Universidad Católica de Santa María

Facultad de Ciencias e Ingenierías Físicas y Formales

Escuela Profesional de Ingeniería de Sistemas

PRONÓSTICO DE COSECHA DE CEBOLLA ROJA MEDIATE EL USO DE REDES NEURONALES PARA LA MEJORA DE LA RENTABILIDAD EN AREQUIPA -2021.

Tesis presentada por la Bachiller:

Lazo Portugal, Ramiro Franchesco

Para optar el Título Profesional de

Ingeniero de Sistemas con especialidad en Ingeniería del Software

Asesor:

Dr. Sulla Torres, José Alfredo

Arequipa- Perú

2022

UCSM-ERP

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

INGENIERIA DE SISTEMAS

CON ESPECIALIDAD EN INGENIERIA DEL SOFTWARE

TITULACIÓN CON TESIS

DICTAMEN APROBACIÓN DE BORRADOR

Arequipa, 30 de Julio del 2022

Dictamen: 003963-C-EPIS-2022

Visto el borrador del expediente 003963, presentado por:

2015201181 - LAZO PORTUGAL RAMIRO FRANCHESCO

Titulado:

PRONÓSTICO DE COSECHA DE LA CEBOLLA ROJA MEDIANTE EL USO DE REDES NEURONALES PARA LA MEJORA DE LA RENTABILIDAD EN AREQUIPA -2021

Nuestro dictamen es:

APROBADO

1561 - GUEVARA PUENTE DE LA VEGA KARIMDICTAMINADOR

1568 - ROSAS PAREDES KARINA DICTAMINADOR

1910 - CASTRO GUTIERREZ EVELING GLORIADICTAMINADOR

DEDICATORIAS

La presente investigación la dedico principalmente a mis padres, por los años de trabajo, sacrificio y amor que me ayudaron a llegar hasta aquí. Valoro todo lo que me han enseñado y lo que soy se los debo a ustedes. Son mi principal motivación y los responsables de todos mis éxitos.

Agradezco a todos los docentes que me enseñaron a lo largo de mi vida universitaria, pues gracias a su conocimiento y apoyo me ayudaron a crecer como persona y como profesional.

AGRADECIMIENTOS

Agradezco al Ing. Jose Sulla y al Mg. Ricardo Valdez por su asesoría y su apoyo.

Agradezo a todas las personas que me guiaron y apoyaron para la elaboración de la presente investigación.

RESUMEN

En este trabajo de investigación se propone una serie de recomendaciones que le permita a los agricultores dedicados al cultivo de cebolla roja en Arequipa mejorar la rentabilidad del cultivo de esta hortaliza y evitar mayores pérdidas a los mismos. Para esta investigación se siguieron las fases de una metodología de minería de datos (Cross-Industry Standard Process for Data mining) para la gestión de la información que será procesada y que servirá para la elaboración de las recomendaciones finales.

Se obtuvo los datos de cosecha de hace 10 años hasta la actualidad en un formato .xls desde los instrumentos de gestión, estadísticos y de producción del MIDAGRI (Ministerio de Desarrollo Agrario y Riego) para luego ser preprocesados mediante una herramienta para el aprendizaje automático y minería de datos. Posteriormente se realizó el preprocesamiento de estos datos y su migración a una base de datos para su procesamiento.

Para la creación de un modelo de pronóstico se definieron objetivos de minería de datos que se encuentran alineados con los objetivos propuestos para la investigación y por cada uno de estos objetivos se crearon modelos que luego fueron evaluados para su posterior mejora en términos de calidad y parametrización con la herramienta de minería de datos escogida para el modelamiento. Se analizaron los resultados obtenidos de la ejecución de los modelos para poder confirmar o descartar las teorías propuestas en la investigación acerca de las variables que afectan la rentabilidad de la siembra y cosecha de cebolla roja en la región Arequipa.

Finalmente se propusieron algunas formas de aplicar la información obtenida a la práctica para que esta sea utilizada en beneficio de los agricultores arequipeños y se pueda considerar la creación de planes de cosecha para evitar la escasez o sobreproducción de cebolla.

Palabras claves: inteligencia artificial, redes neuronales, cebolla roja, cosecha, predicción, modelo

ABSTRACT

In this research work, a series of recommendations is proposed that allows farmers dedicated to the cultivation of red onion in Arequipa to improve the profitability of the cultivation of this vegetable and avoid greater losses to them. For this research, the phases of a data mining methodology (Cross-Industry Standard Process for Data mining) were followed for the management of the information that will be processed and that will serve for the elaboration of the final recommendations.

Harvest data from 10 years ago to the present was obtained in an .xls format from the management, statistical and production instruments of the MIDAGRI (Ministry of Agrarian Development and Irrigation) to later be pre-processed using a tool for automatic learning and data mining. Subsequently, this data was pre-processed and migrated to a database for processing.

For the creation of a forecast model, data mining objectives were defined that are aligned with the objectives proposed for the investigation and for each of these objectives' models were created that were then evaluated for their subsequent improvement in terms of quality and parameterization. with the data mining tool chosen for the modeling. The results obtained from the execution of the models were analyzed in order to confirm or discard the theories proposed in the research about the variables that affect the profitability of the sowing and harvesting of red onion in the Arequipa region.

Finally, some ways of applying the information obtained to the practice were proposed so that it is used for the benefit of Arequipa farmers and the creation of harvest plans to avoid the scarcity or overproduction of onion can be considered.

Keywords: artificial intelligence, neural networks, harvest, red onion, prediction

ÍNDICE

INTRODUCCIÓN

La cebolla es una de las hortalizas más cultivadas en todo el mundo debido a la existencia de grandes variedades de esta que son fácilmente adaptables a una gran diversidad de climas. En el Perú, por campaña, se destina alrededor de 17070 ha. de terreno para la siembra de esta hortaliza y la superficie cosechada es de 16817 ha. con un rendimiento promedio de 39.3 toneladas por hectárea (SIEA,2020). La variedad de cebolla que se comercializa en mayor cantidad en el Mercado Mayorista de Lima (el mercado Mayorista más grande de Lima-Perú y el destino de la mayor parte de cosechas cebolleras) es la cebolla roja o morada.

Arequipa se ubica como el primer productor de cebolla en todo el Perú con una participación mayor al 50% en el país y una producción de 332.5 mil toneladas por campaña, convirtiéndose así en el principal productor de esta hortaliza en el Perú. Además, el 90 % del volumen de cebolla cosechada en Arequipa por campaña es de una de las variedades más populares en el mercado nacional: la cebolla roja.

El alto grado de conservación que posee la cebolla roja, su adaptabilidad a diversos climas y la gran cantidad de usos que tiene, han aumentado su popularidad en la última década, lo que la convierte en una de las hortalizas más aptas para cultivar en Arequipa. Sin embargo, el principal problema que enfrentan los agricultores arequipeños para el cultivo de esta variedad de cebolla es el bajo precio de venta cuando esta es ofrecida a los mercados locales y nacionales afectando la estructura económica y productiva de esta variedad de cebolla (Bermudes, 2019). Sumado al precio de venta, el costo de inversión destinado al cultivo de esta variedad de cebolla ocasiona que en algunos años haya una sobreproducción de cebolla mientras que en otros el volumen producido sea muy bajo, ocasionando así una gran pérdida para los agricultores.

Otro factor con el que los agricultores arequipeños deben lidiar en la época de cosecha es el clima: la temporada principal de cosecha de la cebolla roja empieza en el mes de diciembre hasta abril, fecha en la que las precipitaciones fluviales llegan a la región lo que provoca que el ritmo de cosecha de la hortaliza se acelere para evitar que la producción se malogre por el contacto con el agua.

Dada la coyuntura, muchos agricultores optan por acopiar y guardar sus productos a la espera de un mejor precio y con el fin de evitar que su cosecha se vea afectada por la lluvia. Esto significa un

costo adicional de inversión, que como ya se mencionó anteriormente, puede ocasionar una gran pérdida económica para los agricultores debido a una posible sobreproducción o una baja producción de cebolla.

Por lo tanto, tener conocimiento de algunas de estas variables en una época temprana de cosecha o incluso en la etapa de siembra de esta variedad de cebolla conocida como roja o morada, permitirá que los agricultores dedicados al cultivo de esta hortaliza puedan contar con información que les ayude a la creación de un plan de siembra y a evitar mayores pérdidas económicas de este sector agrícola. Si esta información es utilizada por los agricultores, se podría definir un precio base que permita mejorar la rentabilidad del cultivo de esta hortaliza y mejorar la estructura económica de la misma.

El trabajo de investigación se divide en 4 capítulos en los que se tratan los siguientes temas:

En el **capítulo 1**, se presenta el marco teórico del proyecto, estado del arte y las bases teóricas de la investigación, estudio económico definido según el presupuesto de actividades de la metodología, gestión de costos del proyecto.

En el **capítulo 2**, se realiza el diseño e implementación de la propuesta metodológica, las fases correspondientes y su aplicación con el problema del negocio.

En el **capítulo 3**, se muestra los resultados de haber aplicado la metodología planteada además de una discusión sobre los resultados obtenidos.

Finalmente se definen las conclusiones y recomendaciones para trabajos futuros.

CAPÍTULO I PLANTEAMIENTO TEÓRICO

Objetivos de la investigación

1. General

Generar un modelo de red neuronal para estudiar la relación entre las variables de cosecha de cebolla roja y su precio en el mercado, usando datos reales de la zona agrícola El Cural, en Arequipa durante el periodo de cosecha 2021- 2022.

2. Específicos

2.1.1 Seleccionar de las variables de entrada que determinarán el mejor modelo de predicción (climatológicas, geográficas, etc. entre otros).

2.1.2 Analizar el precio de venta de cebolla roja obtenidos producto de la predicción para revelar la rentabilidad de la siembra de esta hortaliza.

2.1.3 Estimar el porcentaje de cosecha que es descartado por factores naturales (enfermedades y clima) que afectan el desarrollo normal del bulbo de cebolla ocasionando un déficit en la producción total de esta hortaliza.

2.1.4 Analizar los resultados obtenidos para estudiar su interpretabilidad para su evaluación y elaboración de reglas.

Justificación e importancia

2.1. Justificación Metodológica

En el aspecto metodológico, el presente trabajo de investigación aporta herramientas para el desarrollo de un modelo de predicción de cosecha de cebolla roja utilizando redes neuronales. El propósito es el de poder tener conocimiento del volumen de producción de cebolla y el precio correspondiente por campaña yasí establecer una relación entre las variables de cosecha de cebolla roja y el respectivo precio en el mercado local y/o nacional.

2.2. Justificación Temática

La cebolla es una de las hortalizas con uno de los más altos niveles de importanciaeconómica después de la papa, además de poseer un valor social inestimable y seruna de las verduras más consumidas por todo el mundo (Amaya y Méndez, 2013).La cebolla en el Perú se constituye como un elemento imprescindible en la gastronomía peruana y su alto nivel de consumo debería traducirse de forma positiva en el crecimiento del sector agrícola y para todos los participantes de estacadena.

2.3. Justificación Social

El presente trabajo se justifica socialmente de acuerdo con el objetivo de favorecer a los agricultores arequipeños que se dedican al cultivo de cebolla roja mejorando su situación económica que muchas veces se ve amenazada por el elevado costo de inversión que supone el cultivo de esta hortaliza.

De igual manera, pretende favorecer a toda la comunidad dedicada al cultivo de cebolla roja para transformarla en un grupo innovador, que se encuentre enfocado al buen desempeño del agricultor y por el cual busca beneficiar a quienes serán los consumidores finales ya que recibirán productos de mejor calidad.

2.4. Justificación Económica

Bermudes (2019) indica que el precio de la cebolla no ha tenido un aumento muy significativo a lo largo de los últimos 6 años debido al comportamiento variable de este por la sobreproducción o baja producción de cebolla que perjudica a los participantes de la cadena productiva y económica de esta hortaliza. El precio de la cebolla en el Mercado Mayorista de Lima durante los últimos 6 años continua con la tendencia ascendente de Bermudes (2019) pero con el mismo comportamiento variable: El precio promedio en el 2015 fue S/ 1.23/Kg, en el 2016 fue S/ 0.67/Kg, en el 2017 fue de S/ 1.31/Kg, en el 2018 y 2019 tuvo un precio de S/ 0.88/Kg y S/ 1.27/Kg respectivamente y en el 2020 el precio promedio de la cebolla después de la última campaña de cosecha alcanzó los S/ 1.09/Kg.

Descripción de la solución

Durante la investigación se estudiará a profundidad las variables de cosecha escogidas: Climatológicas (porcentaje de precipitaciones fluviales, humedad, calor y frío), Naturales

(enfermedades y plagas), Estadísticas (volumen de cebolla cosechada, volumen de cebolla descartada, área de cebolla sembrada) y de Tipo (riego por goteo o gravedad). De ser necesario, se agregarán más variables que permitan darle mayor precisión al modelo.

Esta información será recolectada con ayudo de expertos en el área agrícola y de fuentes estadísticas del mismo sector. Como el modelo deberá recibir datos de entrada para su entrenamiento, se utilizarán datos históricos de los últimos 10 años con las variables de cosecha ya mencionadas.

Después de haber configurado los parámetros necesarios para el modelo y este haya sido entrenado con los datos históricos se procederá a la obtención de las reglas que ayudarán al análisis y elaboración de conclusiones y recomendaciones finales.

Aporte

El propósito de esta investigación es el de poder brindar más información a los agricultores arequipeños dedicados al cultivo de cebolla que les permita proponer mejores estrategias para enfrentar la volatilidad del precio de esta hortaliza en el mercado nacional y de esta forma contribuir con un aporte económico significativo para sus productores.

Preguntas de investigación

2.5. Pregunta general

- ¿Se podrá crear un modelo de red neuronal de cosecha de cebolla roja para encontrar las causas que generan los bajos precios de esta hortaliza en los últimos años utilizando datos de cosechas de campañas pasadas?

2.6. Preguntas específicas

- ¿Es posible utilizar las variables de entrada seleccionadas para la implementación y entrenamiento del modelo de pronóstico de cebolla?

- ¿Se podrá analizar los datos obtenidos de la predicción de cebolla para poder revelar la rentabilidad de esta hortaliza?
- ¿Será posible estimar el porcentaje de cosecha descartado por factores naturales como enfermedades, plagas y factores climatológicos que representan un déficit en la producción total de cebolla roja?
- ¿De qué forma el análisis de los resultados obtenidos servirán para estudiar su interpretabilidad para su posterior evaluación y elaboración de reglas?

FUNDAMENTOS TEÓRICOS

ESTADO DEL ARTE

En el trabajo de José A. Martínez-Casasnovas y Xavier Bordes Aymerich del 2005 nos ubicamos en España, donde la viña se ha convertido en uno de los cultivos más importantes debido al auge que tiene este producto y a los grandes avances que se han implementado para poder obtener un producto de gran calidad para la elaboración de vinos. La utilización de tecnologías de la información geográfica para la obtención de información de la variedad de cultivo y las variables de producción representan una nueva estrategia para la obtención de una mayor cosecha y minimizar el impacto ambiental (Martínez y Bordes 2005).

Para el trabajo de predicción de cosecha de la vid realizado entre la Universidad de Lleida y la finca de la empresa Codorniú en España, se deseaba poder predecir el rendimiento de distintas variedades de viñas. Para ello se utilizó el número de brotes, número de racimos, peso de poda, vigor de cultivo y cosecha de un año previo como las variables de cultivo que ayudarían a estimar el rendimiento de diferentes variedades de viña de forma localizada

En este modelo, las variables ya mencionadas fueron muestreadas de acuerdo con un marco de 10 x 10 cepas. El objetivo de este procedimiento es el de utilizar un método geoestadístico para la obtención de una cobertura continua de las variables de entrada y de esta forma poder predecir dichos valores. Para la elaboración del modelo de predicción se usó una técnica de regresión multivariante donde se considera también el uso de un mapa de cosecha por variedades del año anterior. Además del mapa de cosecha utilizado por Martínez y Bordes (2005) se utilizó una imagen de Quickbird 2 del año 2014, fecha en la que se iniciaba la etapa de maduración de las viñas.

Para Martínez y Bordes (2005) los resultados obtenidos del modelo de predicción indican la posibilidad de obtener modelos con una alta correlación donde intervengan, además de las variables de vigor de los cultivos, los determinantes de fertilidad como el número de brotes, el número de racimos y el número de yemas (variables de cultivo).

La aplicación de las redes neuronales ha ido ganando mucha aceptación en la industria de los alimentos. Técnicas como la clasificación, el reconocimiento de patrones y la predicción de cosechas resultan de gran utilidad para poder estimar el rendimiento de diferentes productos agrarios. Su utilización ha permitido mejorar la calidad de los productos y el modo en el que se cultivan. En el trabajo de Gustavo Andrés Figueredo y Ávila Javier Antonio Ballesteros-Ricaurte acerca de identificación del estado de madurez de las frutas con redes neuronales artificiales, se presentan los conceptos básicos de las redes neuronales para el procesamiento de imágenes aplicado a la revisión y clasificación de las frutas en Colombia.

Antes que nada, se puede definir una RNA como un sistema adaptable a las necesidades de un contexto que puede realizar paralelamente varios procesos para mezclar técnicas que ayudan al procesamiento de la información. Como su nombre lo indica, su característica principal, pretende imitar el comportamiento del cerebro y de las características funcionales de las redes neuronales biológicas (Figueredo y Ávila, 2015). A lo largo de los años han aparecido nuevos modelos de redes neuronales lo que ha permitido el desarrollo de sistemas más complejos y eficientes. La creación de nuevos modelos de redes neuronales se debe principalmente a las necesidades que ha traído la evolución de la tecnología, nuevos problemas matemáticos y la inagotable necesidad de los humanos por comprender la naturaleza (Figueredo et al., 2015).

