	157.17	7.26
	viernes	118.94	5.58	72.14	3.86	123.60	5.79
Mandarina	martes	414.56	17.38	493.76	20.39	411.58	17.28
	jueves	633.01	26.50	427.62	18.79	616.36	25.65
	viernes	560.01	24.96	578.69	26.25	560.27	24.98
Naranja	martes	301.06	29.63	134.92	12.79	300.65	29.62
	jueves	224.20	21.81	162.97	15.95	224.20	21.81
	viernes	265.00	25.65	118.77	13.36	264.00	25.60
Papa	martes	175.76	12.75	568.82	49.85	175.28	12.62
	jueves	137.71	10.89	649.08	57.21	138.57	10.97
	viernes	211.49	14.72	395.39	32.52	212.73	14.82
Piña	martes	50.18	7.23	48.30	6.40	50.20	7.24
	jueves	44.10	5.76	44.32	5.88	43.98	5.75
	viernes	49.51	6.91	54.96	7.47	49.66	6.93
Plátano	martes	253.95	12.29	309.51	15.28	256.19	12.43
	jueves	264.61	13.49	321.67	17.05	273.73	14.13
	viernes	310.47	16.21	348.07	18.65	315.18	16.53
Tomate	martes	759.81	28.30	768.82	21.01	767.10	27.79
	jueves	732.71	23.41	821.94	26.46	740.14	23.28
	viernes	860.42	29.52	635.74	31.82	871.62	29.59
Yuca	martes	134.09	15.65	109.75	12.90	133.51	15.59
	jueves	126.63	14.98	78.05	9.26	129.97	15.44
	viernes	67.86	8.12	36.46	3.74	67.48	7.87

Con base en los resultados que se exponen en la anterior tabla, se encuentra que no existe una técnica que permita realizar un pronóstico preciso para todos los productos, siendo unos modelos mejores con determinados productos que otros; sin embargo, el modelo RNA *ff* se ajusta a una mayor cantidad de productos para cada día. En la Figura 40 se puede observar el número de producto-día que es explicado con mayor precisión por cada modelo desarrollado.

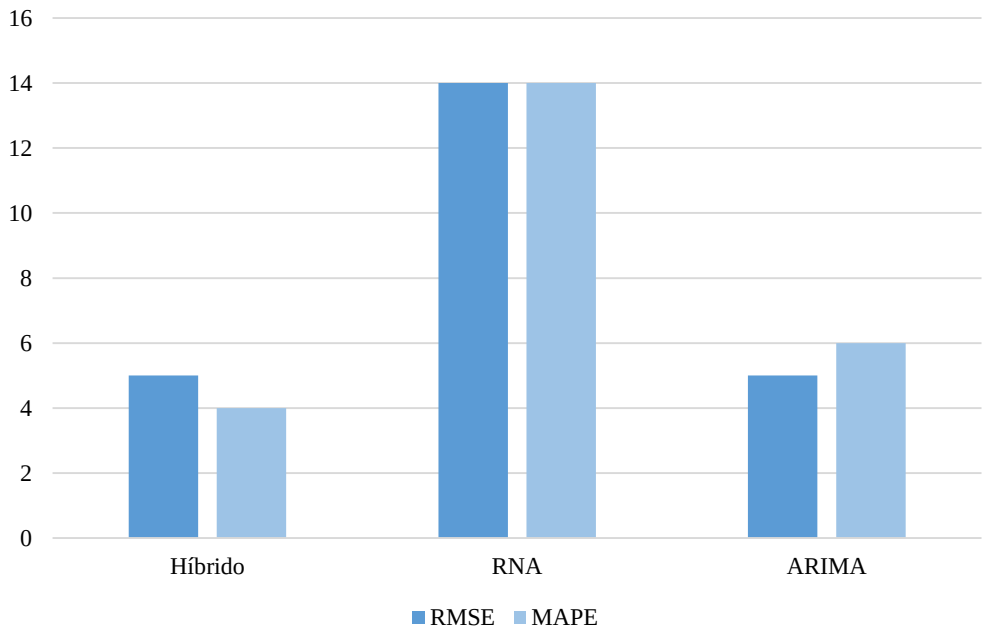

Figura 40: Número de productos ajustados según los modelos desarrollados. Adaptado de R versión 3.4.2 (2018

7.3 Modelo de Red Neuronal Artficial con aprendizaje *bp*

La estructura del modelo se encuentra compuesto por tres capas, la primera constituida por las entradas del modelo, la segunda es una capa con un número n (que varía desde 1 hasta 10) de nodos o neuronas ocultas y un nodo de salida perteneciente al valor pronosticado de la serie (variable respuesta). Este modelo de red neuronal se caracteriza por que la red se ajusta mediante la comparación del error generado a partir del modelo y el valor real del dato lo que se conoce como Red Neuronal Artificial con retropropagación.

7.3.1 Rezagos significativos y tratamiento de la base de datos. Antes de procesar los datos dentro de la Red se debe realizar un ajuste a la base de datos con el fin de identificar cuáles son los rezagos significativos en la serie y a partir de estos generar las variables de entrada de la red neuronal. Para determinar el número de rezagos significativos se hace uso de la gráfica de autocorrelación parcial; estos rezagos son identificados a través de las gráficas autocorrelación parcial del primer modelo.

A continuación, se puede observar la gráfica con los rezagos que son utilizados como entradas para el día martes del banano:

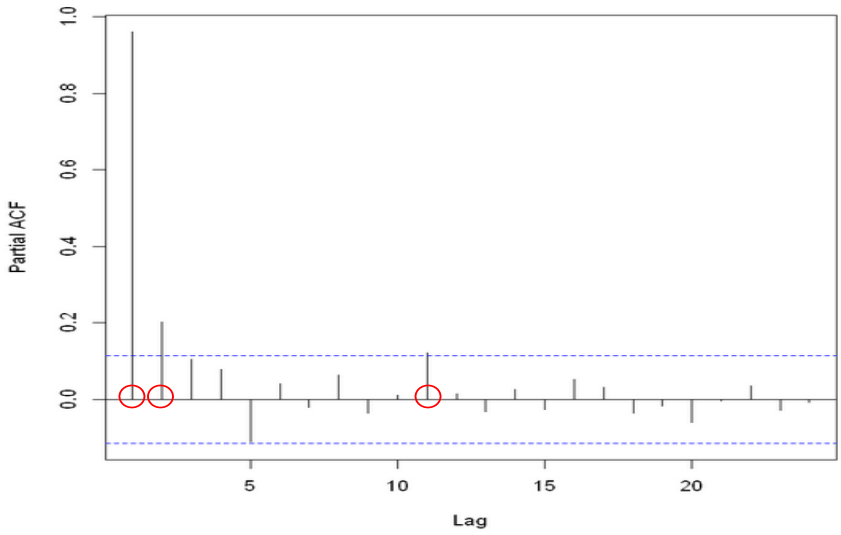

Figura 41: Rezagos significativos en la serie temporal. Adaptado de R versión 3.4.2 (2018)

Se observa que con base en la figura anterior los rezagos que pueden ser considerados como significativos son el rezago uno, dos y once ya que superan los límites de significancia; una vez definidos los rezagos se modifica la base datos original para ser procesada dentro de la red.

A continuación, se observa la base de datos que es obtenida luego de generados los cambios en la base de datos original para los primeros 5 datos a pronosticar del banano en el día martes (la base de datos es ajustada de forma manual). En la Figura 42 se encuentran los valores reales de la

base de datos (Banano martes y - Bamy), la cual se alimenta a si misma con sus valores rezagados 1, 2 y 11.

Bamy	X_1	X_2	X_{11}
889	883	783	717
883	889	883	867
883	883	889	792
850	883	883	733
883	850	883	750

Bamy: Variable a pronosticar o dependiente

Figura 42: Base de datos a ingresar en la RNA *bp*

7.3.2 Preprocesamiento. En la fase de preprocesamiento se busca adaptar la base de datos a ingresar según los requerimientos que exija la red neuronal. Se realiza una normalización [0-1] con el objetivo de que la red responda a ellos con mejor precisión debido principalmente a que la función de activación *sigmoide* se encuentra comprendida en este rango. La ecuación 28 normaliza los datos.

$$Datos\ normalizados = \frac{(Data - valor\ mínimo\ de\ la\ data)}{(valor\ máximo\ de\ la\ data - valor\ \text{mínimo}\ de\ la\ data)} \quad (28)$$

7.3.3 División de datos. Para el tratamiento dentro de la red se deben dividir los datos en dos subconjuntos, el primero consiste en un conjunto de entrenamiento y el segundo es un conjunto de prueba (con el que se compara la capacidad de pronóstico de la red generada con los datos de entrenamiento). El conjunto de entrenamiento consta del 80% del total de los datos (307) y el 20% restante constituye el conjunto de datos de prueba del cual, 15 datos son reservados para la comparación del pronóstico de 15 pasos adelante. En la Figura 43 se ilustra la división de los datos mencionada anteriormente:

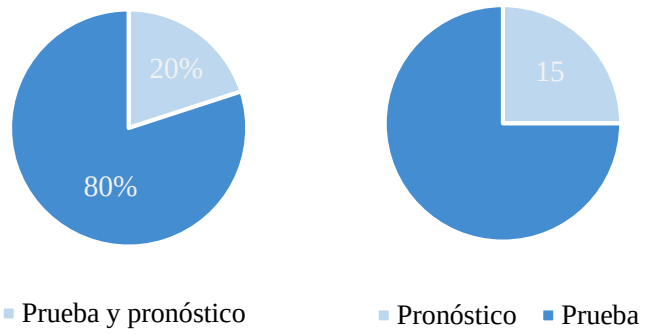

Figura 43: División de los datos para el modelo RNA *bp*

7.3.4 Datos de entrenamiento y prueba. El conjunto de datos de entrenamiento y prueba que se definen para la red neuronal no son los mismos para las series de todos los productos, debido a que los modelos que se desarrollan en los días respectivos depende del número de rezagos significativos presentes en esa serie. El total de datos que se utilizan para el desarrollo de la red en cada producto según las categorías (entrenamiento y prueba) se encuentran en la Tabla 17.

Tabla 17
Datos de entrenamiento y prueba para cada producto

Producto	Día	Entrenamiento	Prueba	Total de datos
Banano	martes	237	59	296
	Jueves	245	61	306
	viernes	245	61	306
Mandarina	martes	227	57	284
	Jueves	239	60	299
	viernes	230	57	287
Naranja	martes	227	57	284
	Jueves	227	57	284
	viernes	234	58	292
Papa	martes	234	59	293
	Jueves	234	58	292
	viernes	238	59	297
Piña	martes	226	57	283
	Jueves	234	59	293
	viernes	233	58	291
Plátano	martes	244	61	305
	Jueves	244	61	305
	viernes	229	57	286
Tomate	martes	233	58	291
	Jueves	226	57	283
	viernes	233	58	291
Yuca	Martes	234	59	293
	Jueves	236	59	295
	Viernes	238	59	297

7.3.5 Desarrollo de la Red Neuronal Artificial *bp*. Para el desarrollo de la red neuronal se deben tener en cuenta los parámetros que exige la red; estos parámetros se encuentran mencionados en el marco teórico y la explicación del uso de cada uno de ellos. En esta fase se generan las múltiples redes con un total de nodos que varía desde 1 hasta 10 posibles neuronas, y un número de variables entradas que disminuye en uno cada vez que son procesadas las 10 neuronas. En la Tabla 18 se puede observar los parámetros que se utilizaron para definir lo modelos:

Tabla 18
Parámetros del modelo RNA bp

Parámetro	Valor
Formula	y = f(número de rezagos significativos)
Datos	Entrenamiento
Nodos en la capa oculta	N [1-10]
Tipo de salida	Linear para definir una salida numérica
Tasa de aprendizaje	0.005
Función de activación	Logística Sigmoide
Número de repeticiones	10
Algoritmo de aprendizaje	Backpropagation
Número máximo de pasos	100.000

7.3.6 Evaluación de la Red Neuronal Artificial *bp*. La validación de la red se lleva a cabo mediante el uso de los datos a pronosticar; estos datos se utilizan con el fin de verificar si los modelos de red generados con los datos de entrenamiento pueden ser generalizados, si los modelos se ajustan con base en los errores de las métricas seleccionadas se puede definir al modelo desarrollado como preciso, de lo contrario, se debe ajustar cada modelo. Según los resultados de los errores de las métricas que se exponen en la Tabla 19, los modelos neuronales según el número de neuronas y variables pueden ser catalogados como precisos. En la Figura 44 se expone una de las redes generadas por la función para los datos del banano del día martes con una red de dos nodos en la capa oculta, tres en la capa de entrada y una en la salida y dos *Bias* (círculos azules) que permiten ajustar los datos mejorando la predicción:

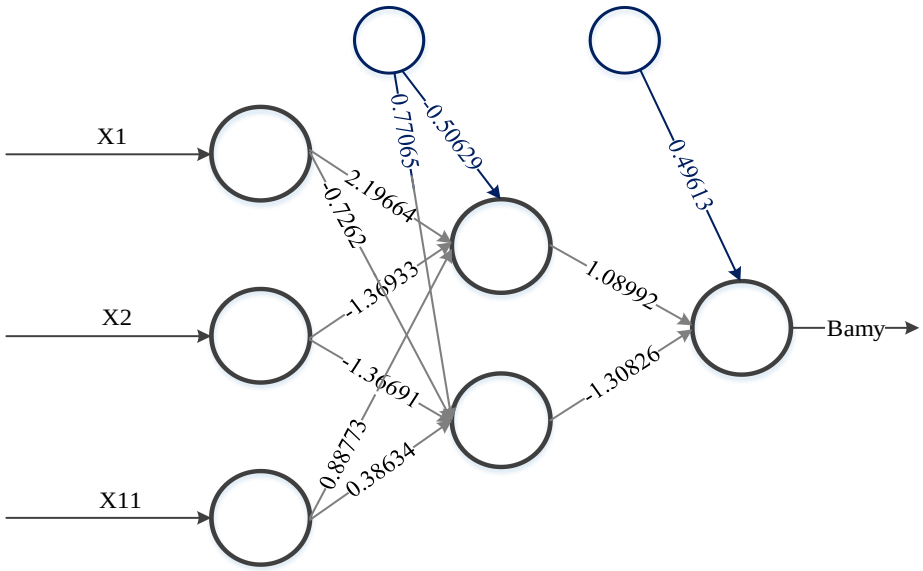

Figura 44: Red Neuronal Artificial *bp* del banano en el día martes. Adaptado de R versión 3.4.2 (2018)

7.3.7 Predicción. En esta última fase se desarrolla la predicción de los precios para los 15 datos del conjunto de prueba. Debido a que el modelo que se desarrolla solo pronostica un valor adelante se generó una función bucle que permite utilizar el dato generado en el periodo *t* (un día adelante) como un rezago en las predicciones posteriores $\boldsymbol{t+m}$ (desde dos días hasta 15 días adelante). La Figura 45 permite observar el procedimiento mencionado con los datos normalizados para el día martes del banano pudiendo observarse la actualización de los datos pronosticados en las variables rezagadas 4 pasos adelante.