En esta investigación se trabajaron con tomates, olivos, manzanas, lechugas y pimientos dulces. Cada uno de estos productos fueron trabajados con técnicas de predicción de diferentes autores de modelos de redes neuronales. Por ejemplo, para las manzanas, se guiaron del modelo de red neuronal que buscaba predecir la pérdida de agua y la ganancia de sólido de la manzana.

Para este modelo, se trabajan con seis variables: tiempo de inmersión, área de superficie, ratio de la masa de la fruta, temperatura, nivel de concentración de la solución y nivel de agitación.

Figueredo y Ávila (2015) revelaron la gran cantidad de investigaciones referidas a la aplicación de redes neuronales artificiales y algoritmos para la inspección, clasificación y predicción de los cambios en diversas frutas y hortalizas.

Para este trabajo se utilizaron como referencia muchos modelos de redes neuronales como una opción para la clasificación, las cuales fueron eficientes y efectivas.

Los autores de muchos de estos modelos consideran que estos dependen bastante de los datos de entrada y de su respectivo tratamiento para la arquitectura del modelo y la selección de los algoritmos de entrenamiento. Estos elementos aseguran una mejor clasificación, por lo que la calidad de estos podrá determinar el tipo de clasificación que se obtendrá de acuerdo con el manejo que se les haya dado.

Es así como las redes neuronales se han convertido en una de las herramientas más usadas en la actualidad por su capacidad de otorgar a los algoritmos la posibilidad de pensar y aprender como si fuera un cerebro y llevar a cabo las acciones de uno. Para Aldo Peláez Toledo, una de las ventajas más importantes es la gran robustez para soportar el ruido en varias industrias con escenarios desfavorables lo que las convierte en una tecnología dinámica capaz de adaptarse a condiciones cambiantes (Peláez Toledo, 2019).

En su trabajo: Redes neuronales artificiales, una herramienta útil para el procesamiento de datos de cosecha, aborda el caso de una cosechadora CASE IH-A8000 que posee un alto nivel de automatización gracias a la gran cantidad de sensores que posee lo que le permite el seguimiento de muchas variables de cultivo de gran importancia para el análisis de información de cosecha. Como era de esperarse, la cantidad de datos de la cosecha al utilizar la CASE IH-A8000 son enormes: Más de 35000 datos de un aproximado de 50 variables en solo dos días de uso.

Cabe considerar, por otra parte, que el número de cosechadoras de este modelo se han incrementado y a pesar de que el tratamiento de la información y datos de cosecha no es común en el país, en la zona donde se realizó esta investigación han aparecido ciertos indicios de un avance en este campo. Para el inicio del procedimiento se escogieron 52 variables de las cosechadoras CASE IH-A8000 en la biblioteca de aprendizaje supervisado (WEKA) para a continuación seleccionar los atributos con la opción del software que permite explorar usar cada conjunto de datos para utilizar diferentes arquitecturas para las pruebas.

Para la selección de los datos se hace uso de “evaluadores”, cada uno con distintos métodos de búsqueda. El primer evaluador utiliza un método que tiene como característica principal la búsqueda de los atributos en sus respectivos subespacios con un algoritmo de búsqueda hacia arriba aumentada con rastreo hacia atrás. El segundo evaluador, se combina con el anterior llamado “GreedyStepwise” para la búsqueda en el espacio de atributos.

El tercero utiliza el método "Ranker" para ordenar los atributos y el método "ClassifierAttributeEval" para evaluar el valor de un atributo dado por un usuario. El cuarto y quinto evaluador utilizan el método "CorrelationAttributeEval" y "PrincipalComponents" respectivamente.

Utilizando los resultados de estos evaluadores se evalúan los grupos de datos para 2 ambientes diferentes. Estos ambientes son las estructuras que tienen 3 neuronas en las capas ocultas y la restante con 2 capas de 3 neuronas. En esta parte se introdujeron solo datos numéricos en el WEKA para proceder al procesamiento con Matlab.

Estos conjuntos de datos de entrada se hicieron con un método inductivo para poder seleccionar variables únicamente numéricas y usando el método Levenberg-Marquardt. Los resultados que se obtuvieron en Matlab son ligeramente diferentes a los obtenidos en WEKA, pues las variables son diferentes en ambos casos. Peláez (2019) llega a las siguientes conclusiones de su trabajo: el método que se utilizó en Matlab es una mejora del "backpropagation" usado en el WEKA, aunque este último ofrece mejores resultados por la alta correlación que existe.

El 70% de las RNA usan MLP con backpropagation o Levenberg-Marquardt. La mayoría de estas RNA usan Matlab o WEKA. En el caso de este último, su popularidad se debe principalmente a que es posible implementar varios métodos de selección de variables para entrenar una RNA.

En este tipo de investigaciones, el paso para la selección de atributos es fundamental antes del procesamiento de la información.

Centrándonos en el contexto nacional, el estudio de Luis Burgos Chinchay y Judith Mendoza Vallejos (2018) aborda un análisis de todos los participantes de la cadena de producción brindándole una mayor importancia al papel del agricultor.

Este trabajo utiliza como contexto el norte del país y toma como fuentes de información secundarias a los estudios y estadísticas del Perú para establecer una hipótesis sobre el comportamiento y la realidad del agricultor que se encuentra dedicado a la comercialización de esta hortaliza (Mendoza & Burgos, 2018).

El presente estudio aborda el sector del cultivo de cebolla roja en el Perú y realiza un análisis de todos los participantes de la cadena de producción brindándole una mayor importancia al papel del agricultor. Este trabajo utiliza como contexto el norte del país y toma como fuentes de información secundarias a los estudios y estadísticas del Perú para establecer una hipótesis sobre el

comportamiento y la realidad del agricultor que se encuentra dedicado a la comercialización de esta hortaliza (Mendoza & Burgos, 2018).

Para este trabajo se definieron 5 factores de riesgo que afectan a la cadena de producción de cebolla roja. Mendoza y Burgos (2018) los definen como fuerzas las fuerzas competitivas y son las siguientes:

1. **La capacidad de negociación entre los compradores y los clientes:** En el caso del cultivo de la cebolla roja, los clientes comunes son los distribuidores mayoristas que son los encargados de acopiar la cebolla. Son estos quienes tienen más poder de negociación ya que son los que definen el precio y quienes tienen mayor influencia en la conformación del precio dentro de la estructura de comercialización.
2. **La capacidad de negociación de los proveedores:** Los proveedores principales de insumos para el cultivo de cebolla roja encontramos a los vendedores de fertilizantes y las empresas dedicadas a la venta de semillas. En el primer caso, al tratarse de un *comodity*, es decir son bienes primarios y en cualquier otra parte se encontrará al mismo precio y la misma calidad. En el caso de la obtención de las semillas es un panorama distinto: El financiamiento para las semillas en la mayoría de los casos es propio. Muchos de los agricultores rechazan la variedad de semilla híbrida. Es más sencillo para los agricultores producir su propia semilla e incluso comprarla a otros por un precio menor de casi 50% menos.
3. **Nuevos competidores:** Hay la posibilidad que los ingresos sean bajos y existan nuevos interesados en entrar al negocio del cultivo de la cebolla roja. Los vendedores mayoristas creen que tienen riesgo de sufrir una pérdida más significativa en caso de que el precio bajara, pues perderían su capital y lógicamente el precio de venta no sería el más alto para recuperar el gasto. La inversión en el cultivo de cebolla es baja, lo que hace más alta la probabilidad de tener nuevos competidores.
4. **Productos sustitutos:** En el mercado, la cebolla es única por su sabor ya que representa la base de la comida nacional. Sin embargo, dependiendo de los márgenes de los productores, estos pueden optar por la siembra de otros cultivos dependiendo del mercado. A pesar de esto, la demanda de cebolla siempre existirá (al menos localmente), el riesgo es bajo.

5. **Diferencias entre los competidores:** El inicio de toda la cadena de producción de la cebolla roja se origina en los productores. Los agricultores no poseen transporte propio para llevar sus productos a un mercado local y normalmente venden en chacra cuando los vendedores minoristas o mayoristas están interesados en su producto. Los agricultores dedicados al cultivo de cebolla roja por lo general tienen muchos años de experiencia y siempre están en la búsqueda de nuevas formas de siembra, cosecha y de aplicar otras metodologías para mejorar la calidad de la cebolla. Es por esta razón que siempre buscan la forma de cultivar con la menor inversión posible. Como se ha mencionado anteriormente, el agricultor se limita a cosechar su producto y venderlo en chacra y no es capaz de dar un paso más para integrarse, lo que limita aún más su poder de negociación, que ya es bastante bajo de por sí.

El precio de la cebolla y sus volúmenes son los factores determinantes para saber que participante de la cadena se quedará finalmente con el mayor valor.

Pero casi siempre, el actor más perjudicado por el bajo precio es el agricultor. Sin embargo, quien aporta un mayor valor competitivo al producto es el acopiador por la influencia que tiene para determinar el precio final.

Mendoza y Burgos (2018) sostienen que la asociación de los agricultores podría permitir una integración vertical para que estos mismos se conviertan en acopiadores para darle un valor agregado al producto mediante la mejora de procesos y la transferencia de tecnología. Esto incluso les permitiría a los agricultores decidir a quienes venden y a quienes no o inclusive decidir en qué licitaciones deberían participar.

En el campo del reconocimiento de imágenes también se ha logrado un gran avance. Algoritmo para classificação de plantas de milho atacadas pela lagarta do cartucho es el proyecto de investigación de Darly G. de Sena Júnior, Francisco de A. de C. Pinto, Daniel M. de Queiroz, Evandro C. Mantovanies para procesar y posteriormente analizar las imágenes de las plantas de maíz atacadas por plagas de insectos. Normalmente para el control de plagas se usan técnicas para poder monitorear la evolución de estas para poder crear medidas de respuesta en contra de la población de plagas (De Sena, Pinto, De Queiroz, Mantovani, 2001). Se usan también técnicas de visión artificial con imágenes de alta resolución que son obtenidas de cámaras instaladas previamente en equipos y maquinaria agrícola.

Para este tipo de sistemas, se debe integrar ordenador junto con las señales de una cámara las cuales son procesadas por los algoritmos para extraer información de las imágenes que va capturando. Sin embargo, es complicado poder segmentar la imagen de un campo de cultivo por su complejidad: tamaño, color y forma. Para estos casos De Sena et al. (2001) se apoyaron de diversas investigaciones y métodos para poder determinar las condiciones del terreno, el estado del cultivo, plagas, estrés hídrico, entre otros. Para poder reducir el tiempo de procesamiento y los requisitos computacionales se subdividió el procesamiento de imágenes digitales en partes: una fase de clasificación de baja resolución y la otra etapa de clasificación con alta resolución.

La clasificación con alta resolución tenía el objetivo de eliminar las porciones de imagen que contenían muchos píxeles, residuos de suelo y/o hojas.

La imagen originalmente tenía 480 filas y 640 columnas. Se descartaron gran cantidad de cada lado para reducir la dimensión de los bloques. Las láminas fotosintéticas absorben la energía electromagnética lo que permite que se refleje el color del suelo y las hojas atacadas por la oruga. En la clasificación con baja resolución se subdividieron los bloques de la primera clasificación. Cada uno de estos fue subdividido por medio de una red neuronal artificial utilizando como datos de entrada las características del color producto de las clasificaciones.

Para probar el algoritmo se utilizó la imagen de una planta de maíz afectada por el ataque de una oruga. La imagen fue procesada con el algoritmo desarrollado para evaluar la eficiencia de clasificación de imágenes. Para este caso De Sena et al. (2001) utilizó una matriz de errores o tabla de contingencia para identificar el error general entre todas las categorías.

Como conclusiones de este trabajo, se consideró que, a mayor número de neuronas en las capas ocultas, se obtendrá un mejor resultado con las pruebas de entrenamiento. El método de recuento de píxeles que representaban las hojas que no habían sido afectadas por plagas fue de gran utilidad para la siguiente clasificación, teniendo un menor número de píxeles para la siguiente clasificación.

Siguiendo en el campo del aprendizaje automático y haciendo una revisión a las técnicas de aprendizaje orientadas a la predicción de variables, analizamos la investigación de Fernando Villada, Nicolás Muñoz y Edwin García-Quintero, que busca predecir el comportamiento de las variables de precio del oro utilizando redes neuronales. Esta investigación aborda problemas tales como el robo de las reservas de metal, precio del petróleo, índices de los mercados financieros que afectan el

precio del oro. Los datos extraídos fueron extraídos de Londres (LBMA, 2015) del mercado del oro y de la plataforma Bloomberg para los precios de petróleo para Estados Unidos y datos del índice del dólar americano.

El modelo de red neuronal utilizada fue de propagación hacia adelante (Villada, Muñoz y García, 2016), por las características que ofrece como aproximador universal. Pero para este punto se probaron diferentes arquitecturas de una capa oculta con la misma cantidad de neuronas producto del resultado igual a la semisuma del número total de entradas y el número de salidas. El proceso de entrenamiento para la investigación consistió en la incrementación de la cantidad de neuronas de la capa oculta hasta obtener la mejor arquitectura. Para la selección de la mejor arquitectura se consideraron algunas medidas de desempeño como el error absoluto promedio porcentual y la raíz cuadrada del error promedio cuadrático. Estos dos valores son los más utilizados en investigaciones dedicadas a pronósticos.

El proceso de entrenamiento es de un tipo de aprendizaje supervisado y para la investigación se escogió el 62% de forma aleatoria para el entrenamiento y obtener el error mínimo deseado. Otro 20% fue utilizado para la validación y el 18% restante fueron utilizados para probar la estructura de red neuronal los que no fueron incluidos en el entrenamiento ni en la validación.

El objetivo principal de este trabajo de investigación fue el de atender las necesidades de operadores de mercado de oro mediante la predicción de su valor en los próximos 22 días. Los resultados de le ejecución de la RNA muestran que la variable de precio puede ser modelada satisfactoriamente utilizando la serie de precios como única variable de entrada. Al incluir una segunda variable de entrada fue el índice de del dólar estadounidense lo que le permitió a la red neuronal aumentar su desempeño y arroja un menor error en la predicción. La tercera variable de entrada se incluyó el índice de Standard and Poor 500 por ser una variable representativa en el mercado de acciones de Nueva York. Adicionalmente la estructura fue modificada nuevamente para incluir el índice DXY y la serie de precios del petróleo. Sin embargo, el mejor indicador de desempeño lo obtuvo el modelo con las primeras tres variables de entrada de los precios del oro, los índices de Standard and Poor 500 y DXY para los precios del petróleo.

Las configuraciones realizadas mostraron bajos errores en la partición de entrenamiento y en los de validación con una estructura de cinco y seis neuronas en la capa oculta. La utilización de una sola variable de entrada muestra una buena aproximación entre los datos reales y los obtenidos en la red.

Si bien los resultados mejoraron con la adición de la segunda variable, los errores disminuyeron considerablemente con las variables DYX de precios del petróleo y si bien los resultados son un poco más altos que los reportados, están dentro del rango admitido para el contexto. Con esto, Villada et al. (2016) confirman la aplicabilidad de las RNA para el mercado de bienes básicos y se convierte de esta forma en una herramienta muy útil para bancos, gobiernos, operadores ya que el modelo podría entregar señales de precios para la elaboración de planes de compras y ventas del metal con una mejor precisión.

Una investigación similar es la de Alfredo Bellido y Max Schwarz. Dicha investigación tuvo como finalidad utilizar una herramienta de aprendizaje automático para predecir el comportamiento de rendimiento y riesgo del conjunto de activos financieros (Bellido y Schwarz, 2019). Los datos de entrada utilizados en la investigación para la estimación de valores de rendimiento y riesgo de la cartera de acciones peruana desde el periodo 2010 hasta el 2016. La técnica utilizada para esta investigación fueron la de redes neuronales con un perceptrón multicapa con regresión con 3 capas ocultas: una con 21 nodos, una oculta con 85 nodos y 2 en la capa de salida. El algoritmo utilizado incluyó también una función de logística de activación con un optimizador LBFGS con una taza de aprendizaje de 0.01. Esto se utilizó para establecer patrones financieros, comerciales, operacionales para la predicción del comportamiento del mercado.

Las variables seleccionadas para el algoritmo son el precio del activo, el EBITDA, grado de liquidez, el coeficiente beta de riesgo, rotación del patrimonio, rentabilidad porcentual, rotación del activo entre otros. Para el caso de la rentabilidad porcentual se utilizó una rentabilidad real de los periodos 2010-2016 representando el corte de 25.9544%. La fuente de datos utilizada para la extracción de estas variables fueron las bases de datos de Economática, Bloomberg, la Superintendencia del Mercado de Valores peruana (SMV) y la Bolsa de Valores de Lima (BVL).

Dentro de los modelos generados, se encontró una red neuronal capaz de aproximar la predicción del rendimiento y riesgo con un 76% de eficacia para los activos seleccionados en el periodo 2010-2016. Producto de la investigación se concluye que la velocidad de rotación de los activos y la capacidad de generación de caja son las variables con mayor nivel de influencia para determinar las mejores combinaciones de rendimiento y riesgo para los activos financieros utilizados en la investigación, independientemente del mercado de operación.

Asimismo, existen algunas variaciones de acuerdo con el tipo de mercado en el que opera el activo y un nivel de correlación débil con respecto a la naturaleza del activo. Bellido y Schwarz (2019) concluyen que las redes neuronales pueden representar una herramienta para la estimación y predicción de las mejores combinaciones de rentabilidad y riesgo para los principales activos del mercado en base a los parámetros tradicionales que reportan los activos a los mercados y sobre los parámetros no tradicionales que se especifican en las compañías propietarias de dichos activos.