Variable a predecir	*X1*	*X2*	*X11*
0.74	0.74	0.74	0.73
0.72	0.74	0.74	0.73
0.69	**0.72**	0.74	0.73
0.70	**0.69**	**0.72**	**0.69**
0.72	**0.70**	**0.69**	0.76

Figura 45: Actualización del pronóstico como variables rezagadas

En la ecuación (29) se puede observar de forma gráfica el comportamiento que se quiere relacionar en la figura anterior:

$$\mathbf{0.70} = \mathbf{0.69} + \mathbf{0.72} + \mathbf{0.69}$$

$$Bamy = X1 + X2 + X11 \tag{29}$$

En las tablas que se relacionan en el Apéndice B se pueden observar los errores obtenidos para cada modelo según el número de neuronas y el número de variables; en esa sección se presentan dos tablas por producto según el día, la primera posee los errores de predicción RMSE, la segunda tabla los errores de predicción MAPE.

En la Tabla 19 se encuentra el resumen de los mejores errores obtenidos por producto según el día teniendo en cuenta el número de neuronas utilizadas y el número de variables de entrada.

Tabla 19
Mejor precisión según el número de neuronas n y N variables

Producto	Día	RMSE	MAPE	n	N variables
Banano	martes	102.29	5.06	6	3
	jueves	145.79	6.69	1	1
	viernes	98.54	4.85	1	1
Mandarina	martes	314.75	17.32	9	6
	jueves	196.81	9.31	4	6
	viernes	419.99	18.24	10	4
Naranja	martes	180.23	18.34	10	4
	jueves	266.07	25.46	5	2
	viernes	167.43	18.52	7	6
Papa	martes	164.78	13.68	7	3
	jueves	114.44	9.13	7	3
	viernes	127.25	9.27	10	1
Piña	martes	53.78	6.73	2	2
	jueves	52.67	7.00	3	2
	viernes	63.85	8.5	2	1
Plátano	martes	252.09	12.06	5	2
	jueves	229.38	11.29	3	2
	viernes	266.21	13.67	7	2
Tomate	martes	623.22	21.92	7	1
	jueves	681.35	22.46	7	4
	viernes	651.97	25.44	6	2
Yuca	martes	56.65	5.91	1	3
	jueves	43.01	4.33	2	2
	viernes	19.89	2.14	10	5

n: número de neuronas en la capa oculta

Se puede concluir que, según los resultados expuestos en la anterior tabla, los mejores errores de predicción en un 62.5%, lo presentan las redes neuronales que tengan un número de neuronas en la capa oculta mayores o igual 5.

7.4 Modelo ANFIS

El modelo ANFIS se encuentra compuesto por 5 capas (excluyendo a la capa de entrada); la primera capa está constituida por los grados de pertenencia donde se fusifican los datos, en la segunda capa se producen las señales de salida a partir de los valores de entrada de la capa, en la tercera capa se normaliza la fuerza de disparo proveniente de la segunda capa, en la cuarta se genera el producto entre la fuerza de disparo normalizado y la combinación lineal de las entradas, siendo la última capa la que calcula la salida o resultado total de sistema.

7.4.1 Rezagos significativos y tratamiento de la base de datos. Antes de procesar los datos dentro del modelo se debe aplicar un ajuste a la base de datos con el objetivo de identificar los rezagos significativos que pueden ser utilizados como variables de entrada del modelo; como los datos que son analizados son los producidos a través de la gráfica de autocorrelación parcial del primer modelo, se opta por mantener los rezagos identificados en el segundo modelo (RNA *bp*).

Debido a los requisitos que exige el modelo, la base datos debe ser organizada de tal forma que la última variable es la que se desea pronosticar y la anterior a esta es la primera variable rezagada identificada. A continuación, se expone el resultado de la modificación de la base de datos para los primeros 5 datos a pronosticar del banano en el día martes. En la Figura 46 se encuentran los valores reales de la base de datos (Banano martes y - Bamy), la cual se alimenta a si misma con sus valores rezagados 1, 2 y 11.

X_{11}	X_2	X_1	Bamy
717	783	883	889
867	883	889	883
792	889	883	883
733	883	883	850
750	883	850	883

Bamy: Variable a pronosticar o dependiente

Figura 46: Base de datos a ingresar al modelo ANFIS

7.4.2 División de datos. Para el procesamiento dentro del modelo ANFIS se desarrolla la misma división de datos que la realizada en el modelo de red neuronal, es decir dos subconjuntos (entrenamiento y prueba). El conjunto de entrenamiento consta del 80% del total de los datos (307) y el 20% restante constituye el conjunto de datos de prueba del cual, 15 datos son reservados para la comparación del pronóstico de 15 pasos adelante. En la Figura 43 de la sección del modelo neuronal se puede observar la división de datos mencionada.

7.4.3 Desarrollo del modelo ANFIS. Para el desarrollo del modelo ANFIS se debe partir del uso de los parámetros que son exigidos por este, los cuales son mencionados y explicados en el marco teórico. En la Tabla 20 se pueden observar los parámetros que son utilizados para definir los modelos:

Tabla 20
Parámetros del modelo ANFIS

Parámetro	Valor
Rango	Valor mínimo y máximo en los que se debe generar la predicción
Datos	Entrenamiento
Tamaño del paso	0.01
Tasa de aprendizaje	0.005
Función de pertenencia	Gaussiana
Número de conjuntos difusos	3 (bajo, medio, alto)
Intersección difusa	MIN
Unión difusa	MAX
Función de implicación	Zadeh
Método	ANFIS

7.4.4 Evaluación del modelo ANFIS. La validación del modelo se lleva a cabo mediante el uso de los datos a pronosticar con el objetivo de definir si los modelos pueden ser generalizados a los datos de prueba; esto se realiza a partir de los resultados de las métricas de error de pronóstico. A continuación, se exponen las gráficas de pertenencia del banano del martes con base en el modelo ANFIS:

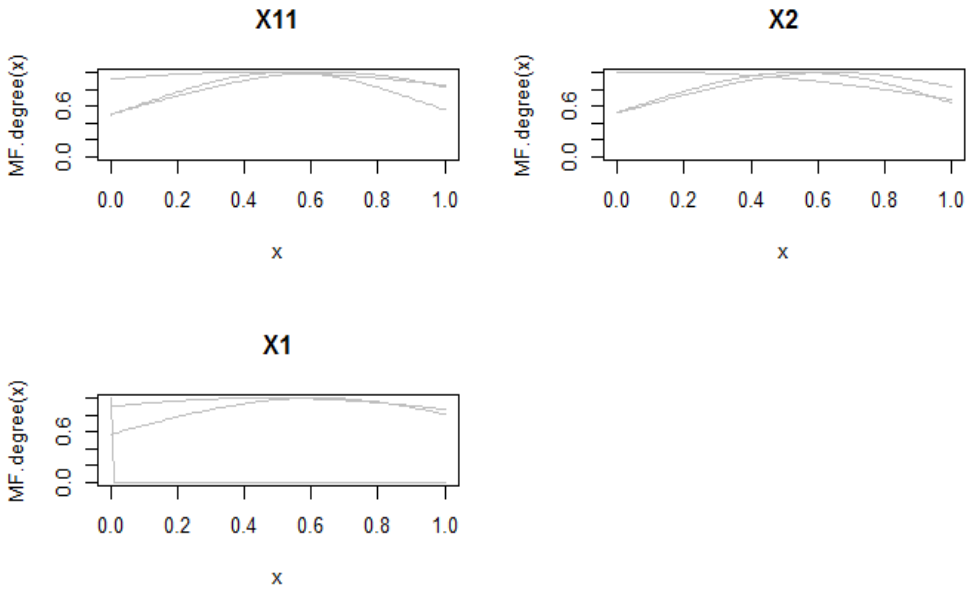

Figura 47: Graficas de la función de pertenencia ANFIS. Adaptado de R versión 3.4.2 (2018)

Se puede observar que los conjuntos difusos no poseen una distribución típica según la forma de la función de pertenencia; esto se obtiene debido a que la función (gaussiana) que es utilizada por el modelo posee dos parámetros, la media y la varianza para cada conjunto difuso (alto, medio y bajo), los cuales son calculados a partir de los datos utilizados por cada variable; sin embargo, esto no demuestra que el modelo que se desarrolla no sea apto, ya que se realiza el mismo procesamiento de los datos con el modelo Wendel Mamdani (WM) con el objetivo de contrastar los resultados de la función de pertenencia. La grafica de pertenencia obtenida por el modelo WM se puede observar en la Figura 48.

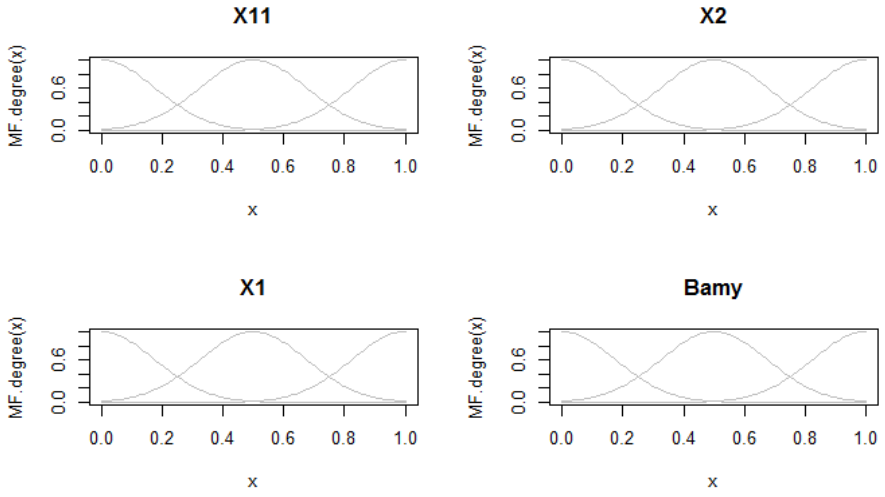

Figura 48: Graficas de la función de pertenencia WM. Adaptado de R versión 3.4.2 (2018)

En la anterior figura aunque se puede observar una mejor distribución con relación a los conjuntos difusos, esta no garantiza que el valor de pronóstico que genere sea el mejor, ya que a diferencia del modelo ANFIS, el modelo WM define una media y varianza estáticas para todos los valores de las variables de entrada lo que puede disminuir su capacidad de predicción, además, el modelo WM no es recomendable utilizarlo para ejercicios de predicción, ya que incrementan los errores de pronóstico (Mendoza & Mazo, 2009). En la Figura 49 y la Figura 50 las etiquetas superiores representan los conjuntos borrosos para las tres variables de entrada (pequeña, mediana y grande o *small, medium* y *large*), [1,] representa el tipo de función utilizada (gaussiana), [2,] representa la media y [3,] la varianza.

```
          small     medium      large      small     medium      large          small
[1,] 5.0000000 5.0000000 5.0000000 5.0000000 5.0000000 5.0000000 5.000000e+00
[2,] 0.4194846 0.5233705 0.6545069 0.1238659 0.5460437 0.6490952 2.606327e-91
[3,] 0.9998901 0.4424604 0.5622241 0.9998746 0.4841841 0.5672478 1.000000e-03
[4,]        NA        NA        NA        NA        NA        NA           NA
[5,]        NA        NA        NA        NA        NA        NA           NA
         medium     large
[1,] 5.0000000 5.0000000
[2,] 0.4671374 0.6177056
[3,] 0.9996445 0.5846875
[4,]        NA        NA
[5,]        NA        NA
```

Figura 49: Media y varianza ANFIS. Adaptado de R versión 3.4.2 (2018)

```
     small medium large small medium large small medium large
[1,] 5.000  5.000 5.000 5.000  5.000 5.000 5.000  5.000 5.000
[2,] 0.000  0.500 1.000 0.000  0.500 1.000 0.000  0.500 1.000
[3,] 0.175  0.175 0.175 0.175  0.175 0.175 0.175  0.175 0.175
[4,]    NA     NA    NA    NA     NA    NA    NA     NA    NA
[5,]    NA     NA    NA    NA     NA    NA    NA     NA    NA
```

Figura 50: Media y varianza WM. Adaptado de R versión 3.4.2 (2018)

Con base en estos resultados la función genera las reglas de decisión por modelo, donde el modelo WM produce una etiqueta (bajo, medio o alto) que debe ser defusificada, el modelo ANFIS genera una serie de funciones según el cumplimiento de la regla, donde $var1$, $var2$ y $var3$ representan los valores que multiplican a los rezagos que son utilizados como variables de entrada y $const$ el valor constante de la función. La Figura 51 y la Figura 52 exponen los resultados de las reglas generadas por cada modelo.

```
  V1  V2 V3     V4  V5 V6 V7     V8  V9 V10 V11    V12  V13
1 IF X11 is medium and X2 is medium and  X1  is medium THEN
2 IF X11 is  small and X2 is medium and  X1  is medium THEN
3 IF X11 is  small and X2 is  small and  X1  is  small THEN
4 IF X11 is  small and X2 is  small and  X1  is medium THEN
5 IF X11 is  large and X2 is  large and  X1  is  large THEN
6 IF X11 is medium and X2 is  large and  X1  is  large THEN
7 IF X11 is  large and X2 is  large and  X1  is medium THEN
8 IF X11 is  large and X2 is medium and  X1  is  large THEN
9 IF X11 is medium and X2 is medium and  X1  is  large THEN
The linear equations on consequent parts of fuzzy IF-THEN rules:
            var.1      var.2       var.3         const
 [1,]  0.005972654 0.67127227 0.173827985 -0.1250444125
 [2,]  0.253479836 0.22244776 0.332945416  0.0006603167
 [3,]  0.252928824 0.38289626 0.659018884  0.7882117592
 [4,]  0.165596610 0.06276663 0.001953828 -0.3841105211
 [5,]  0.109871776 0.59863366 0.791655591  0.5433784609
 [6,]  0.520979923 0.17677083 0.545033951 -0.1226286806
 [7,]  0.580683181 0.48094403 0.405502645 -0.2975403405
 [8,]  0.733406664 0.25931025 0.214798506  0.6659881874
 [9,] -0.050710714 0.04655537 0.062115060  0.2321603962
```