Pero son Julio Rojas y Víctor Vásquez quienes lograron utilizar redes neuronales para el campo agrícola. Este es el caso del yacón, un tipo de raíz que sirve como un tipo de alimento tradicional y que en varias partes de Sudamérica tiene propiedades que promueven la salud. Su popularidad ha traspasado fronteras y debido a nuevas regulaciones de los alimentos, se han realizado estudios a fin de comprobar si el yacón cumple con los requisitos reglamentarios y poder estar autorizado en Europa. Gracias al historial bien documentado, el uso seguro y su composición se llegó a la conclusión que no existe preocupación sobre algún efecto perjudicial o sustancia peligrosa.

La teoría del yacón como alimento que promueve a la salud se basa en la información existente que señala que esta raíz contiene fructooligosacáridos (FOS) que es un tipo particular de azúcares de baja digestión que brindan pocas calorías al organismo no elevando el nivel de glucosa en la sangre (Rojas y Vásquez, 2012). Asimismo, los FOS tiene otras características que ayudan a la prevención de caries, reducción de niveles séricos, colesterol y lípidos y su papel como estimulante de crecimiento de algunas bifidobacterias. De esta forma, Rojas y Vásquez (2012), desean saber la capacidad de predicción de una red neuronal aplicado al efecto de la concentración y temperatura de la solución de fructooligosacáridos en la masa de volumen, humedad y solidos en cubos de yacón osmodeshidratados. Las variables utilizadas para la generación del modelo de red neuronal y que afectan el proceso de osmodeshidratación son: la temperatura, el nivel de concentración de la solución, naturaleza del agente, presión, relación masa del producto y volumen de solución y agitación.

Para este proceso se preparó jarabe de inserción de FOS con el 5% de humedad para obtener las concentraciones deseadas. Las raíces de yacón fueron adquiridas en el mercado de Trujillo, La Libertad. Luego de ser lavadas, se utilizaron temperaturas de 30°, 40° y 50 ° con concentraciones de 30, 40, 50 y 60 % p/p de jarabe. Para el modelo se seleccionaron 2 variables de entrada (concentración y temperatura) y 6 variables de salida (masa, humedad, volumen, sólidos, difusividad

efectiva media con y sin encogimiento). Para la evaluación del modelo se sumergieron los cubos de yacón en un jarabe de FOS con temperaturas entre los 30° C a 45° C. De las 36 unidades experimentales se escogió de forma aleatoria 27 muestras de yacón para el entrenamiento, es decir el 75% y 9 unidades (25%) para la validación. Para el entrenamiento estos datos fueron normalizados y posteriormente el valor fue dividido entre el valor de la humedad. Para la investigación se utilizó un modelo de red neuronal de tipo Feedforwad, con los algoritmos de entrenamiento *Backpropagation* y ajuste de pesos *Levenberg-Marquardt*. Para la evaluación de la eficiencia de la red neuronal se utilizó el Error Cuadrático Medio (ECM).

La validación del modelo se realizó con la herramienta MatLab 7.0 para calcular los valores predichos por la RNA y definir los coeficientes de correlación entre los valores predichos con los esperados. Se desnormalizaron los valores predichos para evaluar la capacidad predictiva del modelo. Luego de haber probado 5 ecuaciones se utilizó el error porcentual como criterio de comparación entre los valores predichos por la RNA y los generados por la ecuación de regresión.

Con la RNA de tipo *Feedforward* con los algoritmos utilizados en la investigación para el entrenamiento y ajuste de pesos ha logrado predecir exitosamente los valores correspondientes a las variables de salida obteniendo como resultado global un error global promedio de 3.44 % y con un nivel de correlación mayores a 0.99 en comparación de las variables experimentales. Las mejores predicciones realizadas por el coeficiente de correlación y el error porcentual fueron la variable de masa (0.9977 y 1.68%) y para la variable de humedad (0.9988 y 1.64 %). El menor error cuadrático medio obtenido de la predicción de la red neuronal fue de 9.247×10^{-5}.

El problema de la rentabilidad de la cebolla roja es abordado por la investigación de Rocío Luz Medina Canto (2012): Ilabaya es un distrito conocido por producir orégano en la parte alta y por su producción de cebolla en la parte baja. La cebolla producida en la parte baja de Ilabaya es apreciada por los comercializadores debido a su resistencia al momento de ser transportada (Medina, 2012). En el caso de Ilabaya, se tiene un rendimiento de 37 toneladas por hectárea. Esta cifra es muy positiva pues el rendimiento nacional promedio de la cebolla es de 30 toneladas por hectárea.

Sin embargo, Medina (2012) indica que existe una falla en la cadena productiva de esta hortaliza lo que provoca la degeneración esta variedad de cebolla. Se definen 3 causas principales de esta falla: la mala selección de los bulbos, un mal manejo de semilleros y la gestión de la postcosecha. Además,

en esta zona de cultivo se utiliza tecnología de bajo nivel: se riegan los cultivos por gravedad, no se hace un control fitosanitario adecuado y de fertilización.

Sumado a todo esto, la condición económica de las familias dedicadas al cultivo de cebolla roja en Ilabaya es baja. Esta se ve afectada a su vez de los costos elevados, la volatilidad de los precios del mercado interno y la desorganización de los mercados. Los agricultores que tienen un nivel intermedio de desarrollo tecnológico tienen una ganancia mayor, pero esto dependerá siempre del precio establecido por el mercado y los vendedores mayoristas y minoristas. Se puede concluir también que la rentabilidad económica de la cebolla roja en Ilabaya está influenciada por costos de cultivo, los precios de venta y el rendimiento de los cultivos.

Esta última variable depende de enfermedades (raíz rosada, botrytis, entre otros) y plagas (gusano, trips, entre otros) que afectarán el rendimiento total del cultivo de cebolla pues disminuirán la producción y productividad de los cultivos. Medina (2012) establece una serie de recomendaciones luego de realizar un análisis de la rentabilidad de la cebolla roja de Ilabaya.

La mejora del proceso comercializador de cebolla roja en la cuenca del rio Locumba podrá impulsar una mejora en las relaciones de los productores y mejorar su situación económica. Para este proceso se debe de establecer costos de producción como punto principal, además de establecer márgenes de rentabilidad de acuerdo con la oferta y la demanda del mercado.

Es muy importante sensibilizar a los agricultores para que puedan conocer las ventajas y desventajas de la comercialización asociativa. Esto permitirá que los agricultores asociados puedan tener mayores oportunidades para comercializar con mercados mayoristas de forma asociativa y así contar con el volumen necesario de cebolla y reducir los costos de cultivo. Se requiere capacitación constante a los productores para clasificar y empaquetar la cebolla.

La capacitación y asesoramiento de los agricultores asociados para la consolidación de herramientas para la gestión empresarial es otra de las recomendaciones de Medina (2012) para fortalecer la cadena productiva de la cebolla roja en Ilabaya. No es el único trabajo que aborda esta problemática. Quizás el más reciente es el de William Óscar López Miranda, Efraín Mamani Coila, María del Pilar Oda Ortiz, Percy Enrique Rubina Cárdenas, donde nos explican que, durante los últimos 20 años, la cantidad de hectáreas de cultivo de cebolla se ha triplicado.

Esto se debe a que la cantidad de terreno dedicado al cultivo de esta hortaliza aumentado y la producción de igual forma. Sin embargo, el periodo de siembra no se realiza de forma planificada.

Estos dependen directamente de los resultados obtenidos del año anterior, aunque estos puedan ser buenos o malos. López et al. (2018) establece la visión del subsector dedicado al cultivo de la cebolla, donde afirma que ambas están orientadas a alcanzar la competitividad de los participantes de la cadena productiva. Para su cumplimiento se necesita la aplicación de buenas prácticas agrícolas y de manufactura.

Estas dos actividades buscan mejorar la calidad de la cebolla y que todas sus actividades sean sostenibles. En el capítulo III de esta investigación se presentaron tres matrices: MEFE, MPC y MPR. La MEFE indicó que existen oportunidades de mejora en el sector cebollero. Esto podría significar grandes ventajas comerciales para los participantes de la cadena productiva de la cebolla. El MPC demostró una mejor posición de la cebolla frente a otros productos, tales como el ajo o el porro. Por último, la MPR señaló que actualmente en el comercio de cebolla, España muestra un mayor nivel de competencia que Perú, Estados Unidos y México.

En el Perú, la producción de cebolla se encuentra orientada a cubrir el mercado interno (López et al., 2018). Si bien la variedad más popular es la roja, la cebolla amarilla es un producto con muchas posibilidades para ser exportado a Estados Unidos.

Un aspecto positivo de esta situación es que en las temporadas donde en Estados Unidos no se produce este vegetal, la costa peruana tendría la capacidad para poder albergar regiones con condiciones apropiadas para el cultivo de esta variedad de cebolla. El subsector de la cebolla tiene un mercado con una baja participación, pero con una industria de rápido crecimiento.

Se han propuesto varias estrategias para la exportación de cebollas en sus variedades amarilla y roja. Para que esto pueda funcionar se tienen algunos requisitos:

a) Invertir en tecnología, procesos y comercialización del producto.
b) Implementación de certificaciones de buenas prácticas agrícolas
c) Capacitar en todos los niveles de la cadena productiva de la cebolla
d) Lograr una integración vertical y horizontal de los participantes de la cadena productiva.

Asimismo, la estabilidad política y económica del país han permitido un buen ambiente de negocios. Sin embargo, aún se requieren reformas tecnológicas y de infraestructura (López et al., 2018).

Arequipa, es un departamento que reúne las condiciones para ser un "cluster". Esto gracias a la alta producción de cebolla a comparación de otras zonas del Perú. Si esto se da, la competitividad del subsector cebollero aumentaría considerablemente. López et al. (2018) concluye que la producción no planificada de la cebolla genera distorsiones entre oferta y demanda de este producto. Esto afecta directamente la rentabilidad del productor y disminuye el bienestar de la cadena productiva.

Se mencionan una serie de debilidades que afronta el subsector cebollero: Poca coordinación, falta de comunicación, escasa o nula planificación (de productores, gremios e instituciones públicas) y un escaso conocimiento por parte de los productores acerca de las necesidades del mercado nacional y extranjero. López et al. (2018) recomienda el establecimiento de alianzas entre asociaciones de diversos subsectores agrícolas.

Esto con el fin de planificar las campañas de siembra de acuerdo con estudios de mercado para poder satisfacer la demanda de estos subsectores.

Es necesario, además, promover el uso de sistemas de información agrícolas y realizar investigaciones de mercados previas al proceso de siembra con el fin de evitar distorsiones en el mercado que afecta principalmente el precio y termina por afectar la estabilidad económica de los productores.

BASES TEÓRICAS DE LA INVESTIGACIÓN

Aprendizaje Automático

El aprendizaje automático o Machine Learning (ML) es un proceso por el que se usan modelos matemáticos para lograr hacer que un equipo pueda aprender sin la necesidad de brindarle instrucciones directas (Mhaskar y Poggio, 2016). El aprendizaje automático hace uso de algunos algoritmos para la identificación de patrones para que posteriormente, estos sean utilizados para la creación de modelos de datos que sean capaces de realizar predicciones.

La experiencia y los datos hacen que los resultados del aprendizaje automático son más precisos. La adaptabilidad del aprendizaje automático lo convierte en la mejor opción en situaciones en donde nos encontramos con datos que cambian constantemente, la constante transformación de alguna solicitud o tarea o de cualquier otro tipo de solución cuya codificación podría parecer imposible.

Inteligencia Artificial

Dentro de las múltiples definiciones que podría tener la IA (Inteligencia Artificial) se puede resumir como la capacidad para que los computadores puedan realizar actividades que normalmente necesitan del razonamiento humana. Esta capacidad es posible gracias al uso de algoritmos que le permiten aprender de los datos tal y como lo haría un ser humano.

Pero a diferencia del ser humano, las computadoras no necesitan descansar y son capaces de analizar gran cantidad de datos a la vez (Lasse, 2018). Por si no fuera mucho, los errores que pueden cometer estas máquinas al momento del procesamiento son significativamente menor a los errores cometidos por los seres humanos en las mismas actividades.

La idea de que la IA es que las computadoras y/o programas puedan aprender y tomar decisiones representa un importante avance y demuestran el crecimiento exponencial que tienen sus procesos. Gracias a estas características, las aplicaciones de inteligencia artificial pueden ejecutar una gran cantidad de actividades que antes estaban reservadas únicamente para los humanos.

Redes Neuronales Artificiales

Las redes neuronales artificiales, es una de las múltiples técnicas de la Inteligencia Artificial ya que pretende desarrollar soluciones tecnológicas que puedan imitar el comportamiento del cerebro humano. Específicamente las redes neuronales artificiales buscan imitar la capacidad de pensar del cerebro humano y esta pueda ser utilizada en diferentes campos, como para el reconocimiento de imágenes a través de patrones.

Esta capacidad de las redes neuronales artificiales ha inspirado a muchos investigadores a intentar modelar el funcionamiento del cerebro humano en el computador. Actualmente las redes neuronales artificiales son utilizadas en diferentes campos

Se podría resumir que las redes neuronales artificiales buscan imitar el comportamiento de las neuronas humanas que procesan información y puedan ser capaces de resolver funciones no lineales que pertenecen a sistemas con un alto nivel de complejidad y que generalmente son impredecibles o imposibles de modelar (Rivas y Mazón, 2018).

Aprendizaje supervisado

El aprendizaje supervisado es uno de los grupos de algoritmos de Inteligencia Artificial que trabaja con datos etiquetados. Los datos etiquetados se refieren a los datos para los que ya se conoce una solución de parte del destino.

Este tipo de algoritmos trabaja con datos históricos por el que el algoritmo asigna una etiqueta como salida o función que le permitirá predecir la variable objetivo para realizar otra actividad.

Este tipo de aprendizaje se utiliza en dos tipos de problemas:

- **Problemas de clasificación**: Se caracterizan principalmente por tener una variable cualitativa como respuesta o salida para este tipo de problemas. Las técnicas de clasificación se encargan de predecir la probabilidad de que una observación pueda pertenecer a una categoría o no.
- **Problemas de regresión**: A diferencia de los problemas de clasificación, estos se caracterizan por tener una variable cuantitativa, de esta forma, la solución de un problema de este tipo

Estos dos tipos de problemas de aprendizaje supervisado se diferencias principalmente por el tipo de variable que se desea predecir, mejor conocida como variable objetivo según (Martinez, 2020). Para los problemas de clasificación, la variable objetivo es de tipo categórico y en los casos de regresión, la variable objetivo es numérica.

Aprendizaje no supervisado

A diferencia del aprendizaje supervisado, este tipo de aprendizaje es utilizado para reunir datos no estructurados de acuerdo con características similares propias del conjunto de datos. Para Martinez (2020), se denomina "aprendizaje no supervisado" porque el algoritmo que es utilizado no es guiado, como ocurre con el aprendizaje supervisado. El objetivo de este tipo de aprendizaje es el de encontrar patrones que no eran conocidos en los datos.

Es por este motivo que las predicciones elaboradas con este tipo de algoritmos son muy deficientes, pues no se conoce con exactitud cual deben ser los resultados de estas técnicas.

Los algoritmos de este tipo de aprendizaje manipulan los datos sin conocerlos previamente con el único objetivo de explorarlos para definirlos en el proceso. Las técnicas de aprendizaje no

supervisado durante todo el proceso se encargan de investigar la estructura de la información para la detección de patrones no comunes.

Predicción

Desde el punto de vista de la computación y la informática, la predicción es producto de un sistema de IA que nos informa de la probabilidad de que ocurra algo.

Actualmente estos sistemas de predicción están en varias partes: sugerencias de videos, películas, música de acuerdo con preferencias o experiencias pasadas, la mejora de la precisión meteorológica gracias a los patrones de climas anteriores.

Esto nos quiere decir que, si a estos sistemas se les brinda la información necesaria, estos tendrán un mejor entrenamiento que pueda ser utilizado para tomar decisiones con mejor certeza (Oxford Internet Institute, 2020). Se puede decir que las predicciones basadas en Inteligencia Artificial ayudan a las personas a tomar decisiones en el día a día, aunque no exista una obvia conexión con el futuro.

ESTUDIO ECONÓMICO

Costos fijos y variables

Para la investigación se contempló la inversión en costos fijos de aquellos gastos en servicios (agua, luz, servicio de telefonía), adquisición de equipos, licencias entre otros para el inicio de las actividades del proyecto. Con lo que respecta a los costos variables se consideró los gastos que se requieren hacer para las visitas técnicas y actividades en campo para la extracción de información y para el proceso de análisis y recolección de datos estadísticos, incluyendo en este grupo de costos útiles de escritorio y papelería. La inversión total (costos fijos y variables) del proyecto queda detallado en la tabla 1 con los conceptos detallados y los costos de cada uno

Tabla 1
Estructura de costos variables y fijos del proyecto

Costos Fijos	**S/136.33**
Teléfono	S/24.00
Agua	S/3.00
Luz	S/26.00
Licencia de software (1)	S/41.67
Depreciación Laptop (20%)	S/33.33
Depreciación Equipo Móvil (25 %)	S/8.33
Costos Variables	**S/1,095.00**
Mano de obra especializada	S/800.00
Útiles de escritorio	S/80.00
Papelería	S/30.00
Licencias momentáneas	S/35.00
Movilidad	S/80.00
Viáticos	S/70.00
Total de costos	**S/1,236.33**

Nota: Esta tabla muestran el costo total (variables y fijos) de servicios, equipos, licencias, entre otros, que han sido considerados para el desarrollo del proyecto

Flujo de caja

El flujo de caja se entiende como las entradas y salidas de caja o efectivo que permite reportar los ingresos operativos que fueron proyectados, así como los gastos requeridos para el cumplimiento de las actividades dentro de un proyecto. Ver tabla 2.

- Se considera el **Ingreso por ventas** a los ingresos generados por la creación de la red neuronal y su estudio para el establecimiento de reglas para la mejora de la rentabilidad de la cebolla.
- Los gastos adicionales por la puesta en marcha afectarán directamente el Ingreso y su resultado serán las **Ventas Fijas** que refleja la ganancia verdadera del proyecto.
- La **Utilidad Bruta** se definió como la diferencia de los costos (variables y fijos) y las **Ventas Netas**.

Tabla 2
Flujo de caja del Proyecto

Ingreso por Ventas	**S/3800.00**
• Costo de ventas	**S/ 250.00**
Ventas Netas	**S/ 3550.00**
Costos Fijos	S/ 141.33
• Teléfono	S/ 24.00
• Agua	S/ 3.00
• Luz	S/ 26.00
• Licencias de software	S/ 41.67
• Depreciación PC (20%)	S/ 33.33
• Depreciación eq. Móvil (25 %)	S/ 13.33
Costos Variables	**S/ 1495.00**
• Mano de obra especializada	S/ 1200.00
• Útiles de escritorio	S/ 80.00
• Papelería	S/ 30.00
• Licencias momentáneas	S/ 35.00
• Movilidad	S/ 80.00
• Viáticos	S/ 70.00
Total Costos	**S/ 1236.33**
Utilidad Bruta	**S/ 2313.67**
• Impuestos (15 %)	S/ 347.05
Utilidad Operativa	**S/ 1966.62**
• Reparto de utilidades	S/ 1000.00
Utilidad Neta	**S/ 966.62**

Nota: Esta tabla muestra las entradas y salidas de efectivo para el proyecto que reflejan los ingresos y costos de este.

Análisis de Rentabilidad

La finalidad principal del uso de esta herramienta de inteligencia artificial es el de brindar información para definir un mejor plan de cosecha que pretenda reducir los costos operativos del cultivo de cebolla y aumentar la rentabilidad para los agricultores de cebolla en un 4%, dada por la Utilidad Operativa de la diferencia de los Costos Operativos de la cosecha y la siembra entre los ingresos por la Venta Total de la producción de la extensión de terreno referencial (2.3 hectáreas).

En la tabla 3 se describe detalladamente los conceptos de los costos operativos de la siembra y cosecha y el ingreso por venta total una producción de 81500 kilos, que sería la producción correspondiente a la extensión de terreno referencial de 2.3 hectáreas.

Tabla 3

Análisis de rentabilidad del proyecto – Campaña 2021 y 2022

	Campaña 2020	**Campaña 2021**
Ingreso por venta total	S/48,900.00	S/40,750.00
Costos operativos siembra	S/16,200.00	S/15,600.00
Mano de obra	S/9,300.00	S/9,800.00
Fertilizantes	S/6,000.00	S/4,500.00
Otros	S/900.00	S/1,300.00
Costos operativos cosecha	S/13,364.00	S/8,063.00
Mano de obra	S/10,320.00	S/6,230.00
Otros	S/1,642.00	S/433.00
Total costos	S/1,402.00	S/1,400.00
Utilidad operativa	S/29,564.00	S/23,663.00
Rentabilidad (%)	**39.54%**	**41.93%**

Nota: Esta tabla muestra la rentabilidad de la siembra y cosecha de cebolla roja en Arequipa para el periodo 2021-2022

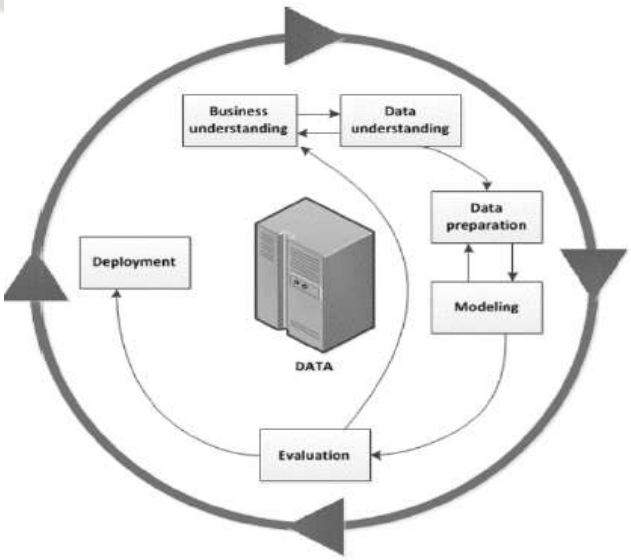
Para la campaña del 2020, la rentabilidad se halló teniendo en cuenta que no se hace uso de ninguna herramienta de inteligencia artificial. Para la campaña del 2021 se considera el uso de la información obtenida por la red neuronal para la aplicación a medidas correctivas en el plan de cosecha y siembra.

Esta comparativa sirve para ver el impacto que tendría el uso de un modelo de red neuronal en el cultivo de cebolla roja, que como se muestra en la tabla 3, mejora un 2.39% la rentabilidad de la siembra y cosecha, buscando aumentar esta cifra con una extensión de terreno mayor y lógicamente, la participación de una mayor cantidad de agricultores.

CAPÍTULO II
DISEÑO E IMPLEMENTACIÓN DE LA METODOLOGÍA

Para la implementación de la Red Neuronal y la obtención de resultados se utilizarán los pasos de modelado de procesos CRISP-DM (Cross-Industry Standard Process for Data Mining) donde se ejecutarán las 6 fases correspondientes de este proceso cíclico: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelamiento, Evaluación y Despliegue (Chapman et al., 2000). La metodología CRISP-DM tiene como base el proceso de análisis de los datos de forma análoga, muy parecido al ciclo de vida de software. Esta metodología consiste en 6 fases que se muestran en la Figura 1 y tiene una naturaleza bidireccional cíclica: En cada fase de esta metodología se busca obtener un resultado de calidad para la siguiente etapa o tarea que se deba ejecutar. La metodología CRISP-DM tiene como base el proceso de análisis de los datos de forma análoga, muy parecido al ciclo de vida de desarrollo de software. En esta metodología consiste en 6 fases que se muestran en la Figura 1. El seguimiento a estas fases no es rígido y siempre se permite movilizarse entre las diferentes fases del proyecto.

Figura 1
Fases de la metodología de modelado de procesos CRISP-DM



Nota: La figura representa el ciclo de vida del ciclo de vida de minería de datos. Tomado de Guía de CRISP-DM de IBM SPSS Modeler de IBM Documentation, 2021.

FASES DE LA METODOLOGÍA

Comprensión del negocio

En esta fase inicial está orientada a definir y entender los objetivos que se tiene en el proyecto. Esta información ayudará a la definición del problema, los requerimientos y en la creación de un plan preliminar para cumplir con los objetivos.

Comprensión de los datos

Esta fase se inicia con la recolección de datos e incluye la familiarización de estos. Para ello, en esta fase se incluyen las siguientes actividades:

- Identificación de los problemas de calidad.
- Descubrimiento de conocimiento primario de estos datos.
- Descubrimiento de subconjuntos de datos para la formulación de hipótesis.

Preparación de los datos

Implica todas las actividades requeridas para la construcción del conjunto final de datos (los datos que serán utilizados en las herramientas de modelado). En esta etapa se incluyen la selección de tablas, registros, atributos, así como la transformación y limpieza de los datos para las herramientas.

Modelamiento

En esta fase se seleccionan las técnicas de modelamiento más adecuadas para abordar el problema y se configuran los parámetros a utilizar a sus óptimos valores. Normalmente en proyectos que utilizan esta metodología es común volver a esta etapa para la reconfiguración de los parámetros para la obtención de un mejor modelo.

Evaluación

En esta fase del proyecto se tienen uno o varios modelos que cumplen con la calidad necesaria desde la perspectiva del análisis de los datos. Se recomienda evaluar los resultados obtenidos con los objetivos definidos. Al finalizar esta fase se debe tener una idea clara de la decisión que se tomará para la aplicación de los resultados del análisis de los datos.

Despliegue

La entrega del conocimiento obtenido del modelo, su gestión y la entrega al cliente para su utilización es la idea principal en esta última fase del proyecto. No se trata de una etapa rígida donde se entrega un informe con los resultados obtenidos y la evaluación, pues podría tratarse de un análisis continuo y la realización periódica para la gestión de los datos dentro de una organización.

Cada una de estas fases contiene una serie de actividades que son descritas en la tabla 4. Cada uno de estos pasos son de vital importancia para el avance entre cada fase y se recomienda cumplir con todas las actividades descritas en cada fase para el cumplimiento de los objetivos planteados en la investigación.

Tabla 4
Fases y Actividades de Metodología CRISP – DM para el proyecto

FASE	ACTIVIDAD
COMPRENSIÓN DEL NEGOCIO	1. Determinar objetivos del negocio 2. Evaluación de la situación actual 3. Determinar objetivos de minería de datos 4. Generar un plan preliminar del proyecto
COMPRENSIÓN DE LOS DATOS	1. Recopilación de los datos 2. Definición de los datos 3. Verificación de los datos
PREPARACIÓN DE LOS DATOS	1. Selección de los datos 2. Limpieza de los datos 3. Construcción de los datos 4. Integración de los datos 5. Formato de los datos
MODELADO	1. Selección de la técnica de modelado 2. Construcción del modelo
EVALUACIÓN	1. Evaluación del modelo 2. Evaluación de resultados 3. Revisión del proceso
DESPLIEGUE	1. Desplegar el plan 2. Monitorear y mantener 3. Generar informe final 4. Revisión del proyecto

Nota: Esta tabla muestra las fases y las actividades a completar por cada etapa del modelo de proceso CRISP-DM

APLICACIÓN DE LA METODOLOGÍA PROPUESTA

Comprensión del negocio

En esta primera fase de la metodología se irá revisando cada una de las actividades cuyo objetivo principal es el de definir los objetivos y requerimientos del proyecto con una perspectiva de negocio, para en las siguientes fases, convertir dichos objetivos en metas a cumplir desde el punto de vista técnico.

Determinación de objetivos del negocio

Como se ha mencionado anteriormente el propósito de esta investigación es el de pronosticar mediante el uso de un modelo de red neuronal la información necesaria para la mejora de la rentabilidad del cultivo de cebolla roja en la ciudad de Arequipa.

El objetivo es el de proporcionar material informativo para que los gremios, asociaciones y trabajadores independientes puedan trabajar juntamente con un plan de siembra y cosecha bien elaborado y que les permita a los agricultores dedicados al cultivo de cebolla poder continuar con sus actividades sin afectar su economía. De esta forma:

- Generar un modelo de red neuronal para estudiar la relación entre las variables de cosecha de cebolla roja y su precio en el mercado y verificar su exactitud usando datos reales de la zona agrícola El Cural, en Arequipa durante el periodo de cosecha 2021- 2022.
- Seleccionar de las variables de entrada que determinarán el mejor modelo de predicción (climatológicas, geográficas, etc. entre otros).
- Analizar los resultados obtenidos para estudiar su interpretabilidad para su evaluación y elaboración de reglas.

Evaluación de la situación actual

No existe alguna organización privada o algún gremio de agricultores que tenga alguna base de datos con información detallada de campañas de cultivo de cebolla o información de las variables de cosecha de ese producto en temporadas pasadas.

Se sabe de la existencia de sistemas de información del MIDAGRI (Ministerio de Agricultura y Riego) y de canales de atención al ciudadano para la consulta y acceso a información pública de diferentes productos agrícolas.

Se utilizó la información del Sistema de Información de Abastecimiento y precios (SISAP) de los mercados mayoristas en las ciudades de Tumbes, Piura, Chiclayo, Trujillo, Huaraz, Ica, Arequipa, Moquegua, Tacna, Cajamarca, Chachapoyas, Huánuco, Cerro de Pasco, Huancayo, Huancavelica, Huamanga, Abancay, Cusco, Puno, Iquitos, Tarapoto, Pucallpa, Puerto Maldonado, Andahuaylas, Jaén y Chota, los mercados mayoristas de Lima Metropolitana y los 26 centros de comercialización mayorista de productos, para la obtención de los diferentes precios y las equivalencias en kilos de los productos de cebolla roja comercializados en las regiones mencionadas.

Dicha información es procesada todos los lunes, miércoles y viernes con una vigencia de 2 días. Asimismo, se incluyó la información climatológica del SENAMHI donde se escogió las variables de precipitación acumulada, temperatura máxima, temperatura mínima, año, mes y día recolectada en tiempo real por la estación meteorológica convencional de Huasacache.

Determinar objetivos de minería de datos

Los objetivos definidos en términos de minería de datos son:

a) Predecir el precio promedio, mínimo y máximo de la cebolla roja (por kilo) en función del año y mes escogidos.

b) Predecir la cantidad de hectáreas de cebolla roja sembradas, cosechadas y la producción total.

Criterios de éxito de la investigación

Se estableció como criterio de éxito de la investigación la posibilidad de predecir los precios de venta de la cebolla roja en los mercados nacionales en épocas tempranas de acuerdo con los datos de entrada de la campaña en curso (cantidad de hectáreas sembradas, producción de campañas anteriores, hectáreas cosechadas en campañas anteriores).

Otro criterio de éxito a considerar al finalizar la investigación serán los resultados de la utilización de dicha información y el cumplimiento de los objetivos secundarios de esta investigación.

Plan preliminar del proyecto

El proyecto será dividido en etapas para hacer más sencilla la organización y tener una estimación real de realización de este se muestran en la Tabla 5.

Tabla 5
Plan Preliminar del Proyecto

FASE	ACTIVIDAD	TIEMPO ESTIMADO
1	Análisis de los datos y ejecución de un primer filtrado de datos.	1 semana
2	Estudio inicial y elaboración de primeras impresiones	1 semana
3	Preparación de los datos (ETL). Carga de información en un gestor de base de datos.	2 semanas
4	Ejecución de consultas para mostrar información representativa.	1 semana
5	Selección de técnicas de modelado. Modelamiento y ejecución de técnicas a los datos. Comparación de técnicas	2 semanas
6	Análisis de resultados	1 semana
7	Elaboración de informes de resultados de acuerdo con los objetivos establecidos.	1 semana
8	Presentación de resultados finales	4 días

Nota: Esta tabla resume las actividades descritas en la tabla 4 para cada una de las fases del modelo de proceso CRISP-DM y el tiempo estimado para cada una.

Comprensión de los datos

En esta segunda etapa de la metodología CRISP-DM se realiza la recolección preliminar de los datos para tener una perspectiva más cercana del problema, familiarizarse con los datos y determinar la calidad de estos. En esta parte se desea tener identificadas las relaciones entre los datos para la formulación de las primeras hipótesis.

Recopilación de los datos iniciales

Los datos que fueron utilizados en este proyecto son referentes a las variables de cosecha de cebolla roja durante los últimos 11 años que incluyen precio (mínimo, promedio, máximo), tipo de producto (equivalencia en kilos), año, mes y tipo de comerciante (minorista o mayorista) entre otros.

Para esto se utilizó datos reales del Sistema de Información de Abastecimiento y precios (SISAP) que agrupa un resumen de los precios mínimos, promedios y máximos de cebolla roja en los mercados mayoristas de diversas regiones del país y datos climatológicos de los últimos años del SENAMHI.

No se requirió de la generación de datos aleatorios, pero si se realizó el preprocesamiento de estos para mover los datos contenidos en una hoja de cálculo con extensión. para formatearlos, limpiarlos y cargarlos a una base de datos para su análisis.

A continuación, se lista los datos adquiridos desde el SISAP y del SENAMHI:

a) **Año y mes**: Mes y año específico en el que se registraron los precios mínimos, promedios y máximos de la cebolla en mercados mayoristas del país.
b) **Producto**: Cada nombre de producto se refiere a la forma de presentación (saco, malla, saco mediano, entre otros) de la cebolla y depende directamente de la equivalencia o peso del producto.
c) **Comerciante**: Los comerciantes dedicados a la venta de cebolla pueden ser mayoristas o minoristas.
d) **Precio promedio, precio mínimo y precio máximo:** Límites de los precios que alcanzó la cebolla roja en sus diferentes presentaciones o equivalencias (en kilos) en los diferentes mercados mayoristas y minoristas del país.
e) **Terreno sembrado:** Extensión de terreno representado en hectáreas en el que se sembró cebolla roja en la ciudad de Arequipa
f) **Terreno cosechado:** Extensión de terreno de cebolla roja que se cosechó al final de la temporada.
g) **Producción:** Cantidad de toneladas de cebolla roja recolectada al finalizar la campaña de cosecha.

h) **Precipitación acumulada:** Cantidad de lluvia o cualquier forma de material acuoso que cae de las nubes y equivale a la medición del agua que se acumularía en una superficie natural.

i) **Temperatura máxima:** El más alto nivel de temperatura en grados Celsius (C°) registrada en una fecha específica en una estación meteorológica.

j) **Temperatura mínima:** Nivel más bajo de temperatura en grados Celsius (C°) registrada en una fecha específica en una estación meteorológica.

Estos datos serán recogidos durante el proceso de extraídos, transformados y cargados a tablas durante el proceso ETL:

1. Producto
2. Precio
3. Tiempo
4. Clima
5. Campaña

Definición de los datos

Los datos se encuentran almacenados en una hoja de cálculo de Excel y corresponden a datos de cosecha y datos climatológicos desde el año 2011 hasta la actualidad (2021).

Exploración de los datos

Una vez que los datos hayan sido descritos se procede a la etapa de exploración. Esta actividad implica la aplicación de pruebas estadísticas que buscan revelar propiedades ocultas de los datos, la creación de tablas de contingencia y gráficos de distribución. Este análisis preliminar para la comprobación de la consistencia y completitud de los datos.

La fecha, el tipo de producto y los precios serían las variables más relevantes en la creación del modelo de pronóstico. En la tabla 6 se presentan los datos del año de cosecha, el precio promedio, precio mínimo y el tipo de producto (kilo, malla o saco) del mes de marzo de los últimos 3 años.

Tabla 6

Relación de variables de cosecha: año, precio promedio y precio mínimo y tipo de producto.