Figura 51: Reglas generadas por el modelo ANFIS. Adaptado de R versión 3.4.2 (2018)

```
   V1  V2 V3     V4  V5 V6 V7     V8  V9 V10 V11    V12  V13  V14 V15
1  IF X11 is medium and X2 is medium and  X1  is medium THEN Bamy  is
2  IF X11 is  small and X2 is medium and  X1  is medium THEN Bamy  is
3  IF X11 is  small and X2 is  small and  X1  is  small THEN Bamy  is
4  IF X11 is  small and X2 is  small and  X1  is  small THEN Bamy  is
5  IF X11 is  small and X2 is  small and  X1  is medium THEN Bamy  is
6  IF X11 is  large and X2 is  large and  X1  is  large THEN Bamy  is
7  IF X11 is  large and X2 is  large and  X1  is  large THEN Bamy  is
8  IF X11 is medium and X2 is  large and  X1  is  large THEN Bamy  is
9  IF X11 is medium and X2 is medium and  X1  is medium THEN Bamy  is
10 IF X11 is  large and X2 is  large and  X1  is medium THEN Bamy  is
11 IF X11 is  large and X2 is medium and  X1  is  large THEN Bamy  is
12 IF X11 is medium and X2 is medium and  X1  is  large THEN Bamy  is
      V16
1  medium
2  medium
3   small
4  medium
5  medium
6   large
7  medium
8   large
9   large
10  large
11  large
12  large
```

Figura 52: Reglas generadas por el modelo WM. Adaptado de R versión 3.4.2 (2018)

Teniendo en cuenta lo mencionado anteriormente, se producen los errores según las métricas con lo que se obtiene los resultados que se pueden observar en la Tabla 21.

Tabla 21
Errores del ANFIS y WM para el Banano comercializado el día martes

Modelo	RMSE	MAPE
ANFIS	97.17	4.51
WM	123	6.21

7.4.5 Predicción. En esta última fase se desarrolla la predicción de los precios para los 15 datos del conjunto de prueba. Como en el caso del modelo RNA *bp* la función sólo predice un paso adelante, se genera una función bucle que permite utilizar el dato generado en el periodo *t* (un día adelante) como un rezago en las predicciones posteriores $\boldsymbol{t+m}$ (desde dos días hasta 15 días adelante). La Figura 53 permite observar el procedimiento mencionado con el banano del día martes pudiendo observarse la actualización de los datos pronosticados en las variables rezagadas 4 pasos adelante.

X11	*X2*	*X1*	Variable a predecir
1595.86	1715.81	1718.86	**1706.78**
1626.74	1718.86	**1706.78**	**1709.42**
1645.30	**1706.78**	**1709.42**	**1712.52**
1645.30	**1709.42**	**1712.52**	1714.35
1637.82	**1712.52**	1714.35	1713.79

Figura 53: Actualización de pronóstico como variables rezagadas

En la ecuación (30) se puede observar de forma gráfica el comportamiento que se quiere relacionar en la tabla anterior:

$$\mathbf{1712.52} = \mathbf{1709.42} + \mathbf{1706.78} + \mathbf{1645.30}$$
$$Bamy = X1 + X2 + X11 \quad (30)$$

En la Tabla *22* se exponen los resultados de error según el número rezagos usados. Esto se realiza debido a que el modelo con determinados productos no genera resultados de predicción, sino que reemplaza el valor mayor definido en el rango, por el que se debe pronosticar según la cantidad de rezagos que son utilizados (por ejemplo, el número de rezagos 5, 6, 7 y 8 utilizados en la mandarina). Los números que aparecen en la parte superior de la tabla hacen referencia al número de rezagos y los números resaltados en azul representan el mejor valor de predicción de los modelos.

Tabla 22
Predicción según el número de variables utilizadas en el modelo ANFIS

Producto	Día		2	3	4	5	6	7	8
Banano	martes	R	263.45	96.08					
		M	13.47	4.34					
	jueves	R	148.98						
		M	7.01						
	viernes	R	81.96						
		M	4.28						
Mandarina	martes	R	857.30	389.16	936.55	1227.33	1227.33	1227.33	1227.33
		M	35.98	16.72	60.40	81.67	81.67	81.67	81.67
	jueves	R	917.53	938.14	436.95	978.34	1190.83		
		M	38.91	42.15	19.43	62.31	76.92		
	viernes	R	556.57	351.30	770.95	1257.68			
		M	29.01	19.51	48.23	82.97			
Naranja	martes	R	343.87	220.92	173.91	419.87			
		M	33.67	21.66	21.19	58.29			
	jueves	R	446.29	125.14					
		M	15.99	49.31					
	viernes	R	260.67	171.64	308.16	428.98	431.35		
		M	25.61	15.71	42.63	59.65	59.96		
Papa	martes	R	144.56	805.39	1089.55	1089.55	1089.55		
		M	10.12	70.78	98.66	98.66	98.66		
	jueves	R	145.48	794.87	1092.24	1092.24	1092.24		
		M	11.78	69.38	99.30	99.30	99.30		
	viernes	R	117.83	815.23					
		M	8.81	73.30					
Piña	martes	R	67.43	140.22	233.37	233.37			
		M	8.63	20.83	36.86	36.86			
	jueves	R	69.13	106.66					
		M	9.63	15.89					
	viernes	R	68.22	207.17	265.52				
		M	9.11	32.34	42.92				

Continuación de la Tabla 22:

Producto	Día		2	3	4	5	6	7	8
Plátano	martes	R	239.31						
		M	11.49						
	jueves	R	141.47						
		M	6.62						
	viernes	R	202.38	467.07					
		M	9.79	23.66					
Tomate	martes	R	660.24	932.07					
		M	17.88	48.77					
	jueves	R	742.16	1184.13	1606.91	1606.91			
		M	24.17	60.58	87.36	87.36			
	viernes	R	1579.91	813.64	1549.62	1549.62	1549.62		
		M	66.88	43.58	89.57	89.57	89.57		
Yuca	martes	R	210.21	207.36	227.41				
		M	24.49	24.11	26.81				
	jueves	R	254.94	681.98	817.62	827.07			
		M	32.12	83.48	103.54	104.87			
	viernes	R	224.94	688.99	909.403	932.93	932.93		
		M	28.68	84.21	115.80	119.22	119.22		

Nota. R y M: Errores RMSE y MAPE respectivamente

Se puede observar con base en los resultados de la anterior tabla, que los mejores resultados de error para los modelos se obtienen cuando son utilizados entre 2 y 3 rezagos como entradas, a excepción de la mandarina del día jueves y la naranja del día martes que se obtienen con cuatro rezagos de entrada (no se exponen los resultados con un rezago ya que el modelo no acepta una sola entrada).

8. Resultados finales y comparación

En la presente sección del documento se expondrá el resumen de los resultados de los modelos de pronóstico para todos los productos con base en las dos métricas evaluadas; estos resultados se encuentran registrados en la Tabla 23.

Tabla 23
Resumen de errores de predicción de los tres modelos propuestos en el proyecto

Productos	Día	ARIMA-RNA *ff*		RNA *bp*		ANFIS	
		RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
Banano	martes	171.12	8.19	102.29	5.06	96.08	4.34
	jueves	153.01	7.08	145.79	6.69	148.98	7.01
	viernes	118.94	5.58	98.54	4.85	81.96	4.28
Mandarina	martes	414.56	17.38	314.75	17.32	389.16	16.72
	jueves	633.01	26.50	196.81	9.31	436.95	19.43
	viernes	560.01	24.96	419.99	18.24	351.30	19.51
Naranja	martes	301.06	29.63	180.23	18.34	173.91	21.19
	jueves	224.20	21.81	266.07	25.46	125.14	49.31
	viernes	265.00	25.65	167.43	18.52	171.64	15.71
Papa	martes	175.76	12.75	164.78	13.68	144.56	10.12
	jueves	137.71	10.89	114.44	9.13	145.48	11.78
	viernes	211.49	14.72	127.25	9.27	117.83	8.81
Piña	martes	50.18	7.23	53.78	6.73	67.43	8.63
	jueves	44.10	5.76	52.67	7.00	69.13	9.63
	viernes	49.51	6.91	63.85	8.5	68.22	9.11
Plátano	martes	253.95	12.29	252.09	12.06	239.31	11.49
	jueves	264.61	13.49	229.38	11.29	141.47	6.62
	viernes	310.47	16.21	266,21	13.67	202.38	9.79
Tomate	martes	759.81	28.30	623.22	21.92	660.24	17.88
	jueves	732.71	23.41	681.35	22.46	742.16	24.17
	viernes	860.42	29.52	651.97	25.44	813.64	43.58
Yuca	martes	134.09	15.65	56.65	5.91	207.36	24.11
	jueves	126.63	14.98	43.01	4.33	254.94	32.12
	viernes	67.86	8.12	19.89	2.14	224.94	28.68

Con base en estos resultados se desarrolla un análisis de varianza (ANOVA) en el software estadístico Minitab 17 con el objetivo de determinar si los resultados de predicción se ven afectados por los factores: modelo, producto o día. A través del análisis se encuentra que los resultados no se ven afectados ni por el modelo ni por el producto, encontrando de esta forma que el día es el único factor que afecta a los resultados de error. A continuación se presentan los

resultados obtenidos del procesamiento de los errores en el ANOVA (Figura 54 y Figura 55); en estos resultados se puede observar que en el caso de los factores producto y modelo para los dos errores, el p valor no supera el límite de significancia 0.05, con lo que se obtiene que las medias son iguales (se acepta la hipótesis nula), siendo día el único factor que influye según el análisis de varianzas en los resultados de error (p valor = 0.78 para el RMSE y p valor = 0.71 para el MAPE).

Información del factor

Factor	Tipo	Niveles	Valores
Productos	Fijo	8	Banano; Mandarina; Naranja; Papa; Piña; Plátano; Tomate; Yuca
Día	Fijo	3	jueves; martes; viernes
Modelo	Fijo	3	ANFIS; ARIMA-RNA; RNA-bp

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Productos	7	2990447	427207	107,95	0,000
Día	2	1953	977	0,25	0,782
Modelo	2	62379	31189	7,88	0,001
Error	60	237455	3958		
Total	71	3292234			

Resumen del modelo

S	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
62,9093	92,79%	91,47%	89,61%

Figura 54: Resultados ANOVA para el error RMSE. Adaptado de Minitab 17.

Información del factor

Factor	Tipo	Niveles	Valores
Productos	Fijo	8	Banano; Mandarina; Naranja; Papa; Piña; Plátano; Tomate; Yuca
Día	Fijo	3	jueves; martes; viernes
Modelo	Fijo	3	ANFIS; ARIMA-RNA; RNA-bp

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Productos	7	3625,23	517,89	14,13	0,000
Día	2	24,33	12,16	0,33	0,719
Modelo	2	311,01	155,51	4,24	0,019
Error	60	2198,44	36,64		
Total	71	6159,01			

Resumen del modelo

S	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
6,05315	64,31%	57,76%	48,60%

Figura 55: Resultados ANOVA para el error MAPE. Adaptado de Minitab 17.

A través de los diagramas de bloques en las Figura 56 y Figura 57 (los cuales están clasificados según el día de venta), se observa que los productos presentan diferentes variabilidades siendo el martes el día de mayor variabilidad, seguido del día jueves y por último el día viernes lo que puede afectar el resultado de los modelos o que en consecuencia sean diferentes.

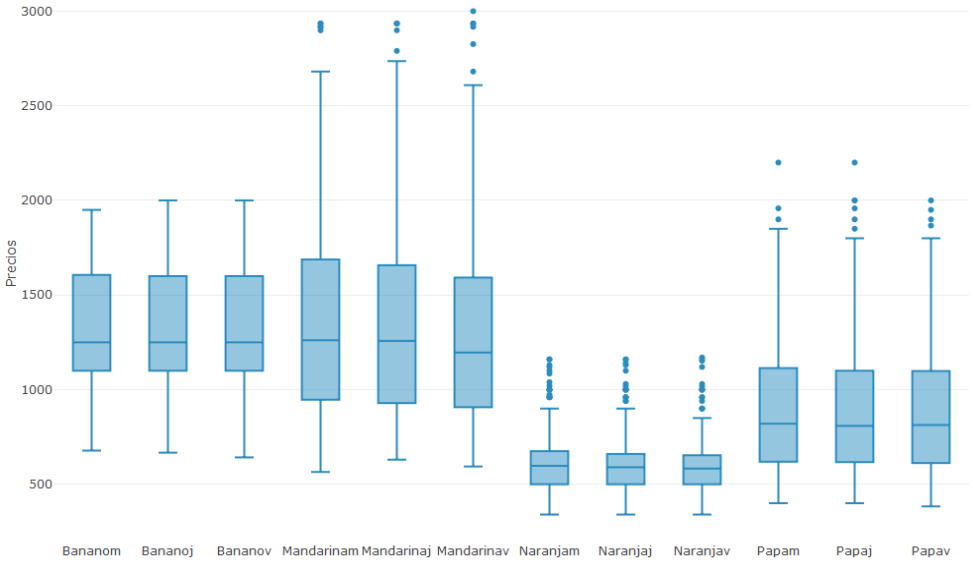

Figura 56: Diagrama de bloques según el día de los primeros 4 productos. Adaptado de R versión 3.4.2 (2018)

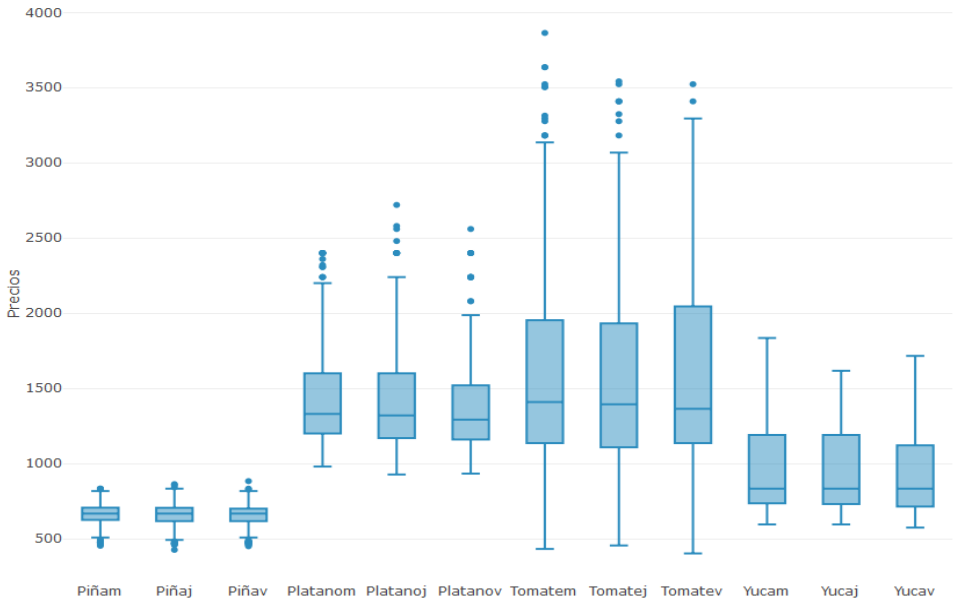

Figura 57: Diagrama de bloques según el día de los últimos 4 productos. Adaptado de R versión 3.4.2 (2018)

Ahora bien, como el factor día no es una variable de diseño (a diferencia de los modelos aplicados y sus variantes analizadas) se sugiere determinar el mejor modelo a partir del tiempo de computo requerido para obtener el resultado final. En la Tabla 24 se presenta el tiempo de procesamiento para el entrenamiento del modelo y la posterior ejecución del pronóstico. El desarrollo de los tres modelos se llevó a cabo en un computador HP Windows 8.1 Pro con un procesador Intel ® Core ™ i7-5500U CPU @ 2.40GHZ, 2.40 GHZ, memoria RAM de 8.00 GB y un sistema operativo de 64 bits, procesador x64.