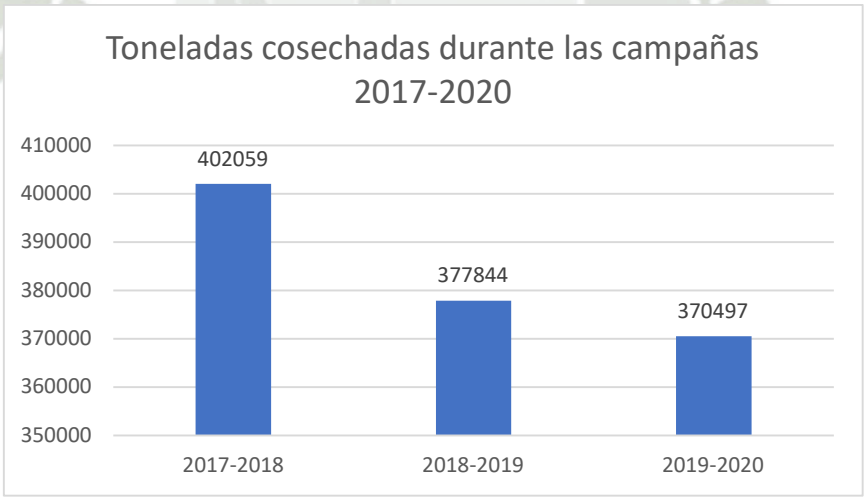
	2018		2019		2020	
Tipo de producto	**P. Prom**	**P. Min**	**P. Prom**	**P. Min**	**P. Prom**	**P. Min**
Kilo	1.04	0.35	1.77	1	1.55	0.4
Malla	86.45	85.2	213.8	211	70.17	65
Saco	55.57	52	74.31	70	50.25	48

Nota: Esta tabla muestra el precio promedio y precio mínimo del kilo, malla y saco de cebolla roja en el mes de marzo en el año 2018,2019 y 2020.

En la tabla 6 se puede apreciar la evolución del precio promedio y precio mínimo en el año 2018, 2019 y 2020, siendo el 2019 el año donde se obtuvo un mejor precio de venta en los mercados mayoristas a nivel nacional, llegando a 1 S/. el kilo de cebolla como precio mínimo. Esto muestra una diferencia de 0.65 centavos con respecto al año 2018 y 0.60 centavos con respecto al 2020.

Figura 2

Relación entre toneladas cosechadas y el año de la campaña (2017-2020)



Nota: Esta figura muestra la relación entre las toneladas cosechadas y la temporada (agosto-julio) desde el 2017 al 2020.

En la Figura 2 se muestra las toneladas cosechadas durante los 3 últimos años y vemos una disminución considerable desde la campaña del 2017.

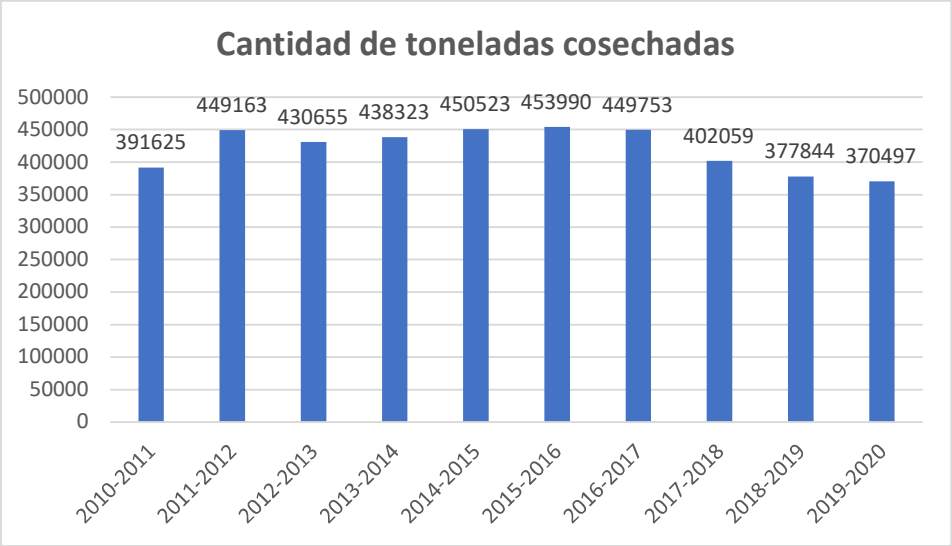
Tal comportamiento puede deberse a una disminución muy brusca de la cantidad de hectáreas sembradas durante estos los 2 últimos años (2019 y 2020).

Sin embargo, a diferencia del 2018, la campaña del 2019 logró obtener un mejor precio de venta, teniendo una diferencia de 0.75 centavos el kilo de cebolla, es decir, en el 2018 el precio mínimo del kilo de cebolla roja llegó hasta los 0.35 S/. mientras que en el 2019 el precio mínimo fue de 1 S/. Sin embargo, en el 2020, se tuvo una producción aún menor que en la campaña del 2019 pero el precio lejos de mantenerse o incrementarse, disminuyó hasta alcanzar los 0.40 S/.

En la Figura 2 se muestra la evolución de la producción total de cebolla roja de los últimos 10 años en la ciudad y como este ha ido disminuyendo con el paso del tiempo.

Figura 3

Cantidad de toneladas cosechadas desde el 2010 hasta el 2020



Nota: Esta figura muestra la relación entre las toneladas cosechadas por año desde el 2010 al 2020.

En la Figura 3 se puede apreciar la producción total de las campañas desde el 2010 hasta el 2020. Desde el 2011 hasta el 2017 no hay una disminución considerable de las toneladas producidas en la extensión de terreno que fueron sembradas durante esos periodos.

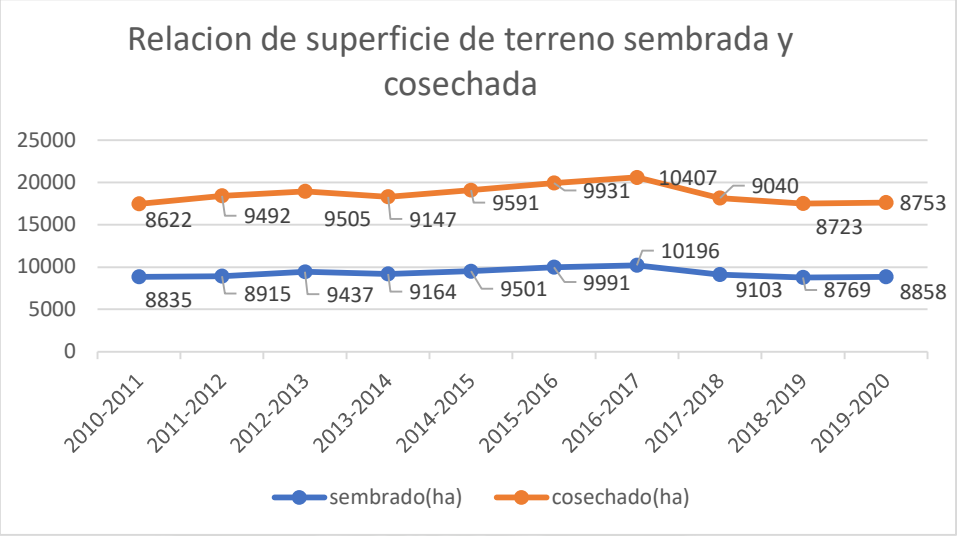
Sin embargo, los 3 últimos años se ve una disminución considerable de esta producción llegando a alcanzar las 370497 toneladas en la campaña del 2020, la cifra más baja de los últimos 10 años. Dicho comportamiento debería tener un impacto positivo en los precios de la cebolla al no tener una producción tan alta como otros años.

Pero en este último caso, para la campaña del 2020 existen factores externos que podrían haber afectado tanto a la producción de la cebolla como su precio: la pandemia y toda la cadena de suministros de alimentos que tuvieron un paro en sus actividades, el alza de precios en este periodo de los fabricantes y proveedores, la disminución de la superficie sembrada de cebolla roja, entre otros. Las plagas y enfermedades son otro factor que interviene en la producción de cebolla cada año.

En la figura 4, se ve la relación entre la cantidad de hectáreas sembradas y la cantidad cosechada en la región Arequipa para establecer la relación entre algún factor que pueda afectar el rendimiento de la extensión de terreno cosechada.

Figura 4

Relación entre la superficie de cebolla sembrada y la cantidad de toneladas cosechadas desde el 2010 hasta el 2020



Nota: Esta figura muestra la relación entre la superficie en hectáreas que fueron sembradas desde el 2010 hasta el 2020 y las hectáreas de terreno cosechadas.

Tabla 7

Relación de hectáreas sembradas, cosechadas, toneladas producidas y rendimiento por hectárea (2010 – 2015)

Periodo	Sembrado (ha)	Cosechado (ha)	Producido (T)	Rendimiento (t/ha)
2010-2011	8835	8622	391625	45.422
2011-2012	8915	9492	449163	47.320
2012-2013	9437	9505	430655	45.308
2013-2014	9164	9147	438323	47.920
2014-2015	9501	9591	450523	46.974

Nota: Esta figura muestra la relación entre la superficie en hectáreas sembradas y cosechadas, así como las toneladas producidas y el rendimiento por hectárea desde el 2010 hasta el 2015

Tabla 8

Relación de hectáreas sembradas, cosechadas, toneladas producidas y rendimiento por hectárea (2015-2020)

	Sembrado (ha)	Cosechado (ha)	Producido (T)	Rendimiento (t/ha)
2015-2016	9991	9931	453990	45.174
2016-2017	10196	10407	449753	43.216
2017-2018	9103	9040	402059	44.476
2018-2019	8769	8723	377844	43.316
2019-2020	8858	8753	370497	42.328

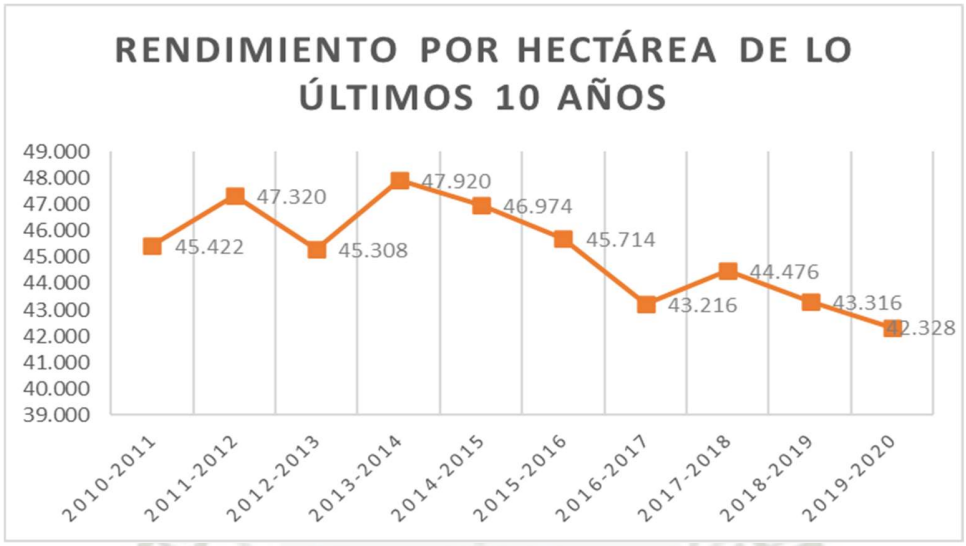
Nota: Esta figura muestra la relación entre la superficie en hectáreas sembradas y cosechadas, así como las toneladas producidas y el rendimiento por hectárea desde el 2015 hasta el 2020

En la tabla 7 se puede apreciar que en los últimos años no hay una diferencia muy grande entre la relación de la superficie sembrada y la cosechada. Esta oscila entre un 0.18% menos de la producción total sembrada y 2.41% en la campaña del 2010-2020, donde a pesar de haber un porcentaje más elevado, el rendimiento por esa superficie sembrada fue relativamente alta a comparación de años posteriores.

Si retrocedemos unos 10 años atrás, podemos ver que el rendimiento por hectárea es superior y que de igual forma la cantidad de hectáreas sembradas es mayor a la superficie de terreno sembrada en las campañas desde el 2015 hasta el 2020 y a pesar de que hay temporadas donde estos datos de siembra son similares, existe un mayor rendimiento por hectárea en la tabla 7 que en la tabla 8.

Figura 5

Relación de la evolución del rendimiento por hectárea de cebolla roja de los últimos 10 años.



Nota: Esta figura muestra la evolución del rendimiento por hectárea de cebolla roja en los últimos 10 años.

En la figura 5 se puede apreciar una disminución considerable del rendimiento en los últimos años de cosecha de cebolla roja por hectárea, siendo la campaña 2013-2014 la mejor en términos de producción con un rendimiento de 47.920 toneladas por hectárea y una producción total de 438323 toneladas de cebolla. Sin embargo, fue la campaña del 2015-2016 donde se logró una mayor producción de cebolla de los últimos 10 años, a pesar de tener un rendimiento menor (46.974 t/ha) a la campaña anterior. En la tabla 9 se muestran otras variables de cosecha de estas dos temporadas de cosecha para ver su relación con el precio de venta.

Tabla 9

Relación de año, mes y precios mínimo, promedio y máximo de la campaña 2014 y 2015

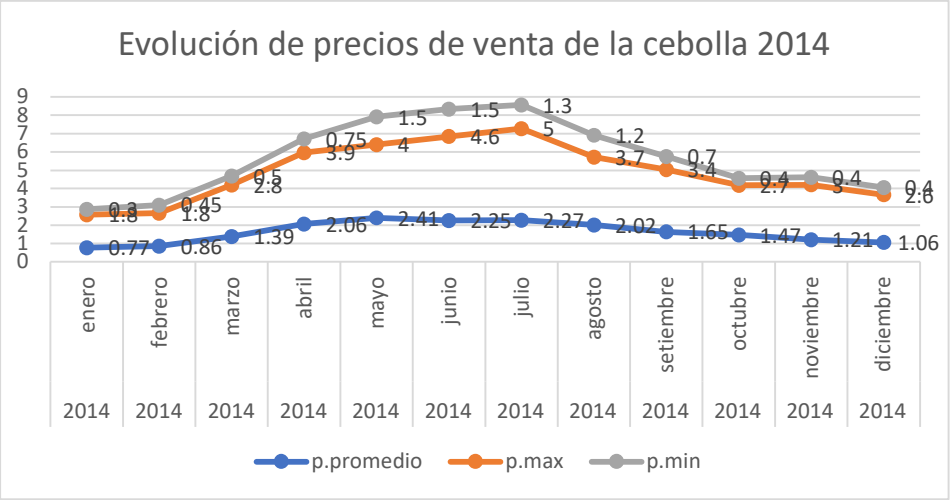
	2014			2015		
	Precio prom.	**Precio min.**	**Precio max.**	**Precio prom.**	**Precio min.**	**Precio max.**
Enero	0.77	0.3	1.8	1.1	0.5	2.6
Febrero	0.86	0.45	1.8	1.17	0.5	2.6
Marzo	1.39	0.5	2.8	1.27	0.6	2.5
Abril	2.06	0.75	3.9	1.34	0.5	2.5

Nota: Esta tabla muestra los precios mínimo, promedio y máximo del 2014 y 2015 donde hubo una mejor producción de cebolla roja.

Como se muestra en la tabla 9, existe un mejor precio de venta por kilo de cebolla roja en la campaña del 2015. En la figura 5 se apreciaba un mejor rendimiento y por tal motivo, una mejor producción de cebolla en el 2014 por lo que es posible que el precio de venta pueda haberse visto afectado por una sobreproducción de cebolla, que en ese año en la región Arequipa fue de 453990 toneladas.

Figura 6

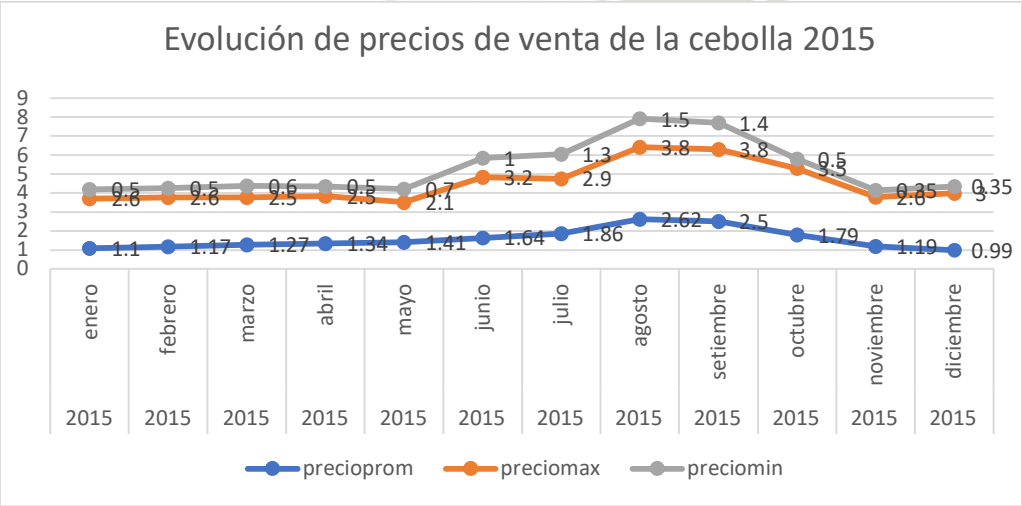
Evolución de los precios de venta de cebolla en el año 2014



Nota: Esta figura muestra la evolución del precio (mínimo, máximo y promedio) por mes de la campaña 2014.

Figura 7

Evolución de los precios de venta de cebolla en el año 2015



Nota: Esta figura muestra la evolución del precio (mínimo, máximo y promedio) por mes de la campaña 2014.

Durante los primeros meses de la campaña de cosecha del 2014, el precio promedio de venta de cebolla roja no sobrepasa la unidad, mientras que en el 2015 se ve un caso contrario. Sin embargo, la evolución del precio de la cebolla evoluciona rápidamente pasados los dos primeros meses del año. En la figura 6 y 7 se realiza una comparación de la evaluación de los precios a lo largo de todo el año 2014 y 2015.

Como se puede aprecias en la figura 6 y 7, la relación existente entre los precios de venta de cebolla a lo largo de los años 2014 y 2015 presenta algunas diferencias en los meses donde se visualiza un incremento en los precios. Se visualiza una curva ascendente en los 4 primeros meses del año 2014 que logra hacer llegar el precio mínimo hasta 1.5 S/. en el mes de julio. Por otro lado, el precio mínimo de la cebolla durante los primeros meses del 2015 no llega a alcanzar los 0.70 centavos, oscilando de esta forma entre los 0.5 S/ y los 0.6 S/ hasta el mes de junio.

Tabla 10

Relación de variables climatológicas, mes y año

Año	Sembrado (ha)	Cosechado (ha)	Producido (t)
2017	10196	10407	449753
2018	9103	9040	402059

Nota: Esta tabla muestra los datos de siembra y cosecha de cebolla roja de la temporada 2017 y 2018.

Tabla 11

Relación de hectáreas sembradas, hectáreas cosechadas, producción en toneladas y año

	Enero		Febrero		Marzo	
	2017	2018	2017	2018	2017	2018
T. Máxima	22.59	22.94	22.90	22.41	22.95	23.97
T. Mínima	12.44	11.51	11.28	11.70	11.66	11.90
Humedad (%)	72.02	60.29	64.63	69.55	74.78	62.79
Precipitación	3.34	0.47	-7.55	-40.11	1.38	0.23

Nota: Esta tabla muestra los datos climatológicos de los primeros meses de los años 2017 y 2018.

La diferencia de producción de la campaña del 2014 a comparación de la del 2015 es de 12,200 toneladas, pero con un precio de venta y un rendimiento mayor. Para este caso se presenta la tabla 10 para analizar la influencia de las variables climatológicas en la producción y rendimiento de la cebolla en los años 2017 y 2018, donde el comportamiento fue muy parecido. Adicionalmente en la tabla 11 se muestra la cantidad de hectáreas sembradas, cosechadas y la producción de estos dos periodos de cosecha.

En la tabla 10 se puede apreciar una mayor cantidad de precipitación acumulada en el 2017 a diferencia que en el 2018 y un mayor nivel de humedad en los meses de enero y marzo en el 2017 a comparación del 2018.

Esto parece ver influido en la producción de esa campaña pues, como se puede apreciar en la tabla 11, la cantidad de toneladas producidas en el 2017 ciertamente es superior a la cantidad producida en el 2018, pero el rendimiento del año 2018 es mayor con un valor de 44.8 t/ha, mientras que en el 2017 se alcanza un rendimiento de 43.21 t/ha.

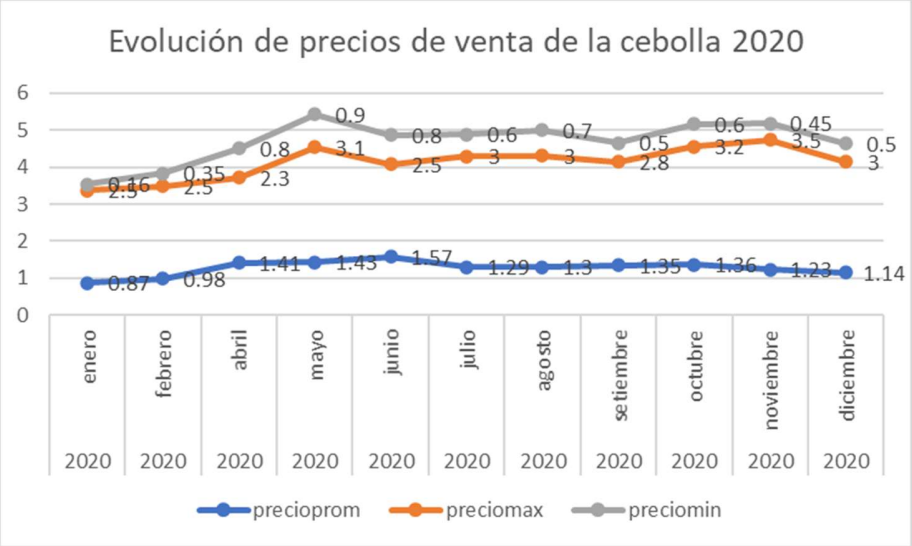
En la tabla 11 se muestra la relación entre las hectáreas sembradas, las hectáreas cosechadas y las toneladas producidas en el 2017 y el 2018, buscando de esa forma, establecer alguna relación entre alguna variable meteorológica y la producción total en una determinada campaña.

Se entiende por la información mostrada en la Tabla 10 demuestra una menor producción en el 2018 pero a pesar de ello el rendimiento por hectárea de este año es superior al del 2017. Se tiene un rendimiento de 44.50 toneladas por hectárea, mientras que en el 2018 se tiene un rendimiento de 43.21 toneladas por hectárea.

Al revisar nuevamente la tabla 9 nos fijamos que la precipitación acumulada del 2018 es más baja que la del año anterior, teniendo un posible déficit de producción por temas climatológicos.

Figura 8

Evolución de los precios de venta de cebolla roja por mes en el año 2020



Nota: Esta figura muestra la evolución del precio de venta de cebolla (mínimo, promedio y máximo) en el 2020.

Preparación de los datos

En la siguiente fase de la metodología se busca preparar los datos para que estos sean los adecuados para la aplicación de técnicas de minería de datos sobre los mismos. De esta forma, esta fase consiste en el filtrado y selección de un conjunto de datos a utilizarse, su posterior limpieza para la mejora de su calidad, la inserción de nuevos datos a partir de los ya que existen y que finalmente estos puedan obtener el formato adecuado y que es requerido por la herramienta que nos ayudará al modelado.

Selección de los datos

Respecto a los registros, se utilizará todo el contenido almacenado en cada tabla que compone la base de datos creada previamente, ya que al ser una base de datos que fue creada especialmente para esta investigación. Sin embargo, al momento del preprocesamiento de la información, se encontró que para la creación de la tabla Clima existían muy pocos registros y valores nulos en los existentes.

Se registra información de fechas que no aportan valor al modelo y la generación de registros para esta etapa no es una opción viable dada la cantidad de registros que tienen estas fallas y su impacto en el modelo podría afectar negativamente. Por este motivo, los campos climatológicos fueron descartados para la generación del modelo.

Asimismo, hay campos dentro de la base de datos que no son necesarios. Ese es el caso del campo Día de la tabla FECHA, ya que, al no ser un campo relevante para el pronóstico de cosecha, se puede prescindir de él, considerando que este atributo era utilizado para el registro de la temperatura máxima, mínima y precipitación acumulada: datos climatológicos que fueron descartados por ser información inadecuada. Los demás registros que han sido insertados en cada tabla son datos reales y necesarios para la creación del modelo. Es importante recordar que todos los registros considerados son necesarios para el cumplimiento de los objetivos planteados.

Los campos seleccionados para el análisis son los siguientes:

- **FECHA**
  - IDFecha
  - Año
  - Mes
  - IdProducto
  - IdCampaña
- **PRODUCTO**
  - IDProducto
  - Producto
  - Equivalencia
- **PRECIO**
  - IDPrecio
  - Precio Promedio
  - Precio Minimo
  - Precio Màximo
  - IDProducto
  - IDFecha
- **CAMPAÑA**
  - IDCampaña
  - Sembrado (Ha)
  - Cosechado (Ha)
  - Producido (Ha)
  - Año
  - Mes

El motivo principal para incluir o excluir alguno de los campos, es, como se mencionó anteriormente, la forma en la que estos datos van a influir en los objetivos planteados previamente, los cuales se definieron en la fase 1 (Comprensión del negocio) y en la importancia de cada una de las variables para el modelo de pronóstico.

Limpieza de los datos

La fuente de datos original contiene todos los campos necesarios para cumplir con los objetivos definidos para la investigación. Luego de su proceso de migración mediante el proceso ETL se contará con una base de datos para el análisis. Se realizará un preprocesamiento de los datos con el fin de validar alguna inconsistencia entre las variables escogidas.

La existencia de campos donde falten valores nulos se justifica en la aparición de información que se desea representar como no existente considerando se cómo datos faltantes. Dichos valores serán tratados al momento de la ejecución de las técnicas de pronóstico al ignorarlas ya que estas no brindan información adicional a la investigación. La herramienta Spoon para la transformación y migración de los datos nos ayudará a eliminar registros vacíos y a la migración a una base de datos real.

Construcción de los datos

Registros generados:

Aparte de estas dos operaciones, no ha sido necesario generar nuevos atributos ni integrar nuevos registros a la base de datos ya que ésta está completa y ha sido creada específicamente para su uso en este proyecto.

Atributos Derivados:

El campo Día de la tabla Fecha no brinda información adicional a la investigación y, al contrario, la presentación de los datos complica la relación con otras tablas al tener únicamente campos que incluyen únicamente el mes y el año, por ejemplo, con la tabla Producto y Precio. Dicho campo fue ignorado y únicamente se consideró el mes y año.

Asimismo, el campo Mes de la tabla Fecha fue convertido en un valor numérico ya que este al ser una variable textual no cumpliría con los requisitos para poder generar el modelo en la herramienta

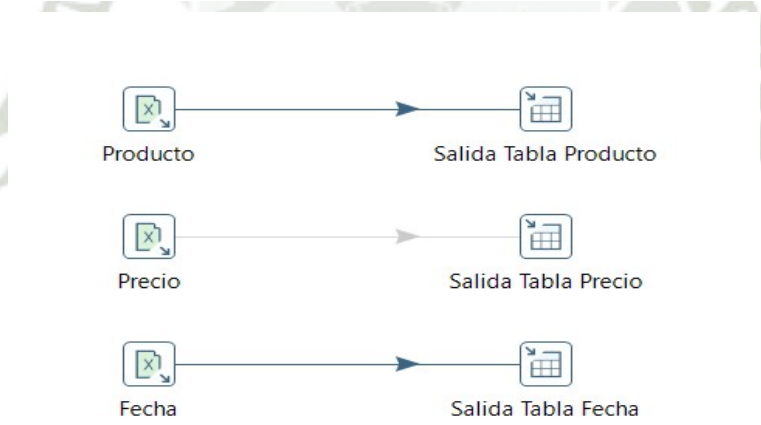
de modelado. De esta forma se hace más útil la información al no tener ni un valor textual o de tipo fecha, solamente numéricos.

Integración de los datos

Como la fuente de datos original era una hoja de cálculo con formato .xls, se utilizó un proceso ETL para realizar el proceso, mover los datos de esta fuente, reformatearlos, limpiarlos y cargarlos a una base de datos, como se ve en la Figura 9.

La herramienta escogida para el modelamiento es el programa SPSS IMB Modeler que a su vez también estará encargada de realizar las tareas de fusión entre distintas tablas de bases de datos si es necesario. De esta forma no hay necesidad de crear nuevas tablas, campos o registros de forma manual.

Figura 9
Diagrama de transformación Spoon para ETL de la fuente de datos



Nota: elaboración propia en base a información obtenida durante la investigación (2021).

Formato de los datos

El campo identificador con la información referente al tipo de producto o equivalencia de la cebolla roja ha sido codificado mediante valores numéricos ya que la herramienta SPSS IMB Modeler nos pide como requisito que los datos sean numéricos.

Al tener equivalencias del producto ya determinadas y no se tiene una gran cantidad de productos se optó por la asignación de un número por cada equivalencia del producto.

De esta forma, los códigos para cada uno de los productos existentes en la base de datos quedaron de la siguiente forma:

a) 1 → Kilo (1 Kg.)
b) 2 → Malla (40 Kg.)
c) 3 → Malla Grande (100 Kg.)
d) 4 → Malla Mediana (46 Kg.)
e) 5 → Saco (40,50,60 y 100 Kg.)
f) 6 → Saco Grande (100 Kg.)
g) 7 → Saco Mediano (46 Kg.)
h) 8 → Saco 1 (90 Kg.)

No fue necesario cambiar el orden de los campos de los registros de la base de datos. Para el preprocesamiento solo se consideró la reorganizó de los mismos en cada tabla. No fue necesario cambiar el formato de algún otro campo, ya que el formato de los campos que serán utilizados en el modelo es admitido por la herramienta de modelamiento.

Modelamiento

En esta fase de la metodología CRISP-DM se escogerá la técnica o conjunto de técnicas más adecuadas para cumplir con los objetivos de minería de datos propuestos en esta investigación.

Luego de la configuración de los parámetros para cada uno de los modelos que se generarán se aplicará las técnicas escogidas sobre los datos del modelo y finalmente se deberá evaluar si el modelo probado ha logrado cumplir con los criterios propuestos para ser considerado como exitosos o no.

Selección de la técnica de modelado

Como la herramienta modelo que se utilizará para la creación del modelo es el software SPSS IMB Modeler se procederá a utilizar alguna de las técnicas de modelado que nos ofrece dicha herramienta y que, a su vez, esta pueda cumplir con los objetivos de la investigación.

De los modelos revisados previamente de la herramienta SPSS IMB Modeler, la mejor técnica que se adapta a los objetivos establecidos es un modelo de red neuronal perceptrón multicapa (MLP) teniendo como referencia el estudio de Mhaskar (2016), ya que al ser una técnica no paramétrica nos ahorra el cumplimiento de supuestos, se puede procesar cualquier tipo de variables (adaptabilidad) y básicamente tiene un comportamiento como un aproximador de funciones.

Sin embargo, como el tipo de problema que se desea resolver es un problema de predicción y los campos de la fuente de datos contienen valores continuos también podría aplicarse un modelo de regresión. De esta forma, correlación de ambos modelos se podrá optar por uno o por otro de acuerdo con estos resultados.

Construcción del modelo

A continuación, se procede con la construcción del modelo escogido sobre los datos de entrenamiento. En esta sección se procede a describir la configuración de los parámetros del modelo que se escogerán de la herramienta IBM SPSS Modeler, de igual forma el resultado del modelo o modelos y su respectiva descripción.

a) **Ajuste de Parámetros**

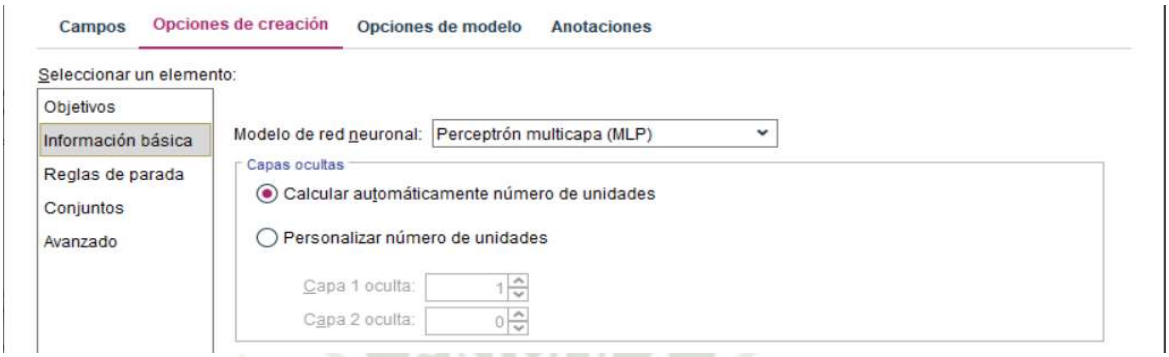
En el punto **3.2.1.3** se definieron dos objetivos para la creación de los modelos.

Estos objetivos estaban orientados a la generación de resultados orientados a la metodología de minería de datos elegida. Los parámetros que serán configurados dependerán de cada objetivo que se trabajará y estos variarán de acuerdo con los objetivos que se desean alcanzar.

En cuanto a los parámetros utilizados, solo se debió modificar la opción de creación del modelo para que la técnica de creación sea la de Perceptrón Multicapa (MLP), que también aparecen en la Figura 10. Los demás valores se conservaron con la configuración predeterminada.

Figura 10

Configuración de parámetros para la generación de modelo Perceptrón Multicapa MLP

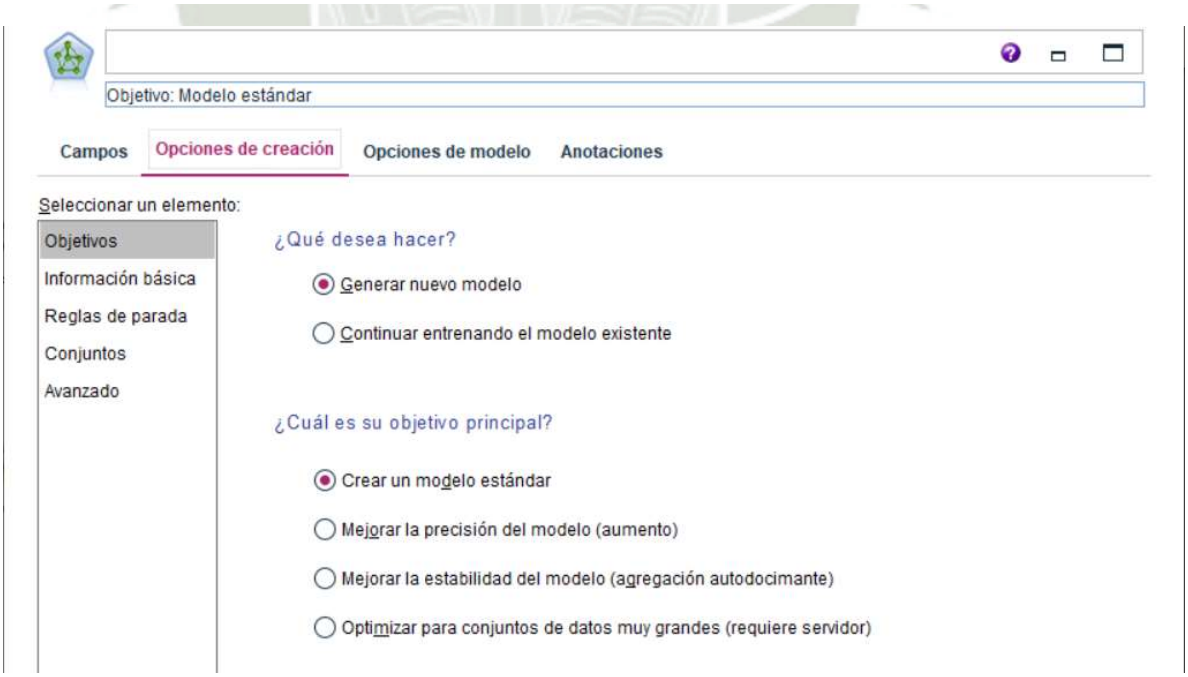


Nota: Esta figura muestra la configuración básica para la generación de un modelo MLP en SPSS Modeler.