Tabla 24
Tiempo de procesamiento por modelo en minutos

		ARIMA-RNA *ff*			RNA *bp*			ANFIS		
Producto	Día	User	system	Elapsed	User	System	elapsed	User	system	elapsed
	m	0.02	0.00	0.02	0.53	0.00	0.53	0.18	0.00	0.19
Banano	j	0.03	0.00	0.03	0.52	0.00	0.59	0.07	0.00	0.07
	v	0.03	0.00	0.03	0.45	0.00	0.54	0.07	0.00	0.07
	m	0.12	0.01	0.13	98.15	0.24	98.86	4.85	0.00	4.85
Mandarina	j	0.04	0.00	0.04	106.57	0.18	109.62	1.81	0.00	1.81
	v	0.03	0.00	0.04	55.97	0.17	57.45	1.05	0.00	1.06
	m	0.13	0.01	0.14	35.91	0.17	43.70	1.00	0.00	1.00
Naranja	j	0.13	0.00	0.13	21.44	0.13	22.04	0.23	0.00	0.23
	v	0.06	0.00	0.06	61.43	0.27	64.34	2.00	0.00	2.01
	m	0.04	0.00	0.05	32.02	0.22	32.74	1.81	0.00	1.82
Papa	j	0.05	0.00	0.05	38.35	0.23	39.81	1.84	0.00	1.85
	v	0.03	0.00	0.03	18.69	0.10	19.33	0.24	0.00	0.24
	m	0.03	0.00	0.03	12.02	0.05	12.50	1.08	0.00	1.09
Piña	j	0.04	0.00	0.04	15.75	0.07	16.32	0.24	0.00	0.24
	v	0.03	0.00	0.03	15.19	0.07	15.51	0.63	0.00	0.63
	m	0.18	0.01	0.19	33.55	0.04	33.66	0.08	0.00	0.08
Plátano	j	0.04	0.00	0.05	14.27	0.06	14.51	0.08	0.00	0.09
	v	0.05	0.00	0.05	23.90	0.10	24.63	0.26	0.00	0.26
	m	0.23	0.01	0.25	44.95	0.63	47.41	0.27	0.00	0.27
Tomate	j	0.06	0.00	0.07	56.14	0.18	60.82	1.57	0.00	1.57
	v	0.05	0.00	0.06	54.98	0.18	56.67	3.02	0.00	3.02
	m	0.05	0.00	0.05	51.64	0.17	55.82	0.56	0.00	0.56
Yuca	j	0.05	0.00	0.06	55.06	0.20	55.61	1.14	0.00	1.14
	v	0.04	0.00	0.04	59.22	0.25	63.50	1.63	0.00	1.63

User es el tiempo de la CPU dedicado a ejecutar las instrucciones del proceso iniciado, *system* es el tiempo de la CPU empleado por el sistema operativo en el que se siguen las instrucciones del

proceso (abrir ficheros, iniciar otros procesos o mirar al reloj del sistema, etc.) y *elapsed* es el tiempo total del proceso desde que se inicia hasta que se termina.

Se puede observar que, con base en los resultados obtenidos anteriormente, el modelo que presenta menor tiempo de procesamiento para cada producto y su respectivo pronóstico es el ARIMA-RNA *ff*, seguido del modelo ANFIS y encontrando por último al modelo RNA *bp*.

9. Difusión académica

Para la difusión de los resultados del proyecto se realiza un artículo (Apéndice C) que describe algunos referentes teóricos de los modelos desarrollados, las respectivas metodologías de cada modelo, la síntesis de los resultados obtenidos, la discusión y conclusiones derivadas de la investigación y el tema del artículo, el cual es el análisis y comparación del rendimiento de los tres modelos basados en los modelos RNA para la predicción de los 8 productos agrícolas seleccionados, los cuales se distribuyen entre los tres los días martes, jueves y viernes. El artículo se someterá a la revista Neurocomputing la cual es seleccionada a través de una consulta en el listado de revistas categorizadas por Scimago como Q1 y Q2, enfocadas en temas de ciencias agrícolas, economía, econometría, finanzas e inteligencia artificial.

Se realiza una ponencia en modalidad poster (Apéndice D) con resultados parciales del proyecto en el día del investigador desarrollado el día 31 de octubre del 2018 los cuales abarcan una revisión de literatura con el objetivo de identificar las principales técnicas de pronósticos de precios agrícolas y observar cómo se encuentra la investigación en este campo, identificar una metodología

de desarrollo de la investigación y seleccionar los principales modelos utilizados en la predicción de precios del sector agropecuario.

Se realiza una ponencia de resultados parciales en el "XII Encuentro de Semilleros de Investigación UNAB 2018 El Intercambio del Conocimiento: Base para el desarrollo" (Apéndice E) evento que se realiza el día 11 de octubre de 2018 en el que se presentan los resultados obtenidos a través del desarrollo de los modelos ARIMA, RNA *ff* y el híbrido ARIMA-RNA *ff*.

Se genera un repositorio en R *markdown* el cual contiene la metodología y codificación del análisis de los datos, limpieza, imputación, pruebas de hipótesis, el desarrollo de todos los modelos abarcados en la investigación y las bases de datos obtenidas después del procesamiento; este repositorio se encuentra en el siguiente enlace de github: https://github.com/JuanDM0106/Tesis.

10. Conclusiones

Durante la revisión de literatura se encuentra que la actividad de predicción a nivel de investigación está relacionada con el comportamiento de diferentes fenómenos en áreas como la ciencia, medicina, economía, ingeniería, sociología y agricultura; donde la predicción de precios en el sector agro se encuentra liderada a nivel mundial por países como China, India, Estados Unidos y Brasil, lo que puede deberse a la importancia que poseen para estas naciones este sector. De los métodos de pronósticos aplicados se encuentra que en los últimos años las técnicas basadas en aprendizaje automático o *machine learning* han tenido un amplio uso, lo cual se fundamenta en su aprendizaje interactivo y en la capacidad que estas poseen para procesar mayores volúmenes de datos sin perder precisión en los resultados.

En cuanto a Colombia, desde el año 2017 el DANE se ha estado desarrollando un proyecto que plantea el seguimiento de los precios de la actividad agrícola mediante un monitoreo de los precios de venta de productos comercializados en las principales centrales de abastos de la nación y generar con ello un acompañamiento a la actividad económica; es importante resaltar que las técnicas aplicadas durante el proyecto del DANE se basan en modelos autorregresivos. Ahora bien, a partir de proyectos como el mencionado anteriormente se puede determinar el interés actual de la nación en el estudio, desarrollo y constante mejora de esta actividad en dicho sector, ya que en Colombia todos los estudios realizados con este enfoque y/o con el uso de técnicas aprendizaje automático están orientados principalmente al comportamiento de precios de productos o bienes que cotizan en la bolsa de valores, acciones y la energía. En consecuencia, el presente proyecto sigue las líneas de interés de la nación en cuanto a la aplicación de modelos de pronósticos de precios en el sector agrícola con la implementación de un procesamiento de datos más especializado en conjunto con el uso de técnicas de aprendizaje automático. Estas últimas han sido seleccionadas ya que presentan mejor precisión que los modelos autoregresivos cuando las series históricas son no lineales. En adición, entre las técnicas más aplicadas se encontraron las redes neuronales artificiales, las máquinas de soporte vectorial y los árboles de decisión, siendo las RNA el grupo de técnicas seleccionadas para la presente investigación debido a su alta aplicación en el campo de la predicción.

Durante el desarrollo metodológico fue realizado un análisis de correlación entre los precios de los productos bajo estudio. A partir de dicho análisis se concluye que la mayoría de los productos poseen una baja relación lineal entre ellos (correlación de sus series de preciso históricas) presentando únicamente una relación alta los productos cítricos (naranja y mandarina) los cuales presentan una correlación de precios superior al 70%); no obstante, es posible que existan

relaciones no lineales, por tanto, para futuros trabajos es necesario desarrollar estudios más profundos como, análisis de causalidades, cointegraciones, minería de datos, diseños de experimentos, entre otros y, de esta forma, se podría considerar como variables exógenas el precio histórico de otros productos.

Por otra parte, durante la aplicación del primer modelo propuesto (ARIMA-RNA *ff*) se encontraron diferencias respecto a lo expuesto en la literatura ya que al comparar el desempeño del ARIMA-RNA *ff* con sus modelos base (RNA *ff* y ARIMA por separado) se concluye que, aunque el híbrido es una alternativa recomendada por Mitra & Paul (2017) durante la presente investigación se encontraron mejores resultados utilizando el modelo RNA *ff*, el cual presenta un mejor desempeño en 14 de 24 productos (8 productos distribuidos entre los días martes, jueves y viernes), el modelo híbrido por su parte sólo en 5 productos según el RMSE y en 4 según el MAPE, y el modelo ARIMA sólo en 5 productos según el RMSE y en 6 según el MAPE. Sin embargo, aunque el modelo RNA *ff* es el que mejor ajuste presenta de los tres modelos, no existe un modelo ideal para todos los productos y, por tanto, es necesario continuar explorando alternativas de técnicas y herramientas.

Dicho esto, al comparar los tres modelos basados en redes neuronales aplicados durante la presente investigación (ARIMA-RNA *ff*, RNA *bp* y ANFIS), se encuentra que mediante la metodología usada por Mitra & Paul (2017) y Mishra & Singh (2013) los mejores modelos de predicción son el RNA *bp* con 11 productos y el ANFIS con 10 productos respectivamente; donde el menor valor de error del modelo RNA *bp* puede deberse a la estructura de su funcionamiento en la cual se generan pesos de forma aleatoria, los que se ajustan continuamente con el ingreso de cada nuevo dato buscando reducir gradualmente la diferencia entre el valor de salida de los precios;

mientras que el menor valor de error del modelo ANFIS es consecuencia de la fusificación de datos y la definición de reglas según la categoría a la cual pertenece el dato ingresado, esto permite generar pesos de manera no aleatoria y obtener una salida próxima al valor real de los datos. Se puede enunciar que los dos modelos presentan mejores resultados que la metodología ARIMA-RNA *ff* haciendo uso de las métricas de comparación halladas en la literatura.

Sin embargo, con el propósito de determinar qué tan significativos son las diferencias entre los errores obtenidos por las tres variantes (ARIMA-RNA *ff*, RNA *bp*, ANFIS), se realiza un análisis de varianza donde se observa el efecto que poseen los factores: modelo, producto y día en los resultados, con el que se concluye que el único factor que influye es el día (es decir los efectos de los factores modelo y producto no son significativos en los resultados de error), siendo el día una variable no utilizada en el diseño del modelo. Por lo tanto, para la selección del modelo de mejor pronóstico se hace uso del tiempo de computo (tiempo empleado en el entrenamiento y pronóstico de los modelos) con el que se obtiene que el mejor modelo es el híbrido (ARIMA-RNA *ff*), ya que presenta los menores tiempos de computo, seguido del modelo ANFIS y en último lugar el modelo RNA *bp*.

Finalmente, al analizar los errores de las métricas obtenidas por los modelos se encuentra que la mandarina y el tomate presentan variabilidades altas en comparación con productos como la piña y el banano, dando como resultado un error de pronóstico mayor, lo que se fundamenta en la dificultad del modelo de replicar el comportamiento histórico de los datos, con lo que se obtiene como consecuencia que los precios predichos difieran del valor real en mayor cantidad.

11. Recomendaciones

Para abarcar diferentes enfoques de desarrollo en los modelos de predicción, se recomienda implementar para su explicación el uso de variables exógenas o independientes con el objetivo de observar si otras variables existentes permiten generar una aproximación de pronósticos más precisa que la basada en series temporales, encontrando patrones de comportamiento entre dichas variables. Un análisis y desarrollo de modelos con el uso de variables de este tipo podría permitir identificar posibles características o efectos que producen la interacción de variables en el cambio del valor histórico de los precios y, así, de esta forma establecer sucesos en la economía que afectan los precios del sector agropecuario.

Se recomienda profundizar en el uso de otras técnicas de aprendizaje automático como las máquinas de soporte vectorial o los árboles de decisión, con el fin de determinar cuál puede ser la mejor técnica según el producto que se desea explicar, ya que aunque las redes neuronales han sido las técnicas más utilizadas para la predicción según la revisión de literatura, no se garantiza que sean las mejores para todos los productos, por tanto, se hace necesario el desarrollo de un estudio más profundo.