Los objetivos establecidos para la generación del modelo de red neuronal están orientados a la creación de un nuevo modelo estándar para encontrar las relaciones entre los campos y este tipo de modelos son más fáciles de interpretar y suelen ser más rápidos de puntuar que otros con conjuntos aumentados, agregados o conjuntos grandes. Ver Figura 11.

Figura 11

Configuración de los objetivos de minería de datos para la generación de modelo MLP



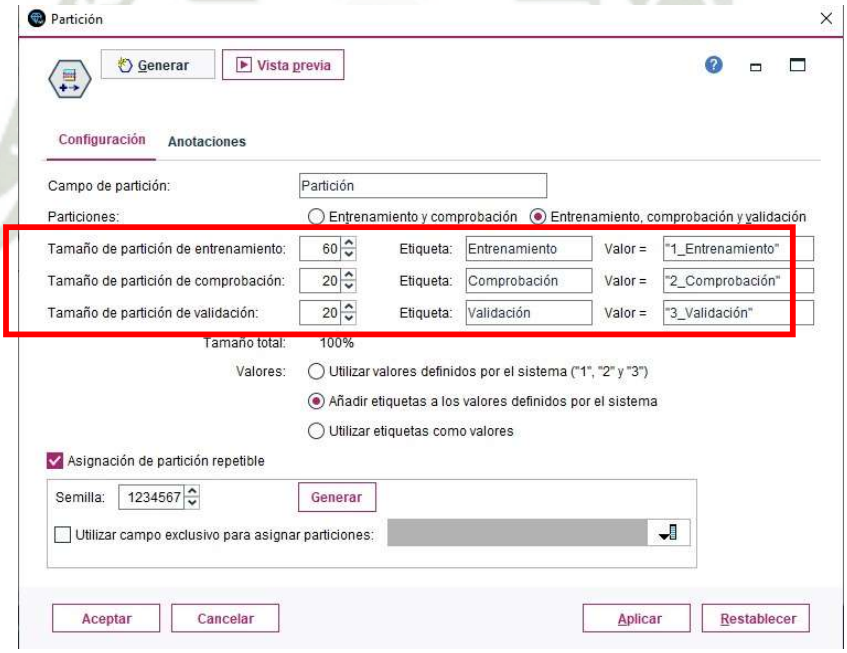
Nota: Esta figura muestra la configuración de los objetivos de predicción para la generación de un modelo MLP en SPSS Modeler.

En el nodo **Partición** se configuró el porcentaje de registros que serían utilizados para el entrenamiento del modelo, la comprobación y finalmente la validación de los modelos generados. El 60% de los registros serían utilizados para el entrenamiento y el 40% restante para la comprobación y validación (20% respectivamente para cada partición de comprobación y validación) Ver figura 12.

Dicha configuración se aplicará a todos los modelos generados y los registros de una partición determinada no estarán incluidos dentro de otra partición, lo que podría aumentar el nivel de error en las particiones con una menor cantidad de registros. La herramienta predeterminadamente asigna una etiqueta a los valores predichos para su comparación con los deseados.

Figura 12

Configuración de la partición para generación del modelo de predicción



Nota: Esta figura muestra la configuración de la partición para el entrenamiento, prueba y validación n de un modelo MLP para los objetivos de minería de datos.

Estos parámetros son los mismos para la generación de los modelos para el cumplimiento de los objetivos establecidos para la generación del modelo de red neuronal.

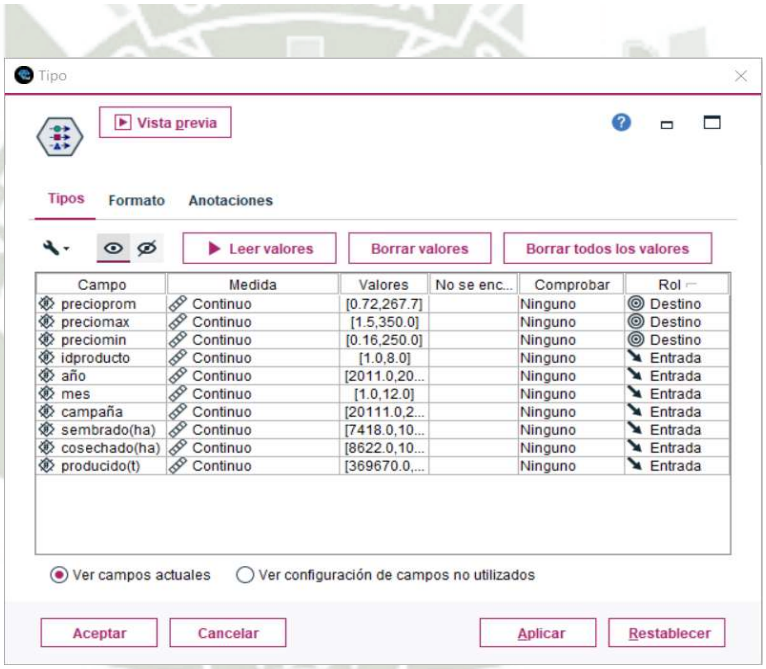
- **Objetivo 1:** Predicción del precio (promedio, mínimo y máximo) de la cebolla roja.

Para este caso, se tienen 3 subobjetivos. Existe la posibilidad que al intentar generar un modelo con estas 3 variables de predicción la precisión del modelo se reduzca considerablemente.

En el nodo Tipo utilizado para el control de metadatos del modelo se configuró cada una de estas variables, escogiendo "precio_minimo", "precio_promedio" y "precio_promedio", como las variables objetivas o "target" que se muestran en la figura 13.

Figura 13

Configuración de los objetivos del modelo para la predicción de precio en SPSS Modeler

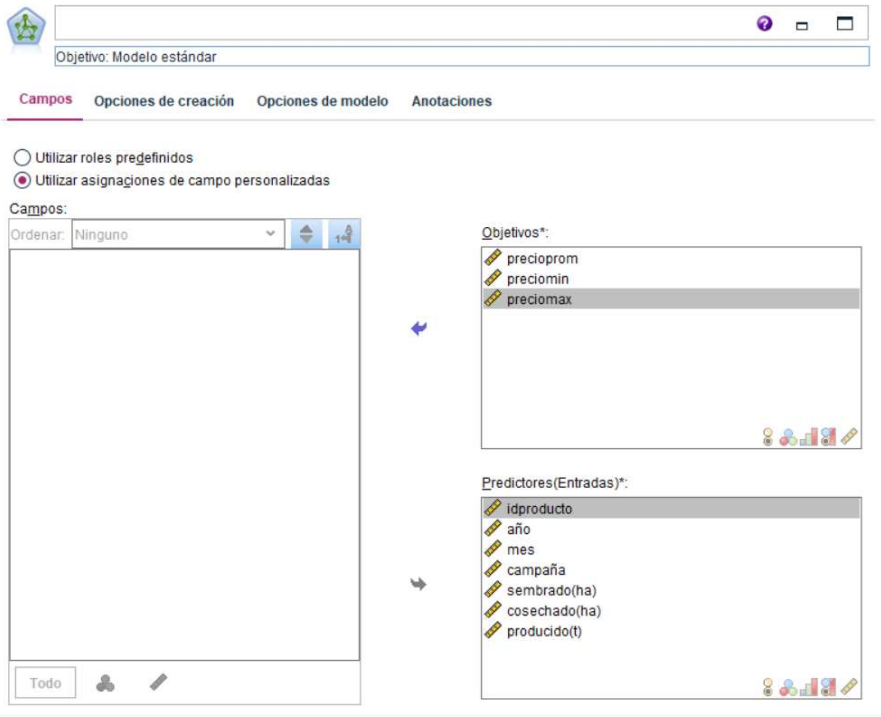


Nota: Esta figura muestra la configuración básica para la generación de un modelo MLP para el primer objetivo de minería de datos.

Se configuran de igual forma los objetivos para el modelo en el asistente de creación del modelo de red neuronal y poder alternarlos entre los valores predictores o entradas para el modelo, ver Figura 14.

Figura 14

Asistente de configuración para creación del modelo para la predicción de precio en SPSS Modeler



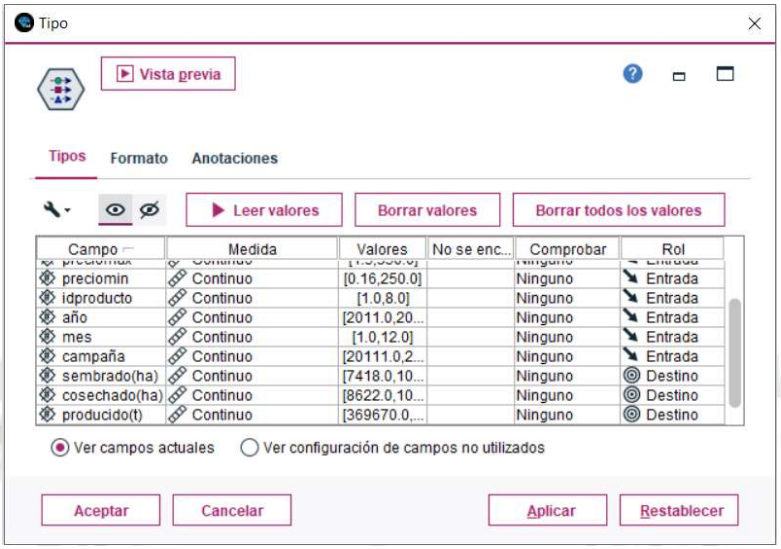
Nota: Esta figura muestra la configuración básica para la generación de un modelo MLP para el primer objetivo de minería de datos.

- **Objetivo 2:** Predecir la cantidad de hectáreas de cebolla roja sembradas, cosechadas y la producción total.

  Al igual que el primer modelo, existe la posibilidad que al intentar generar un modelo con estas 3 variables de predicción la precisión del modelo se reduzca considerablemente. En el nodo Tipo utilizado para el control de metadatos del modelo se configuró cada una de estas variables, escogiendo "cosechado(ha)", "sembrado(ha)" y "producido(t)", como las variables objetivo o "target" para este segundo modelo. Con los mismos datos de las variables predictoras y las variables "target" se configura el asistente de creación del modelo de red neuronal. ver Figura 15 y 16.

Figura 15

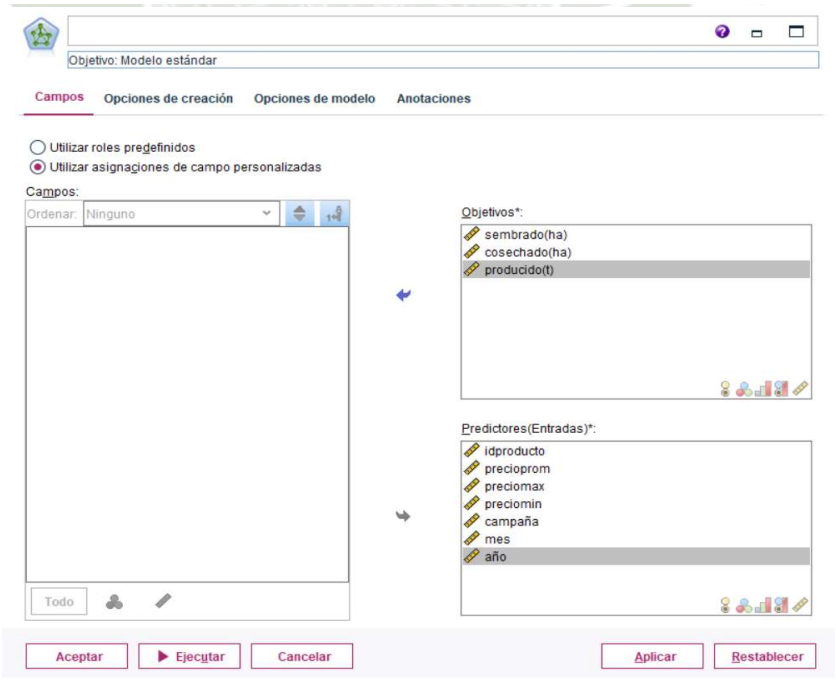
Configuración de los objetivos del modelo para la predicción de producción en SPSS Modeler



Nota: Esta figura muestra la configuración básica para la generación de un modelo MLP para el segundo objetivo de minería de datos.

Figura 16

Asistente de configuración para creación del modelo de predicción de precio



Nota: Esta figura muestra la configuración básica para la generación de un modelo MLP para el segundo objetivo de minería de datos.

Generación de modelos

Se ejecuta el nodo Red Neuronal para la creación de un modelo predictivo de red neuronal para cada uno de los objetivos de minería de datos establecidos previamente. Se ejecutan dichos modelos sobre un conjunto de datos de entrenamiento del 70%.

Eso nos deja un 30% de datos para el conjunto de pruebas. En total se tienen 1130 registros de los últimos años con los precios, extensión de terreno cosechada, sembrada y producida. Los detalles de la ejecución de cada modelo se muestran a continuación. Adicionalmente al primer análisis del modelo, se adjuntó el coeficiente de correlación de Pearson. La prueba de correlación de Pearson es una prueba realizada que calcula la relación estadística existente entre variables.

El coeficiente de correlación de Pearson puede tomar valores en un rango de +1 a -1, donde un valor mayor a 0 indica una asociación positiva, un valor menor a 0 una asociación negativa. De esta forma, podemos evaluar el grado de influencia de cada una de las variables estudiadas con relación a la variable objetivo para poder confirmar la teoría inicial o modificarla. (J. Hernández, 2018).

Modelo para el objetivo 1:

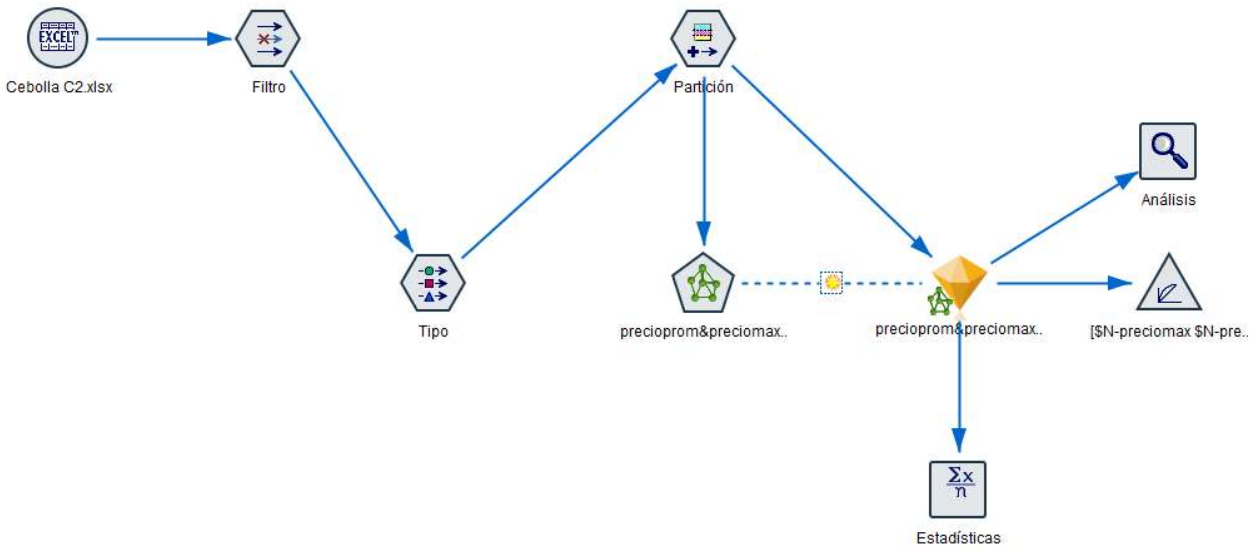
Para la creación del primer modelo en la herramienta SPSS Modeler se utilizaron 7 nodos para la configuración del modelo de red neuronal para el cumplimiento del primer objetivo de minería de datos. El primer nodo es la fuente de datos que se utilizará para la predicción. El nodo "Filtro" nos permite seleccionar el tipo de dato y el rol que cumplirá en la creación del modelo (variable de entrada o variable objetivo).

El nodo "Partición" nos permite definir el porcentaje de datos que serán utilizados para el entrenamiento del modelo y el porcentaje de datos que serán utilizados para la prueba del modelo. En ambos modelos se utiliza el 60% de los registros para la prueba y el 40% restante para la prueba.

El nodo de red neuronal que tiene el nombre del primer objetivo acepta la configuración de la red neuronal, el tipo de algoritmo, reglas de parada, configuración del predictor, entre otros.

Los últimos nodos (Análisis, Evaluación y Estadística) son nodos utilizados para la evaluación del modelo y los datos generados por el mismo. Se consideraron para poder comparar el modelo con otros generados, pero con otra configuración. Ver Figura 17.