Referencias Bibliográficas

Abdulshahed, A. M., Longstaff, A. P., & Fletcher, S. (2015). The application of ANFIS prediction models for thermal error compensation on CNC machine tools. *Applied Soft Computing Journal*, *27*, 158–168. https://doi.org/10.1037/cdp0000045

Acevedo Borrego, A., & Linares Barrantes, M. (2012). El enfoque y rol del ingeniero industrial para la gestión y decisión en el mundo de las organizaciones. *Revista de La Facultad de Ingeniería Industrial*, *15*(1), 2012. Retrieved from https://www.redalyc.org/pdf/816/81624969002.pdf

Acosta Leal, D. A. (2014). Fijación de precios en mercados campesinos de Bogotá. Caso hortalizas frescas de Fúmeque y Chipaque (Cundinamarca). *Universidad Nacional de Colombia*, 116.

Agronet. (2016). *Principales Cultivos por Área Sembrada 2016:* Retrieved from http://www.agronet.gov.co/Documents/SANTANDER2016.pdf

Ahmad, H. A., Dozier, G. V., & Roland Sr., D. A. (2001). Egg price forecasting using neural networks. *Journal of Applied Poultry Research*, *10*(2), 162–171. https://doi.org/10.1093/japr/10.2.162

Ahmad, H. A., & Mariano, M. (2006). Comparison of forecasting methodologies using egg price as a test case. *Poultry Science*, *85*(4), 798–807. https://doi.org/10.1093/ps/85.4.789

Andrade Revelo, S. (2006). *Simulación de un Convertidor Multinivel Apilable controlado con Lógica Difusa*. Universidad de las Américas Puebla, Cholula, Puebla, Mexico. Retrieved from http://catarina.udlap.mx/u_dl_a/tales/documentos/meie/revelo_a_s/capitulo4.pdf

Anjoy, P., & Kumar, R. (2017). Comparative performance of wavelet-based neural network approaches. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-017-3289-9

Ayankoya, K., Calitz, A. P., & Greyling, J. H. (2016). Real-Time Grain Commodities Price Predictions in South Africa: A Big Data and Neural Networks Approach. *Agrekon*, *55*(4),

483–508. https://doi.org/10.1080/03031853.2016.1243060

Bartels, R. (1982). The Rank Version of von Neumann's Ratio Test for Randomness. *Journal of the American Statistical Association*, *77*(377), 40–46. https://doi.org/10.1080/01621459.1982.10477764

Bertona, L. F. (2005). Entrenamiento de redes neuronales basado en algoritmos evolutivos. *Grado, Tesis de Ingenieria Informática, Universidad de Buenos Aires*. https://doi.org/10.1371/journal.pone.0109400

Caeiro, F., & Mateus, A. (2015). randtests: Testing randomness in R.

Camara de Comercio de Bucaramanga. (2016). Producto Interno Bruto por departamentos. Retrieved April 5, 2018, from https://www.camaradirecta.com/temas/documentos pdf/informes de actualidad/2017/pib 2016.pdf

Correa, G. J., & Montoya, L. M. (2013). Aplicación del modelo ANFIS para predicción de series de tiempo. *Revista Digital de La Facultad de Ingenierias" Lámpsakos," 1*(9), 12–25. Retrieved from file:///C:/Users/luigy/Downloads/927-4404-2-PB (1).pdf

Damián, J. M. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones, 55. Retrieved from ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/matich-redesneuronales.pdf

DANE. (2018). Sistema de Información de Precios SIPSA.

de Arce, R., & Mahía, R. (2003). Modelos ARIMA, 31. Retrieved from http://www.uam.es/personal_pdi/economicas/rarce/pdf/Box-Jenkins.PDF

de Mattos Neto, P. S. G., Cavalcanti, G. D. C., & Madeiro, F. (2017). Nonlinear combination method of forecasters applied to PM time series. *Pattern Recognition Letters*, *95*, 65–72. https://doi.org/10.1016/j.patrec.2017.06.008

Departamento Administrativo Nacional de Estadistica. (2017). Principales Indicadores Del Mercado Laboral, 1–34. Retrieved from https://www.dane.gov.co/files/investigaciones/boletines/ech/ech/bol_empleo_abr_17.pdf

Departamento Nacional de planeacion. (2017). Big Data para la predicción de precios agrícolas, 13.

Duan, Q., Zhang, L., Wei, F., Xiao, X., & Wang, L. (2017). Forecasting model and validation for aquatic product price based on time series GA-SVR. *Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering*, *33*(1), 308–314. https://doi.org/10.11975/j.issn.1002-6819.2017.01.042

Fahimifard, S., Salarpour, M., Sabouhi, M., & Shirzady, S. (2009). Application of ANFIS to agricultural Eonomic variables forecasting case study: Poultry retail price. *Journal of Artificial Intelligence*, *2*(2), 65–72. https://doi.org/10.3923/jai.2009.65.72

FAO. (2001). Interpretación y uso de la información de mercados, 99. Retrieved from http://www.fao.org/tempref/docrep/fao/004/x8826S/x8826s00.pdf

FAO. (2004). *Política de desarrollo agrícola : conceptos y principios*. FAO. Retrieved from http://www.fao.org/docrep/007/y5673s/y5673s00.htm#Contents

Fritsch, S., & Guenther, F. (2016). neuralnet: Training of Neural Networks. Retrieved from https://journal.r-project.org/archive/2010/RJ-2010-006/RJ-2010-006.pdf

Galán, H., & Martínez, A. (2006). Inteligencia artificial . Redes neuronales y aplicaciones, 8. Retrieved from http://www.it.uc3m.es/jvillena/irc/practicas/10-11/06mem.pdf

García, S., Ramirez, S., Luengo, J., & Herrera, F. (2016). Big Data : Preprocesamiento y calidad de datos, 17–23. Retrieved from http://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf

Gobernación de Santander. (2017). Santander se mantiene con mayor producción en sus 4 líneas productivas.

Gómez, J., Palarea, J., & Martín, J. (2006). Métodos de inferencia estadística con datos faltantes. Estudio de simulación sobre los procesos en las estimaciones. *Estadística Española*, *48*(162), 241–270.

González Casimiro, M. P. (2009). *Técnicas de predicción económica*. España. Retrieved from https://addi.ehu.es/bitstream/handle/10810/12493/05-09pil.pdf?sequence=1

Guarnizo Lemus, C. (2011). *Metodología para la implementación de controlador difuso tipo Takagi-Sugeno en PLC s7-300*. Medellin, Colombia. Retrieved from http://www.scielo.org.co/pdf/tecn/v15n30/v15n30a05.pdf

Gujarati, D. N., & Porter, D. C. (2010). *Econometria*. *McGraw-Hill*.

Hyndman, R., Athanasopoulos, G., Razbash, S., Schmidt, D., Zhou, Z., & Khan, Y. (2013). forecast: Forecasting functions for time series and linear models [R package version 4.03].

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting : principles and practice*. Retrieved from https://books.google.com.co/books?hl=es&lr=&id=gDuRBAAAQBAJ&oi=fnd&pg=PA7&ots=qKModlDmmK&sig=Fc-Wp_6TT3mjd7Kfwx1QFCaM2fg&redir_esc=y#v=onepage&q&f=false

Jabez Harris, by. (2017). A Machine Learning Approach To Forecasting Consumer Food Prices, (August). Retrieved from http://dalspace.library.dal.ca/handle/10222/73170

Jha, G. K., & Sinha, K. (2013). Agricultural Price Forecasting Using Neural Network Model: An Innovative Information Delivery System. *Agricultural Economics Research Review*, *26*(2), 229--239. Retrieved from https://ageconsearch.umn.edu/bitstream/162150/2/8-GK-Jha.pdf

Jha, G. K., & Sinha, K. (2014). Time-delay neural networks for time series prediction: An application to the monthly wholesale price of oilseeds in India. *Neural Computing and Applications*, *24*(3–4), 563–571. https://doi.org/10.1007/s00521-012-1264-z

Kohzadi, N., Boyd, M. S., Kermanshahi, B., & Kaastra, I. (1996). A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing*, *10*(2), 169–181. https://doi.org/10.1016/0925-2312(95)00020-8

Kotler, P., & Armstrong, G. (2013). *Marketing*. *Marketing*. Mexico. https://doi.org/10.2307/1250103

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. Retrieved from https://datajobs.com/data-science-repo/Supervised-Learning-[SB-Kotsiantis].pdf

Larrañaga, P., Inza, I., & Moujahid, A. (1997). *Tema 8. Redes Neuronales*. Gipuzkoa, Spain . Retrieved from http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t8neuronales.pdf

Li, Z.-M., Cui, L.-G., Xu, S.-W., Weng, L.-Y., Dong, X.-X., Li, G.-Q., & Yu, H.-P. (2013). Prediction model of weekly retail price for eggs based on chaotic neural network. *Journal of Integrative Agriculture*, *12*(12), 2292–2299. https://doi.org/10.1016/S2095-3119(13)60610-3

Llano, L., Hoyos, A., Arias, F., & Velásquez, J. (2007). Comparación del Desempeño de Función de Activación en Redes Feedforward para aproximas Funciones de Datos con y sin ruido. *Avances En Sistemas E Informática*, *4*, 10. Retrieved from http://bdigital.unal.edu.co/15163/1/9756-17596-1-PB.pdf

Martín del Brio, B., & Sanz Molina, A. (2007). *Redes Neuronales y Sistemas Borrosos*. (A. G. E. S. . de CV, Ed.) (Tercera ed). Zaragoza.

Masaro, J. V. (2016). *Predicción de precios mediante modelización multivariada de series de tiempo. Una aplicación al sector lácteo argentino*. Universidad Nacional de Córdoba. Retrieved from https://rdu.unc.edu.ar/bitstream/handle/11086/3391/Vicentin Masaro%2C Jimena. Predicción de precios mediante modelización multivariada de serie de tiempo. Una aplicación al sector lacteo argentino.pdf?sequence=1&isAllowed=y

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data : la revolución de los datos masivos*. Turner. Retrieved from https://books.google.es/books?hl=es&lr=lang_es&id=uO9FbEcaMpkC&oi=fnd&pg=PA11&dq=Big+Data&ots=VZzU7krNxT&sig=r4nFMHc3tlYvqIS9ep_j7tkKMI0#v=onepage&q=Big Data&f=false

Medina, F., & Galván, M. (2007a). *Imputación de datos: teoría y práctica*. Santiago de Chile. Retrieved from https://repositorio.cepal.org/bitstream/handle/11362/4755/1/S0700590_es.pdf

Medina, F., & Galván, M. (2007b). Imputación de datos: teoría y práctica Santiago de Chile, julio de 2007, 84. Retrieved from https://repositorio.cepal.org/bitstream/handle/11362/4755/S0700590_es.pdf

Mena, J. (1999). *Data mining your website*. Digital Press. Retrieved from https://books.google.com.co/books?id=M1iXkVC6JvYC&printsec=frontcover&dq=inauthor:%22Jesus+Mena%22&hl=es-419&sa=X&ved=0ahUKEwi44-HS5KvaAhVBoFMKHcIrD1gQ6wEITTAG#v=onepage&q&f=false

Mendoza, Y., & Mazo, A. (2009). *Análisis del modelo ANFIS en el pronóstico de un título de renta variable*. Universidad Nacional de Colombia.

Ministerio de Agricultura, A. (2016). El 83.5% de los alimentos que consumen los colombianos son producidos por nuestros campesinos - 28 de octubre de 2016. Retrieved April 5, 2018, from http://www.agronet.gov.co/Noticias/Paginas/El-83-5-de-los-alimentos-que-consumen-los-colombianos-son-producidos-por-nuestros-campesinos-.aspx

Ministerio de Agricultura y Desarrollo Rural. (1993). Ley 101 de 1993, *149*(23). Retrieved from https://www.minagricultura.gov.co/Normatividad/Leyes/Ley 101 de 1993.pdf

Ministerio de Agricultura y Desarrollo Rural. (1994). Ley 160 de 1994. Retrieved from https://www.minagricultura.gov.co/Normatividad/Leyes/Ley 160 de 1994.pdf

Mishra, G. C., & Singh, A. (2013). A study on forecasting prices of groundnut oil in Delhi by arima methodology and artificial neural networks. *Agris On-Line Papers in Economics and Informatics*, *5*(3), 25–34. Retrieved from https://econpapers.repec.org/article/agsaolpei/157527.htm

Misra, P. K., & Goswami, K. (2015). Predictability of sugar futures: Evidence from the Indian commodity market. *Agricultural Finance Review*, *75*(4), 552–564. https://doi.org/10.1108/AFR-02-2014-0002

Mitra, D., & Paul, R. K. (2017). Hybrid time-series models for forecasting agricultural commodity prices. *Model Assisted Statistics and Applications*, *12*(3), 255–264. https://doi.org/10.3233/MAS-170400

Mora, J., Pérez, P., & Pérez, S. (2006). Utilización de redes ANFIS y señales de corriente para localización de la zona de falla en sistemas de distribución de energía eléctrica, *26*. Retrieved from http://www.redalyc.org/html/643/64326312/

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, *9*(1), 207–218.

Mustaffa, Z., & Sulaiman, M. H. (2015). Price predictive analysis mechanism utilizing grey wolf optimizer-Least Squares Support Vector Machines. *ARPN Journal of Engineering and Applied Sciences*, *10*(23), 17486–17491.

Ojeda Magaña, B. (2010). Aportación a la extracción de conocimiento aplicada a datos mediante agrupamientos y sistemas difusos, (December), 220. Retrieved from http://oa.upm.es/4838/

Pernaut Ardanaz, M., & Ortiz Felipe, E. J. (2008). *Introducción a la teoría económica*. Caracas: Universidad Católica Andrés Bello. Retrieved from https://books.google.com.co/books?id=yQOjLTNubkcC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R* (Second). New York: Springer. Retrieved from https://www.springer.com/gp/book/9780387759661

Pinheiro, C. A. O., & de Senna, V. (2017). Multivariate analysis and neural networks application to price forecasting in the Brazilian agricultural market | Aplicação de análise multivariada e redes neurais para previsão de preços no mercado agrícola brasileiro. *Ciencia Rural*, *47*(1). https://doi.org/10.1590/0103-8478cr20160077

Pitarque, A., Carlos, J., Francisco, J., Valencia, U. De, Pitarque, A., Carlos, J., … Roy, F. (2000). Las redes neuronales como herramientas estadísticas no paramétricas de clasificación. *Psicothema*, *12*, 459–463. Retrieved from http://www.psicothema.com/pdf/604.pdf

Polyiam, K., & Boonrawd, P. (2017). A hybrid forecasting model of cassava price based on artificial neural network with support vector machine technique. In *2017 3rd International Conference on Information Management (ICIM)* (pp. 123–127). https://doi.org/10.1109/INFOMAN.2017.7950359

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Reina, D. (2008). *Fundamentos de Matemática Difusa*.

Reina, M., Zuluaga, S., Bermúdez, W., & Oviedo, S. (2011). *La Política Comercial del Sector Agrícola en Colombia*. *Cuadernos de Fedesarrollo Número 38* (Vol. 38). Retrieved from http://www.repository.fedesarrollo.org.co/bitstream/handle/11445/161/CDF_No_38_Mayo_2011.pdf?sequence=1

Ribeiro, C. O., & Oliveira, S. M. (2011). A hybrid commodity price-forecasting model applied to the sugar-alcohol sector. *Australian Journal of Agricultural and Resource Economics*, *55*(2), 180–198. https://doi.org/10.1111/j.1467-8489.2011.00534.x

Riza, L. S., Bergmeir, C., Herrera, F., & Benítez, J. M. (2015). {frbs}: Fuzzy Rule-Based Systems for Classification and Regression in {R}. *Journal of Statistical Software*, *65*(6), 1–30.

Sánchez Anzola, N. (2016). Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario. *ODEON*, (9), 113. https://doi.org/10.18601/17941113.n9.04

Sanjay Krishnankutty, A., Torres Blanc, C., & Sánchez Torrubia, M. G. (2015). Introducción a los FIS. Retrieved October 22, 2018, from http://www.dma.fi.upm.es/recursos/aplicaciones/logica_borrosa/web/fuzzy_inferencia/introfis.htm

Secretaría de Agricultura. (2017). Secretaría. Retrieved April 5, 2018, from http://www.santander.gov.co/index.php/secretaria-agricultura

Shih, M. L., Huang, B. W., Chiu, N.-H., Chiu, C., & Hu, W. Y. (2009). Farm price prediction using case-based reasoning approach-A case of broiler industry in Taiwan. *Computers and Electronics in Agriculture*, *66*(1), 70–75. https://doi.org/10.1016/j.compag.2008.12.005

Sievert, C. (2018). plotly for R.

SIPSA, S. de I. de P. y A. del S. A. (2017). Boletín Quincenal, Abastecimiento de Alimentos, 16. Retrieved from https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-de-

informacion/sipsa

Torres Soler, L. C. (2010). *Redes Neuronales*. Bogotá, Colombia. Retrieved from http://disi.unal.edu.co/~lctorress/RedNeu/LiRna008.pdf

Trujillano Cabello, J., Badía Castelló, M., March Llanes, J., Rodríguez Pozo, À., Serviá Goixart, L., & Sorribas Tello, A. (2005). Redes neuronales artificiales en medicina intensiva. Ejemplo de aplicación con las variables del MPM II. *Medicina Intensiva*, *29*(1), 13–20. https://doi.org/10.1016/S0210-5691(05)74198-X

Urna de Cristal. (2016). Con la paz gana el campo. Retrieved April 5, 2018, from http://www.urnadecristal.gov.co/paz-y-campo

Valencia, M., Yáñez, C., & Sánchez, L. (2006). Algoritmo Backpropagation para Redes Neuronales: conceptos y aplicaicones, 14. Retrieved from http://www.repositoriodigital.ipn.mx/bitstream/123456789/8628/1/Archivo que incluye portada%2C índice y texto.pdf

Velásquez, J. D., Zambrano, C., & Vélez Laura. (2011). ARNN: Un paquete para la predicción de series de tiempo usando redes neuronales autorregresivas ARNN: A packages for time series., 5. Retrieved from http://bdigital.unal.edu.co/28850/1/26744-93664-1-PB.pdf

Velásquez, J., Fonnegra, Y., & Villa, F. (2013). Pronóstico de series de tiempo con redes neuronales regularizadas y validación cruzada. *Vínculos*, 267–279. Retrieved from http://revistas.udistrital.edu.co/ojs/index.php/vinculos/article/view/4702

Villavicencio, J. (2011). *Introducción a Series de Tiempo*. Puerto Rico. Retrieved from http://www.estadisticas.gobierno.pr/iepr/LinkClick.aspx?fileticket=4_BxecUaZmg%3D

Wei, T., & Simko, V. (2017). R package "corrplot": Visualization of a Correlation Matrix.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Xiong, T., Li, C., & Bao, Y. (2017). An improved EEMD-based hybrid approach for the short-term forecasting of hog price in China. *Agricultural Economics (Czech Republic)*, *63*(3), 136–148. https://doi.org/10.17221/268/2015-AGRICECON

Xiong, T., Li, C., & Bao, Y. (2017). Seasonal forecasting of agricultural commodity price using a hybrid STL and ELM method: Evidence from the vegetable market in China. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2017.11.053

Xiong, T., Li, C., Bao, Y., Hu, Z., & Zhang, L. (2015). A combination method for interval forecasting of agricultural commodity futures prices. *Knowledge-Based Systems*, *77*, 92–102. https://doi.org/10.1016/j.knosys.2015.01.002

Yasrebi, S. S., & Emami, M. (2008). Application of Artificial Neural Networks (ANNs) in prediction and Interpretation of Pressuremeter Test Results. Retrieved from https://pdfs.semanticscholar.org/6365/182f2a94dcfd05f99dd11a7740b50a2e852c.pdf

Zankova, E. (2016). *High frequency financial time series prediction : machine learning approach*. Faculty of Science and Technology. Retrieved from https://munin.uit.no/bitstream/handle/10037/9255/thesis.pdf?sequence=2

Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News*, *2*(3), 7–10.

Zeleckis, E. (2015). Solid Biofuel – Wood Chips price in Lithuania modeling using artificial Neural Network (ANN) and ARIMA models, (May), 105. Retrieved from file:///C:/Users/luigy/Downloads/19501887 (3).pdf

Zhang, G. P. (2003). *Time series forecasting using a hybrid ARIMA and neural network model*. *Neurocomputing* (Vol. 50).

Zhanga, J. H., Kongb, F. T., Wu, J. Z., Zhu, M. S., Xu, K., & Liu, J. J. (2014). *Tomato prices time series prediction model based on wavelet neural network*. *Applied Mechanics and Materials* (Vol. 644–650). https://doi.org/10.4028/www.scientific.net/AMM.644-650.2636

Zhao, Y., Zhang, Y., & Xu, H. (2012). A hybrid neural network method for detecting structural change in oil-bioenergy crops prices system. *Research Journal of Applied Sciences, Engineering and Technology*, *4*(18), 3363–3367. Retrieved from http://maxwellsci.com/print/rjaset/v4-3363-3367.pdf

Zhu, Q., Pan, L., Yin, Y., & Li, X. (2013). Influence on normalization and magnitude

normalization for price forecasting of agricultural products. *Information Technology Journal*, *12*(15), 3046–3057. https://doi.org/10.3923/itj.2013.3046.3057

APENDICE A

Selección de los productos agrícolas para el estudio

Productos agrícolas seleccionados

A continuación, se nombrará y proporcionara cierta información (periodos de siembra, hectáreas cultivadas y características) acerca de los productos que fueron seleccionados para el estudio del proyecto bajo criterios que son mencionados en el caso de estudio. Permanente o transitorio indica el tipo de producto según el periodo de cultivo, además las áreas de cultivo que se enuncian en cada uno de los productos que se exponen a continuación son obtenidas a partir de los últimos informes generados por la entidad encargada Agronet entre los años 2016 y 2017.

Plátano (permanente)

Figura 58: Plátano. Adaptado de Food Home (2018). https://conservaciondealimentos.com/vegetales/platano-macho/#.W1DFrtJKjIX

El Plátano o *Musa Acuminata* es uno de los productos más importantes a nivel nacional, ya que su cultivo se ha caracterizado por ser reconocido como de regular prioridad, debido a su importancia social y económica en varios aspectos como su siembra, la generación de empleo, la dieta en la población, entre otros (Cayón Salinas & Salazar Alonso, 2001). Además, "su cultivo es llamativo para ser desarrollado en ciertas regiones, ya que requiere de pocos cuidados y soporta condiciones climáticas desfavorables" (SIPSA, 2014, p. 1). En cuanto a la siembra, la época más apropiada para sembrarlo es en el inicio de la temporada de lluvias. La presencia del cultivo del plátano en muchas partes del territorio hace que el manejo de su cadena sea complejo; en la mayor

parte de las regiones los plátanos llegan a las plazas desde sus mismas regiones a excepción de Bogotá. Los precios son regulados por intermedíarios, pues aún no se ha podido organizar de manera que se garantice su comercialización en todas las regiones (Ministerio de Agricultura y desarrollo rural, 2014).

En Santander el área sembrada es de 17.538 hectáreas.

Yuca (transitorio)

Figura 59: Yuca. Adaptado de Fresh Fruits & Vegetables (2015b). http://flp.co/en/cassava

"La Yuca o *Manihot Esculenta Crantz* junto con el maíz, la caña de azúcar y el arroz constituyen las fuentes más importantes de energía en las regiones tropicales del mundo. En el sur del continente la yuca es conocida como mandioca" (Vila Flórez, 2012). La yuca es un cultivo que tiene sus orígenes de América Latina el cual se ha cultivado desde épocas prehistóricas. Su adaptación a diversos ecosistemas, su potencial de producción, la diversidad de sus mercados y usos finales la han convertido en una de las bases de la alimentación para la población rural y en una alternativa de comercialización en centros urbanos (Vila Flórez, 2012).

La Yuca forma parte del sistema alimentario mundial. El consumo de yuca se extiende vigorosamente en el mundo en desarrollo que hoy produce más de la mitad de la cosecha mundial, donde la facilidad del cultivo y el gran contenido de energía la han convertido en un valioso producto comercial para millones de agricultores. La yuca se debe cosechar cuando cumpla el

período según variedad y altura sobre el nivel del mar, así: de 0 a 1000 metros (8 -12 meses) de 1000 a 1500 metros (13-17 meses) de 1500 a 1700 metros (18 - 22 meses) (Agronet, 2006).

En Santander la superficie sembrada de Yuca es de 9248 hectáreas.

Piña (permanente)

Figura 60: Piña. Adaptado de El Diario NY (2017). https://eldíariony.com/guia-de-compras/beneficios-de-la-pina-en-tu-dieta-para-desintoxicar-tu-cuerpo-y-bajar-de-peso/

La piña de nombre científico *Ananas comosus L.* tiene como origen a América del Sur ya que no se conoce con certeza el país de procedencia, pero por su domesticación se cree que puede ser de una zona entre Brasil y Paraguay, de donde se propago a otros países del continente y posteriormente a Europa y Asia. De acuerdo a las cifras de la Encuesta Nacional Agropecuaria (ENA) se registró un total de 8.871 hectáreas sembradas para el cultivo de la piña a nivel nacional, de las cuales el 51.38 % correspondió a el área en edad productiva, de donde se extrajo un total de 125.150 toneladas, así el departamento de Valle del Cauca reportó los mayores volúmenes con un 35.22 %, seguido de Quindío con el 25.29 %, Santander 11.68 %, Cauca 10.82 % y Casanare con 7.20 % y otros con 9.79% (DANE, 2016). Por lo general se realizan dos cosechas que dependiendo de la variedad y de los factores ambientales se pueden dar una primera de los 15 a los 24 meses y otra de los 15 a 18 meses posteriores a esta primera.

En Santander el área sembrada de piña corresponde a 11.517 hectáreas.

Mandarina Común (permanente)

Figura 61: Mandarina. Adaptado de El País (2017).
http://elpaisonline.com/index.php/sociales-2/item/242745-la-mandarina-ayuda-a-prevenir-el-cancer

Citrus reticulata Blanco o Mandarina es una de las especies más conocidas y cultivadas del grupo de los frutos cítricos (Universidad Lasallista, 2012). El grupo de las mandarinas se caracterizan por tener en común su fácil pelado; son árboles de mediano porte, muy ramificados, flor pequeña, abundante, y con frecuencia presentan tendencia a la alternancia (Universidad Lasallista, 2012). Los cítricos registrados para la cadena se encuentran distribuidos en gran parte del territorio nacional, la producción se considera constante, aunque se presenta baja producción entre los meses de marzo, abril, agosto y septiembre (Espinal G, Martínez C, & Peña M, 2005). Si se tiene disponibilidad de riego se puede sembrar en cualquier época del año, en caso contrario la época más adecuada es al inicio de la época lluviosa.

En Santander no se conoce exactamente la cantidad de mandarina cultivada debido a que las entidades encargadas de sus estadísticas la incluyen dentro del grupo denominado cítricos.

Naranja valencia (permanente)

Figura 62: Naranja. Adaptado de Fuente Saludable (2018).
https://www.fuentesaludable.com/efectos-negativos-y-secundarios-de-la-naranja/

La *Citrus Sinensis* o naranja dulce es la especie más importante de cítricos, si se tiene en cuenta su producción mundial (Universidad Lasallista, 2012). Dentro de este grupo se encuentra la naranja valencia siendo naranja dulce tardía más cultivada en las principales regiones citrícolas del mundo y en Colombia.

En Colombia existen dos periodos de baja oferta de la fruta: marzo, abril, agosto y septiembre, encontrándose en los demás meses una adecuada asequibilidad, debido a que el eje cafetero coloca en el mercado su producción en los meses de mayo-julio y octubre-diciembre, y la región de la Orinoquía, específicamente Meta y Casanare, con cosechas en la época de octubre-febrero y julio-agosto; por otra parte, Santander hace su aporte en los meses de diciembre-enero y, mayo-junio, mientras que la Costa Atlántica contribuye con la producción en los meses de marzo a junio (Espinal G et al., 2005). El momento más adecuado para hacer la siembra, es durante el inicio de la época lluviosa, y puede extenderse hasta el mes de julio, tomando en consideración el receso de las lluvias durante el verano; sin embargo, se puede llevar a cabo en este tiempo si se cuenta con riego. La época de cosecha para la naranja valencia es en los meses de febrero y mayo (Ministerio de Agricultura y Ganadería, 1991).

Como el caso de la mandarina el área total sembrada de este producto es desconocida debido a su inclusión en productos cítricos.

Banano (Permanente)

Figura 63: Banano. Adaptado de Fresh Fruits & Vegetables (2015a). http://flp.co/en/baby-banana

Se le considera originario de las regiones tropicales y húmedas de Asia. Según la variedad de la planta el banano alcanza de 3 hasta 7 metros de altura constituyéndose como una planta herbácea; su tallo está formado por pecíolos de hojas curvadas y comprimidas dispuestas en bandas en espiral que desde el centro van formándose sucesivamente nuevas hojas y al extenderse comprimen hacia el exterior las bases de las hojas más viejas (Anacafe, 2011). De acuerdo con la variedad un racimo puede llegar a tener 100 a 400 frutos, cada uno llega a tener de 8 a 20 centímetros de largo con un peso entre 1 y 4 onzas (Anacafe, 2011); a los 14 meses después de la siembra de los rizomas o 4 meses después de aparecer la yema floral, los racimos están listos para ser cosechados. Sin duda, este cultivo cumple un papel importante en la economía agrícola del país. En 2016 Colombia exportó 1.830.388 toneladas de banano cavendish por un valor de 848.7 millones de dólares (Semana, 2017).

El banano posee 3.335 de hectáreas sembradas en el departamento Santandereano.

Papa negra (transitorio)

Figura 64: Papa negra. Adaptado de Green Shop (2017). https://greenshop.com.ar/producto/papa-negra-500gr/

La *Solanum Tuberosum* o papa negra también conocida como “Yema de Huevo” por su carne amarilla, posee la forma redonda en los frutos más pequeños, ovalada en los medianos y alargada en los grandes; se siembra a lo largo de todo el año en un ciclo de cuatro meses, necesita mayor humedad que el resto de las papas antiguas y excepcionalmente con las papas borrallas que pueden cultivarse en costas bajas, lo que se suele realizar para producciones de demanda navideña (Panizo

& Canaria De Semillas, 2015). La papa negra presenta dos subespecies: la *andigena*, que corresponde a las variedades que se cultivan actualmente en los Andes en América del Sur, desde Venezuela hasta el norte de Argentina y la subespecie *tuberosum* distribuida por todo el mundo (DANE, 2017).

Este tubérculo es un alimento muy importante en la dieta de la población mundial, ya que de acuerdo con cifras de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (OCDE-FAO, 2014), se produjo un poco más de 1.298 billones de toneladas participando Colombia con 2.157.568 toneladas.

En Santander existen 5.782 de hectáreas sembradas.

Tomate (Transitorio)

Figura 65: Tomate. Adaptado de ABC (2017). https://www.abc.es/viajar/gastronomia/abci-manzanas-paraiso-201708010826_noticia.html

El *Lycopersicum Esculemtum Mill* o Tomate representa un sector predominante en la economía de la agricultura mundial al ser la hortaliza más importante en el mundo y un producto primordial en la alimentación de muchos países; el tomate compone el 30% de la producción hortícola mundial con 163.963.770 toneladas y aproximadamente 4.6 millones de hectáreas sembradas (Saboyá López, 2016); en Colombia los departamentos con mayor producción (en toneladas) a campo abierto son Norte de Santander 116.507, Santander 52.599 y Boyacá con 44.870 (Saboyá López, 2016). El ciclo del cultivo del tomate es durante siete meses, contados a partir del trasplante hasta lograr el último corte. “La fase reproductiva empieza a partir del inicio de la floración,

pasando luego por la formación y llenado del fruto hasta llegar a la madurez, cuando están listos para la primera cosecha; esta etapa tiene una duración aproximada de 180 días" (DANE, 2014).

El tomate posee 1.963 hectáreas cultivadas en el departamento de Santander.

Apéndice B

Errores RMSE y MAPE del modelo RNA bp

El presente apéndice expone los resultados de las métricas de error de los modelos de redes neuronales teniendo como base el número de nodos utilizados en las capas ocultas y el número de variables utilizadas.

Resultados banano del día martes

RMSE

N	1 Var	2 Var	3 Var
1	184.39	187.30	118.34
2	199.36	164.19	118.91
3	151.93	162.05	123.09
4	199.89	177.57	113.80
5	199.09	180.80	119.90
6	202.50	191.34	102.29
7	224.78	185.72	113.43
8	203.85	183.09	120.45
9	220.55	159.59	154.47
10	233.93	160.45	153.08

MAPE

N	1 Var	2 Var	3 Var
1	8.84	9.03	5.79
2	9.72	7.90	5.78
3	7.30	7.82	5.98
4	9.75	8.49	5.52
5	9.70	8.67	5.85
6	9.91	9.27	5.06
7	11.25	8.96	5.55
8	9.99	8.81	5.87
9	10.99	7.69	7.41
10	11.78	7.73	7.40

Resultados Banano del día jueves

RMSE

N	1 Var
1	145.79
2	171.80
3	206.43
4	208.92
5	199.90
6	214.26
7	213.45
8	206.66
9	167.20
10	178.38

MAPE

N	1 Var
1	6.69
2	7.90
3	10.06
4	10.20
5	9.66
6	10.53
7	10.48
8	10.07
9	7.69
10	8.30

Resultados Banano del día viernes

RMSE

N	1 Var
1	98.54
2	129.61
3	147.32
4	164.27
5	159.25
6	172.43
7	166.63
8	152.37
9	110.96
10	140.71

MAPE

N	1 Var
1	4.85
2	6.12
3	7.16
4	8.19
5	7.88

6	8.69
7	8.29
8	7.46
9	5.34
10	6.76

Resultados Mandarina del día martes

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var	7 Var	8 Var
1	851.43	801.36	712.54	672.45	495.02	433.20	459.49	471.09
2	851.81	781.90	708.40	642.69	514.31	401.69	323.54	454.91
3	854.68	754.45	725.39	603.77	548.80	367.81	443.81	514.19
4	835.90	771.28	715.57	544.16	500.77	348.65	420.98	597.94
5	818.05	749.67	721.37	589.75	541.18	487.16	339.44	436.11
6	841.78	803.72	728.24	494.09	474.10	406.74	439.05	490.13
7	855.04	755.65	730.54	605.29	499.49	358.75	467.019	435.05
8	838.39	699.86	691.35	564.93	595.78	366.07	397.41	452.03
9	817.54	779.31	719.42	681.09	490.84	314.75	440.12	497.26
10	838.77	800.78	679.64	546.92	493.55	677.82	360.29	431.69

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var	7 Var	8 Var
1	35.30	33.28	29.96	28.36	21.30	19.09	20.21	20.53
2	35.24	32.47	29.79	27.35	22.08	17.86	15.32	20.06
3	35.23	31.26	30.44	25.40	23.51	16.66	19.58	22.09
4	34.78	31.91	30.41	24.79	21.88	17.04	18.54	24.82
5	34.28	31.13	30.03	25.28	23.00	20.73	15.55	19.16
6	35.00	33.17	30.79	21.35	20.66	21.35	19.53	21.28
7	35.43	31.42	30.39	25.55	21.22	17.38	20.41	19.13
8	35.07	29.15	28.87	24.29	24.98	18.15	18.47	19.82
9	33.90	32.21	29.97	34.53	21.12	17.32	19.45	21.31
10	34.98	33.50	28.79	23.91	21.43	34.23	15.67	19.01

Resultados Mandarina del día jueves

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	933.22	857.02	822.50	740.66	603.40	515.65
2	931.18	860.90	825.87	718.15	620.18	546.14
3	935.17	871.61	773.55	709.29	629.47	454.21
4	933.26	851.41	700.11	785.66	658.60	196.81
5	924.52	902.12	825.13	652.82	663.83	484.40
6	932.61	906.67	805.68	726.31	528.20	340.81
7	901.26	898.83	840.11	696.08	592.85	526.79
8	924.70	808.17	826.56	814.99	564.18	393.09
9	937.79	803.08	830.13	887.04	638.26	493.31
10	929.32	718.44	832.80	720.72	721.89	596.02

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	40.81	37.01	34.72	31.22	25.84	22.47
2	40.88	37.77	35.13	30.71	26.59	23.74
3	41.22	36.86	32.93	30.50	27.04	19.58
4	41.27	37.23	30.71	33.60	28.32	9.31
5	39.85	40.27	36.05	27.04	28.77	20.88
6	40.96	40.20	35.31	30.05	22.37	14.88
7	40.24	39.62	36.38	29.26	25.35	22.88
8	39.86	34.78	36.22	36.02	24.00	17.25
9	41.61	34.67	34.50	40.12	27.34	21.06
10	41.21	34.06	34.91	34.22	34.41	25.94

Resultados Mandarina del día viernes

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	978.91	768.67	730.91	654.26	657.30
2	953.72	785.96	724.54	538.70	640.31
3	1006.17	763.43	669.11	563.10	617.08
4	975.85	836.40	742.27	619.11	711.07
5	1018.30	739.09	690.36	528.60	714.82
6	999.63	779.26	826.65	604.36	647.91
7	1017.13	758.28	657.74	627.69	601.28
8	1015.07	741.02	714.47	527.17	707.47
9	924.35	748.81	787.94	533.14	653.08
10	948.28	730.24	850.82	419.99	733.19

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	42.78	33.48	32.00	28.35	28.49
2	41.20	33.79	31.69	23.64	27.97
3	44.53	32.56	28.80	24.477	26.92
4	42.16	36.91	31.31	27.11	30.40
5	45.39	32.54	29.47	22.56	31.17
6	44.24	34.18	35.76	26.52	28.12
7	45.32	32.84	28.39	27.20	25.93
8	45.12	32.34	35.00	22.98	31.04
9	40.11	32.80	34.64	23.26	28.47
10	41.05	31.83	36.97	18.24	31.60

Resultados Naranja del día martes

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	262.37	269.53	234.91	210.13	196.35
2	267.80	277.48	236.22	190.99	184.68
3	258.84	273.06	250.52	212.61	186.26
4	250.67	266.12	258.52	188.22	212.40
5	250.69	259.35	239.27	209.33	197.34
6	255.05	261.46	283.26	190.20	194.19
7	251.29	275.28	236.17	181.92	193.15
8	261.68	265.55	230.53	184.28	210.88
9	265.47	262.84	235.61	180.78	192.91
10	268.51	265.20	234.68	180.23	185.89

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	26.10	26.66	23.39	21.00	19.69
2	26.62	27.36	23.52	19.15	18.60
3	25.79	26.98	24.94	21.25	18.93
4	25.01	26.33	25.58	19.00	21.14
5	25.01	25.71	23.85	20.87	19.70
6	25.42	25.90	27.70	19.03	19.46
7	25.07	27.21	23.53	18.76	19.58
8	26.04	26.29	23.00	18.79	20.98
9	26.41	26.04	23.41	18.64	19.35
10	26.69	26.26	23.39	18.34	19.14

Resultados Naranja del día jueves

RMSE

N	1 Var	2 Var	3 Var
1	277.94	273.37	274.58
2	276.75	272.53	269.46
3	270.04	280.41	306.33
4	271.93	274.40	287.39
5	275.16	266.07	281.94
6	269.23	285.72	288.25
7	275.03	267.54	272.34
8	272.96	271.52	271.93
9	272.09	270.76	290.13
10	274.57	272.73	276.25

MAPE

N	1 Var	2 Var	3 Var
1	26.69	26.22	26.35
2	26.51	26.14	25.78
3	25.91	26.97	29.44
4	26.10	26.32	27.69
5	26.38	25.46	26.98
6	25.81	27.55	27.50

7	26.38	25.69	26.12
8	26.19	26.04	26.10
9	26.12	26.00	27.73
10	26.35	26.20	26.49

Resultados Naranja del día viernes

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	263.56	273.02	244.24	209.61	232.88	230.31
2	261.15	271.51	236.91	209.07	225.22	229.15
3	261.85	264.84	250.38	237.51	192.69	283.71
4	253.56	271.39	239.61	192.80	212.08	245.85
5	252.52	265.60	243.91	191.99	251.69	287.10
6	274.62	257.72	216.79	194.65	218.61	193.86
7	252.91	261.87	246.20	197.47	239.40	167.43
8	236.55	241.57	219.95	195.30	269.62	197.50
9	231.72	257.98	216.81	206.26	217.21	225.33
10	238.22	266.73	229.26	201.14	236.29	279.68

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	26.10	27.11	24.42	21.10	23.45	23.21
2	25.84	26.94	23.75	20.98	22.61	23.09
3	25.92	26.23	24.77	23.15	20.51	28.50
4	25.04	26.91	23.61	19.54	20.88	24.35
5	24.94	27.03	24.28	20.03	25.03	28.61
6	27.29	25.49	21.71	19.65	21.86	20.66
7	24.99	25.88	24.27	20.01	23.81	18.52
8	24.79	23.85	22.37	19.39	26.63	20.23
9	22.97	25.49	21.97	20.29	22.04	22.72
10	23.54	26.41	23.11	20.38	23.55	27.87

Resultados Papa del día martes

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	214.37	268.59	229.13	240.35	214.37	209.91
2	209.03	228.89	249.66	227.32	200.26	208.70
3	208.61	220.79	209.47	226.68	208.06	214.43
4	197.65	241.59	213.01	252.71	231.71	193.31
5	216.42	257.08	206.28	224.56	172.54	228.94
6	252.56	218.63	244.30	271.49	184.04	234.31
7	207.34	222.52	164.78	230.75	214.16	248.46
8	214.42	166.16	216.67	193.90	183.41	199.01
9	181.51	206.48	218.93	192.45	226.25	199.03
10	204.47	171.39	202.88	203.91	201.42	198.29

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	18.21	23.56	19.78	20.76	18.37	17.57
2	17.77	19.71	21.68	19.55	17.14	17.38
3	17.72	18.97	17.93	19.51	17.80	17.85
4	16.78	20.83	18.25	21.95	19.96	15.14
5	18.43	22.41	17.63	19.33	14.37	19.51
6	21.80	18.75	21.13	23.81	15.67	19.86
7	17.59	19.12	13.68	19.90	18.37	21.43
8	17.07	13.88	18.60	16.54	15.61	16.60
9	15.27	17.60	18.83	16.43	19.48	16.49
10	17.35	14.39	17.36	17.42	17.21	16.11

Resultados Papa del día jueves

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	128.56	136.12	134.22	119.42	145.52	134.68
2	126.48	137.91	136.35	117.73	138.60	142.98
3	125.68	140.47	116.59	177.58	137.75	156.72
4	131.94	131.99	145.97	143.59	178.04	149.45
5	153.78	130.46	115.96	179.21	141.50	144.37
6	154.53	208.78	190.91	170.55	134.45	151.84
7	132.15	119.16	114.44	121.79	138.36	193.11
8	214.49	115.40	139.63	126.71	129.79	131.31
9	134.29	129.32	146.93	116.62	129.35	133.70
10	156.69	119.72	125.37	119.64	138.05	140.42

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	10.14	10.72	10.58	9.58	11.24	10.52
2	10.037	10.87	10.75	9.51	10.86	11.53
3	9.99	11.06	9.31	14.61	10.78	12.67
4	10.34	10.35	11.49	10.57	14.73	11.26
5	12.35	10.28	9.23	12.70	11.01	11.49
6	12.41	17.83	16.03	12.15	10.61	12.37
7	10.37	9.54	9.13	9.83	11.33	16.04
8	17.21	9.20	10.96	9.77	10.19	10.42
9	10.49	10.17	11.63	9.57	10.35	10.56
10	12.65	9.58	10.02	9.51	10.85	10.81

Resultados Papa del día viernes

RMSE

N	1 Var	2 Var	3 Var
1	127.25	134.43	139.97
2	127.98	134.34	132.79
3	156.04	144.50	134.57
4	136.78	133.06	144.32
5	128.24	135.64	155.12
6	137.95	134.92	136.74
7	142.97	130.32	137.14
8	136.79	133.66	138.17
9	137.27	135.51	141.85
10	127.35	138.63	156.76

MAPE

N	1 Var	2 Var	3 Var
1	9.27	9.82	9.61
2	9.311	9.82	9.67
3	11.63	10.45	9.68
4	10.06	9.82	9.94
5	9.14	9.88	10.40
6	10.14	9.92	9.63
7	10.45	9.42	9.62
8	10.069	9.77	9.65
9	10.09	9.84	9.78
10	9.27	10.04	10.49

Resultados Piña del día martes

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	62.87	59.48	67.02	66.71	74.41
2	62.31	53.78	66.34	64.84	75.33
3	62.51	58.16	61.52	63.85	76.92
4	66.28	62.22	64.94	67.35	70.43
5	66.16	64.90	68.41	72.38	75.02
6	66.95	62.76	66.44	69.76	72.74
7	68.40	59.71	65.27	69.97	73.75
8	67.48	63.28	68.92	66.74	74.01
9	65.09	61.43	69.25	68.69	74.18
10	61.81	59.77	71.25	67.16	75.19

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	7.84	7.25	8.20	8.22	9.77
2	7.74	6.73	8.13	8.10	9.89
3	7.79	7.10	7.59	7.89	10.07
4	8.38	7.54	7.99	8.30	9.25

5	8.40	7.89	8.35	9.32	9.85
6	8.56	7.61	8.14	8.51	9.42
7	8.86	7.27	8.01	8.61	9.74
8	8.69	7.68	8.42	8.20	9.75
9	8.23	7.46	8.47	8.40	9.76
10	7.70	7.28	8.71	8.40	9.89

Resultados Piña del día jueves

RMSE

N	1 Var	2 Var	3 Var
1	59.90	55.34	70.50
2	58.03	55.63	55.15
3	67.52	52.67	77.10
4	65.72	52.77	74.30
5	69.25	53.77	56.46
6	68.68	58.13	75.24
7	69.80	55.34	61.04
8	66.13	56.31	82.46
9	67.68	56.62	72.57
10	69.85	56.23	71.22

MAPE

N	1 Var	2 Var	3 Var
1	7.87	7.33	9.13
2	7.61	7.37	7.13
3	9.13	7.00	10.12
4	8.81	7.01	9.67
5	9.42	7.14	7.30
6	9.33	7.68	9.79
7	9.53	7.34	7.88
8	8.89	7.45	10.84
9	9.16	7.49	9.39
10	9.54	7.45	9.25

Resultados Piña del día viernes

RMSE

N	1 Var	2 Var	3 Var	4 Var
1	68.88	69.93	79.27	78.53
2	63.85	68.97	77.90	74.20
3	69.19	69.50	75.85	79.94
4	68.73	70.69	80.76	83.94
5	70.58	74.89	77.05	80.40
6	70.02	69.92	79.93	84.27
7	69.95	78.31	85.57	75.64
8	69.58	71.63	71.03	83.31
9	69.86	70.21	74.68	78.46
10	68.85	70.18	81.23	78.77

MAPE

N	1 Var	2 Var	3 Var	4 Var
1	9.32	9.45	10.53	10.48
2	8.50	9.32	10.37	9.88
3	9.38	9.39	10.07	10.67
4	9.29	9.56	10.77	11.25
5	9.63	10.19	10.27	10.71
6	9.53	9.45	10.67	11.30
7	9.52	10.76	11.51	10.10
8	9.45	9.69	9.38	11.22
9	9.50	9.49	9.90	10.46
10	9.32	9.48	10.83	10.52

Resultados Plátano del día martes

RMSE

N	1 Var	2 Var
1	299.10	271.15
2	280.22	269.28
3	307.36	299.75
4	295.68	273.46
5	302.66	252.09
6	312.13	305.76
7	321.86	267.51
8	291.87	255.19
9	311.92	286.50
10	258.94	317.68

MAPE

N	1 Var	2 Var
1	14.70	13.16
2	13.63	13.05
3	15.15	14.75
4	14.50	13.29
5	14.89	12.06
6	15.43	15.09
7	15.95	12.95
8	14.31	12.24
9	15.42	14.02
10	12.62	15.74

Resultados Plátano del día jueves

RMSE

N	1 Var	2 Var
1	274.04	254.98
2	264.20	256.63
3	277.89	229.38
4	265.89	250.24
5	255.70	289.93
6	267.53	231.54
7	265.66	234.01
8	277.95	254.98
9	279.11	255.48
10	275.44	265.20

MAPE

N	1 Var	2 Var
1	14.38	12.97
2	13.74	13.07
3	14.62	11.29
4	13.85	12.64
5	13.19	15.16
6	13.95	11.43
7	13.84	11.59
8	14.63	12.97
9	14.70	13.00
10	14.46	13.61

Resultados Plátano del día viernes

RMSE

N	1 Var	2 Var	3 Var
1	327.61	299.19	296.83
2	341.35	299.19	307.46
3	321.69	316.26	270.44
4	309.96	300.27	299.97
5	320.25	299.35	296.77
6	315.54	287.35	313.38
7	337.26	266.21	291.32
8	321.69	283.90	297.74
9	319.53	279.89	312.23
10	330.53	290.95	285.52

MAPE

N	1 Var	2 Var	3 Var
1	17.61	15.71	15.56
2	18.48	15.71	16.38
3	17.24	16.73	13.93
4	16.52	15.76	15.80
5	17.15	15.71	15.73

6	16.87	14.98	16.56
7	18.21	13.67	15.22
8	17.24	14.76	15.59
9	17.11	14.51	16.67
10	41.21	34.06	34.91

Resultados Tomate del día martes

RMSE

N	1 Var	2 Var	3 Var
1	682.66	690.35	693.14
2	682.41	683.26	677.37
3	633.87	661.09	694.21
4	642.19	713.32	687.25
5	689.16	700.59	670.39
6	635.49	673.64	708.85
7	623.22	679.58	730.09
8	643.32	700.19	720.49
9	640.81	655.23	676.41
10	625.27	661.57	722.54

MAPE

N	1 Var	2 Var	3 Var
1	22.51	24.18	24.93
2	22.47	24.25	26.07
3	23.65	24.39	25.28
4	23.39	24.23	25.13
5	21.92	24.45	26.61
6	23.55	25.11	24.87
7	25.12	24.70	24.30
8	23.33	23.85	25.38
9	23.37	24	24.88
10	24.59	24.10	25.36

Resultados Tomate del día jueves

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	783.17	793.72	786.87	816.78	873.50
2	819.01	841.50	774.46	743.35	984.30
3	755.50	821.55	747.07	800.15	774.46
4	733.08	744.24	855.10	762.47	908.54
5	766.98	753.19	760.78	889.69	926.36
6	753.71	761.44	844.65	768.18	864.20
7	682.50	785.33	822.79	681.35	894.31
8	735.39	783.64	750.91	819.75	824.50
9	818.29	773.41	752.49	761.88	865.08
10	833.88	781.67	783.87	764.51	908.70

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	25.20	25.49	25.24	26.23	27.43
2	26.36	27.00	24.85	23.85	30.48
3	24.31	26.36	23.86	25.70	23.94
4	23.49	23.82	27.40	24.42	28.51
5	24.67	24.12	24.38	28.44	28.95
6	24.25	24.43	27.02	24.50	27.16
7	22.96	25.22	26.40	22.46	28.06
8	23.59	25.16	24.05	26.35	25.89
9	26.36	24.82	24.08	24.36	27.18
10	26.81	25.10	25.15	24.52	28.55

Resultados Tomate del día viernes

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	791.99	796.81	772.86	814.48	817.50	811.68
2	772.28	807.02	781.52	798.50	780.08	807.50
3	788.39	738.54	792.71	697.73	797.96	803.36
4	789.17	757.59	736.32	790.71	826.64	812.45
5	763.51	789.37	749.06	820.33	764.97	732.72
6	753.40	651.97	690.17	764.86	803.81	883.79
7	780.84	770.47	729.25	807.56	687.44	818.47
8	775.81	757.49	740.61	814.40	867.90	814.16
9	750.85	715.87	771.19	779.05	789.52	815.46
10	780.25	759.05	777.11	823.19	823.87	830.59

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	26.72	27.04	26.26	27.50	27.69	27.51
2	26.04	27.38	26.60	26.96	26.87	27.37
3	26.58	25.73	26.92	26.90	27.25	27.52
4	26.62	25.72	26.06	26.74	27.92	27.55
5	25.73	26.85	25.89	27.58	27.41	27.72
6	25.44	27.64	26.78	25.97	27.34	29.70
7	26.34	26.17	26.17	27.26	28.12	27.62
8	26.37	25.75	25.99	27.46	29.21	27.53
9	25.51	26.12	26.20	26.30	26.96	27.57
10	26.32	25.76	26.41	27.78	27.87	28.04

Resultados Yuca del día martes

RMSE

N	1 Var	2 Var	3 Var	4 Var
1	184.55	165.30	56.65	130.55
2	162.34	132.18	66.76	82.91
3	233.99	162.71	84.64	141.67
4	218.83	173.48	118.23	254.29
5	227.19	146.12	98.81	166.90
6	224.51	207.33	128.32	268.75
7	247.11	248.78	126.28	164.04
8	231.66	257.66	114.21	202.12
9	183.28	174.62	127.21	209.48
10	212.68	183.28	110.92	161.53

MAPE

N	1 Var	2 Var	3 Var	4 Var
1	21.27	19.35	5.91	15.35
2	18.83	15.61	7.07	9.40
3	27.19	19.06	9.73	16.61
4	25.31	20.25	13.91	29.43
5	26.33	17.20	11.57	19.38
6	26.01	24.26	15.08	31.11
7	28.74	29.04	14.85	19.08
8	26.87	30.14	13.44	23.46
9	22.09	20.37	14.95	24.20
10	24.57	22.09	13.05	18.83

Resultados Yuca del día jueves

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	58.88	63.50	79.99	52.07	95.90
2	56.40	43.01	76.55	83.39	117.75
3	79.73	50.84	62.78	66.95	87.98
4	53.06	111.07	72.14	91.66	101.92
5	71.73	106.11	48.71	69.52	119.01
6	80.65	98.43	66.15	99.81	124.82
7	101.06	62.13	59.21	94.16	99.74
8	79.64	116.54	62.08	61.10	130.10
9	84.68	99.41	84.18	83.46	123.38
10	65.07	105.32	100.55	103.18	93.60

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var
1	5.44	7.53	9.72	5.44	11.72
2	5.08	5.23	9.25	10.12	14.49
3	8.13	6.06	7.52	7.74	10.54
4	4.33	13	8.43	11.27	12.49
5	7.18	12.51	5.76	8.03	14.70
6	8.23	11.64	7.71	12.35	15.43
7	10.44	7.35	6.91	11.61	12.19

8	8.14	13.69	7.24	6.83	16.11
9	8.83	11.78	10.26	10.16	15.27
10	6.38	12.38	12.32	12.76	11.30

Resultados Yuca del día viernes

RMSE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	60.99	28.18	27.83	34.97	27.61	28.99
2	67.43	72.35	59.72	29.70	24.52	52.94
3	75.73	24.54	70.17	30.56	32.68	28.17
4	99.46	94.50	37.49	52.60	46.07	30.09
5	77.56	72.27	84.94	71.29	35.99	46.77
6	90.71	91.41	39.97	57.23	83.65	99.21
7	75.58	68.09	31.82	54.33	22.08	91.54
8	92.40	78.57	30.98	64.20	41.86	66.96
9	66.71	48.28	55.81	53.04	43.60	71.71
10	86.62	97.71	56.39	30.99	19.89	37.99

MAPE

N	1 Var	2 Var	3 Var	4 Var	5 Var	6 Var
1	6.42	2.92	2.14	4.03	2.61	2.92
2	7.24	7.48	6.93	3.01	2.43	6.36
3	8.42	2.44	8.09	3.76	3.87	2.86
4	10.94	9.66	4.45	6.18	5.43	3.07
5	8.57	7.84	9.35	7.83	4.03	5.48
6	10.05	9.67	4.62	6.81	9.51	12.24
7	8.33	7.16	3.74	6.35	2.24	11.29
8	10.22	8.16	3.69	7.56	4.71	8.31
9	7.22	4.97	6.23	6.24	4.27	8.80
10	9.61	10.12	6.52	3.51	2.15	3.99

APÉNDICE D

Certificado de participación del estudiante Juan David Márquez González en el evento “Día del investigador” realizado el 31 de octubre del 2018.

Certificado de participación de la estudiante Laura Vanessa Montaño Pérez en el evento “Día del investigador” realizado el 31 de octubre del 2018.

APENDICE E

Certificado de participación y finalista en el XII encuentro de semilleros de investigación “El Intercambio del Conocimiento: Base para el Desarrollo” del estudiante Juan David Márquez González.

Certificado de participación y finalista en el XII encuentro de semilleros de investigación “El Intercambio del Conocimiento: Base para el Desarrollo” de la estudiante Laura Vanessa Montaño Pérez.