Figura 17
Nodos utilizados para la generación del modelo para el objetivo 1.

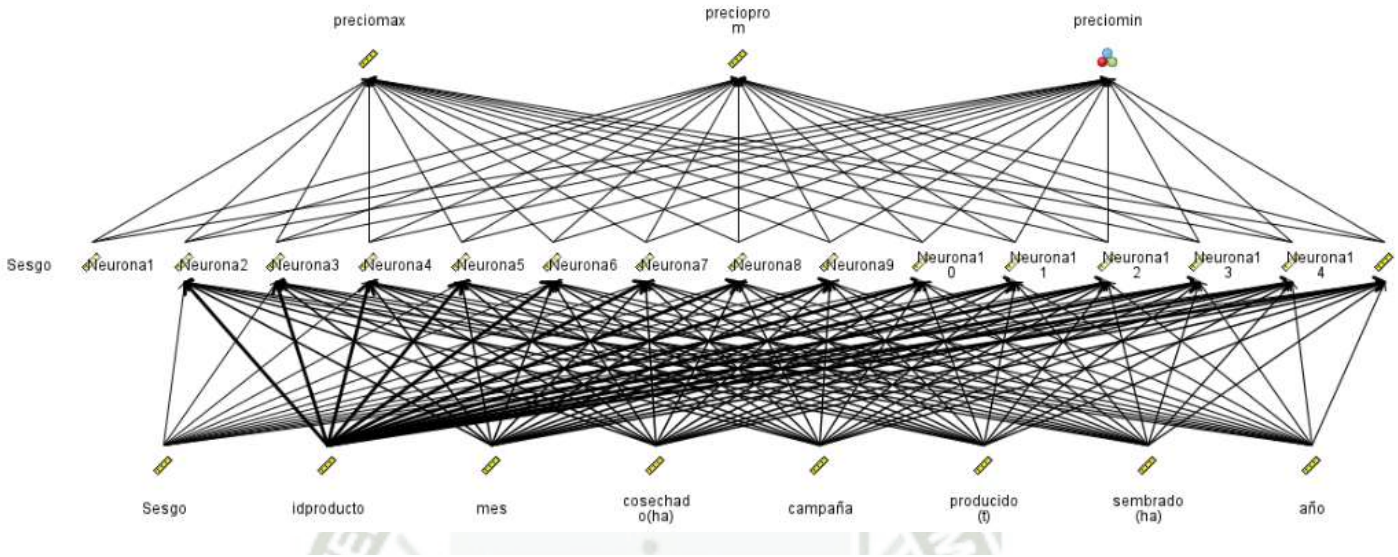


Nota: Esta figura muestra los nodos utilizados para la configuración y generación del modelo de red neuronal para el cumplimiento del objetivo 1 de minería de datos de la investigación.

La ejecución del quinto nodo nos devuelve un modelo de red neuronal con las variables de entrada o predictoras y las variables objetivo, ver Figura 18. El modelo utiliza 15 neuronas en la primera capa el porcentaje de precisión de este primer modelo alcanza el 93.4% luego del entrenamiento y prueba de los datos, cómo se puede apreciar en la Figura 18. Al ejecutar un análisis a cada una de las variables objetivo nos encontramos con una correlación lineal moderada, pues no es mayor al 60% ni menor al 50 %, como se puede apreciar en la tabla 12

Figura 18

Modelo de red neuronal generado para el objetivo 1 de minería de datos.



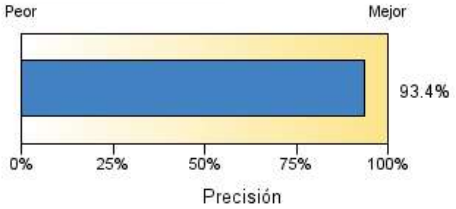
Nota: Esta figura muestra el modelo de red neuronal para el cumplimiento del objetivo 1 de minería de datos de la investigación.

Figura 19

Resumen del modelo generado para la predicción de precios de venta cebolla roja

Resumen del modelo

Objetivos	cosechado(ha) producido(t) sembrado(ha)
Modelo	Perceptrón multicapa
Regla de parada utilizada	El error no se puede disminuir más
Neuronas de la capa oculta 1	15



Nota: Esta figura muestra el resumen del primer modelo generado para el cumplimiento de los objetivos de minería de datos.

Tabla 12

Resumen del modelo para el primer objetivo de minería de datos – Precio Máximo

Partición	Entrenamiento	Comprobación
Error mínimo	-99,827	-82,327
Error máximo	222,964	140,215
Error promedio	-0,352	-0,874
Error absoluto promedio	35,581	35,987
Desviación estándar	45,842	47,066
Correlación lineal	0,607	0,531
Ocurrencias	909	221

Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción del precio máximo.

Tabla 13

Resumen del modelo para el primer objetivo de minería de datos – Precio Mínimo

Partición	Entrenamiento	Comprobación
Error mínimo	-87,588	-77,002
Error máximo	159,56	139,773
Error promedio	0,166	1,475
Error absoluto promedio	24,45	26,497
Desviación estándar	35,439	28,465
Correlación lineal	0,606	0,529
Ocurrencias	909	221

Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción del precio mínimo.

Tabla 14

Resumen del modelo para el primer objetivo de minería de datos – Precio Promedio

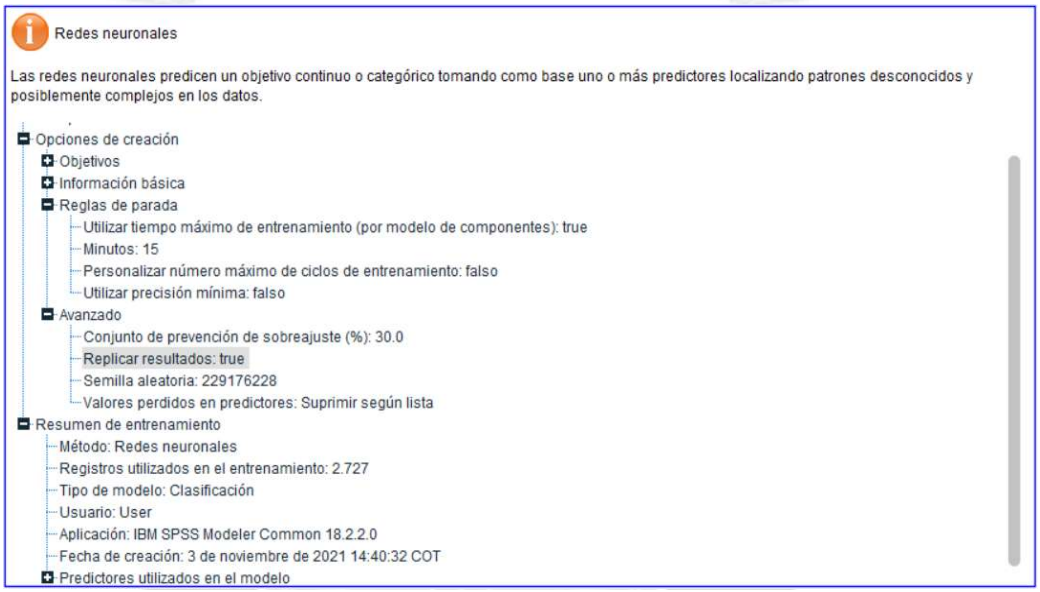
Partición	Entrenamiento	Comprobación
Error mínimo	-92,397	-67,736
Error máximo	161,719	138,75
Error promedio	0,875	1,761
Error absoluto promedio	27,796	29,844
Desviación estándar	39,013	41,049
Correlación lineal	0,623	0,558
Ocurrencias	909	221

Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción del precio promedio.

En las tablas 13 y 14 se muestran los resultados estadísticos de la generación del modelo para los objetivos del precio promedio y precio mínimo. Para todos los objetivos se tienen datos muy similares con un nivel de correlación mayora a 50% pero menor al 70%. Al tener 3 variables objetivo, el coeficiente de relación tiene a reducirse por reducir también la cantidad de variables predictoras.

Figura 20

Resumen de la generación del modelo para la predicción de precios



Nota: Esta tabla muestra el resumen de la generación del modelo para la predicción del precio mínimo, promedio y máximo.

En la figura 20 se puede apreciar el resumen del modelo generado para el cumplimiento del primer objetivo de minería de datos relacionado con la predicción de los precios de venta de cebolla roja (precio máximo, precio mínimo y precio promedio). Como resultados de este modelo se obtuvo un tiempo de ejecución de 15 minutos, el cual fue configurado previamente en el asistente de SPSS Modeler.

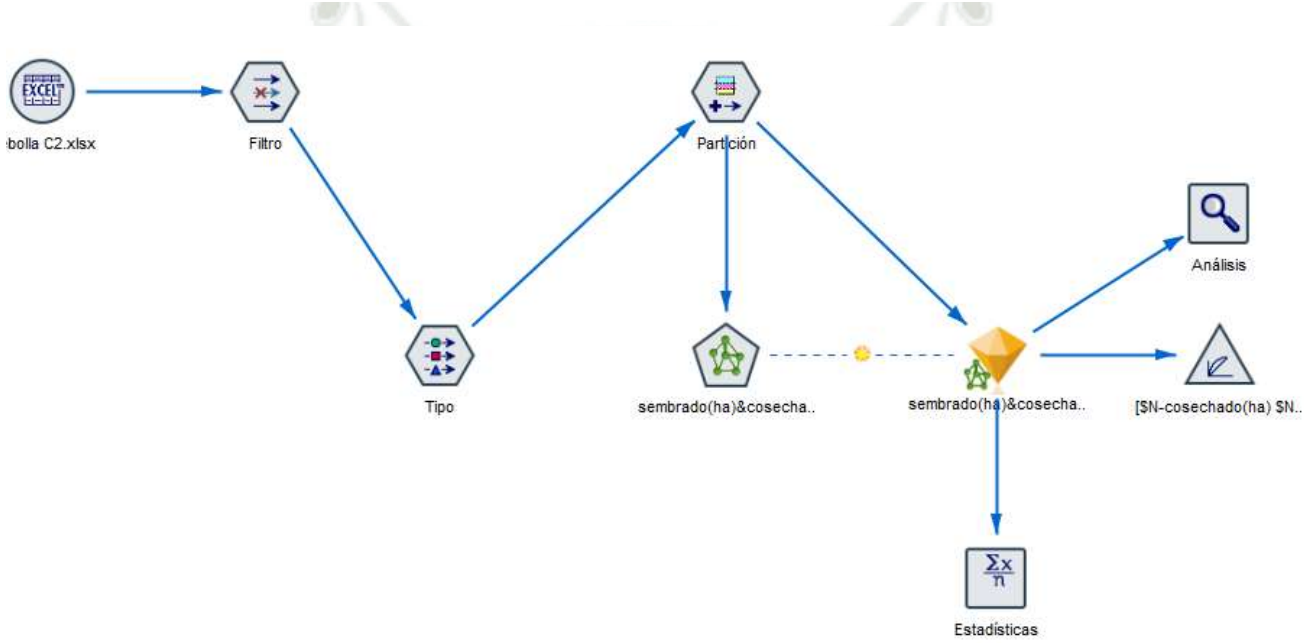
Adicionalmente en la parte de entrenamiento del modelo se utilizaron 2727 registros. Respecto al conjunto de sobreajuste de los registros para la generación del modelo, se destinó únicamente el 70% para el entrenamiento de este y el 30% de los registros para la comprobación o prueba.

Modelo para el objetivo 2:

Los nodos utilizados para la generación del segundo modelo de red neuronal se muestran en la figura 21 y su estructura es similar a la del primer modelo generado. Los únicos cambios que se realizaron fueron en el primer nodo, donde algunas de las variables predictoras cambiaron su rol pasando a ser las variables objetivo y viceversa.

Figura 21

Nodos utilizados para la generación del modelo para el objetivo 2

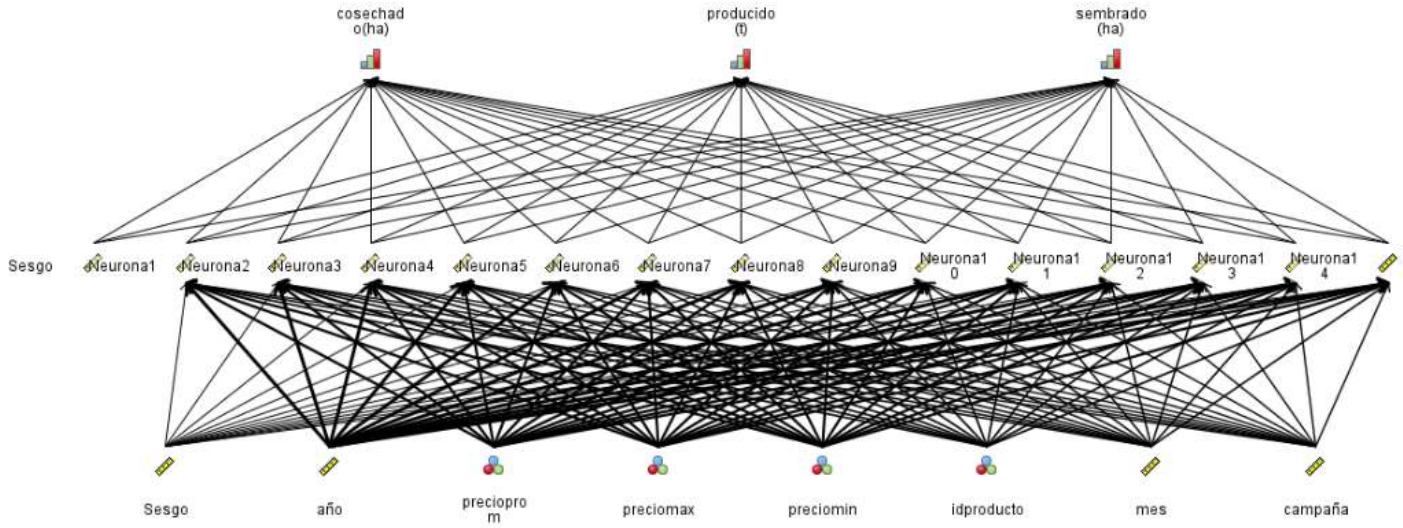


Nota: Esta figura muestra los nodos utilizados para la configuración y generación del modelo de red neuronal para el cumplimiento del objetivo 1 de minería de datos de la investigación.

Al igual que el primer modelo generado, la ejecución del quinto nodo nos devuelve un segundo modelo de red neuronal con las variables de entrada o predictoras y las variables objetivo. Este modelo también tiene 8 neuronas en la primera capa oculta y 4 neuronas en la segunda capa oculta. ver Figura 22. En la figura 23 se puede apreciar el porcentaje de precisión del modelo que alcanza el 57.0% luego del entrenamiento y prueba de los datos. Al ejecutar un análisis a cada una de las variables objetivo nos encontramos con una correlación lineal alta mayor al 90% en los 3 variables objetivos.

Figura 22

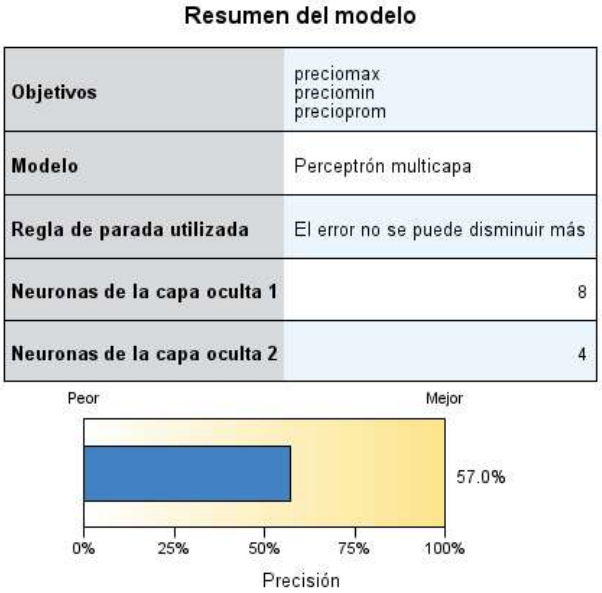
Modelo de red neuronal generado para el objetivo 2 de minería de datos.



Nota: Esta figura muestra el modelo de red neuronal para el cumplimiento del objetivo 2 de minería de datos de la investigación.

Figura 23

Resumen de la generación del modelo para la predicción de variables de producción



Nota: Esta figura muestra el resumen del segundo modelo generado para el cumplimiento de los objetivos de minería de datos.

Tabla 15

Resumen del modelo para el segundo objetivo de minería de datos – Superficie Cosechada

Partición	Entrenamiento	Comprobación
Error mínimo	-1326,64	-1284,248
Error máximo	308,647	315,86
Error promedio	-8,707	-22,551
Error absoluto promedio	128,232	141,08
Desviación estándar	185,194	201,206
Correlación lineal	0,939	0,924
Ocurrencias	793	337

Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción de la superficie cosechada

Tabla 16

Resumen del modelo para el segundo objetivo de minería de datos – Toneladas Producidas

Partición	Entrenamiento	Comprobación
Error mínimo	-40145,29	-38497,951
Error máximo	26409,477	26077,309

Error promedio	-76,435	-245,091
Error absoluto promedio	5976,018	6452,496
Desviación estándar	8404,615	8856,123
Correlación lineal	0,966	0,962
Ocurrencias	793	337

Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción de las toneladas producidas

Tabla 17

Resumen del modelo para el segundo objetivo de minería de datos - Superficie Sembrada

Partición	**Entrenamiento**	**Comprobación**
Error mínimo	-603,063	-766,299
Error máximo	663,401	613,599
Error promedio	-1,672	-12,941
Error absoluto promedio	153,169	153,707
Desviación estándar	193,992	192,039
Correlación lineal	0,949	0,948
Ocurrencias	793	337

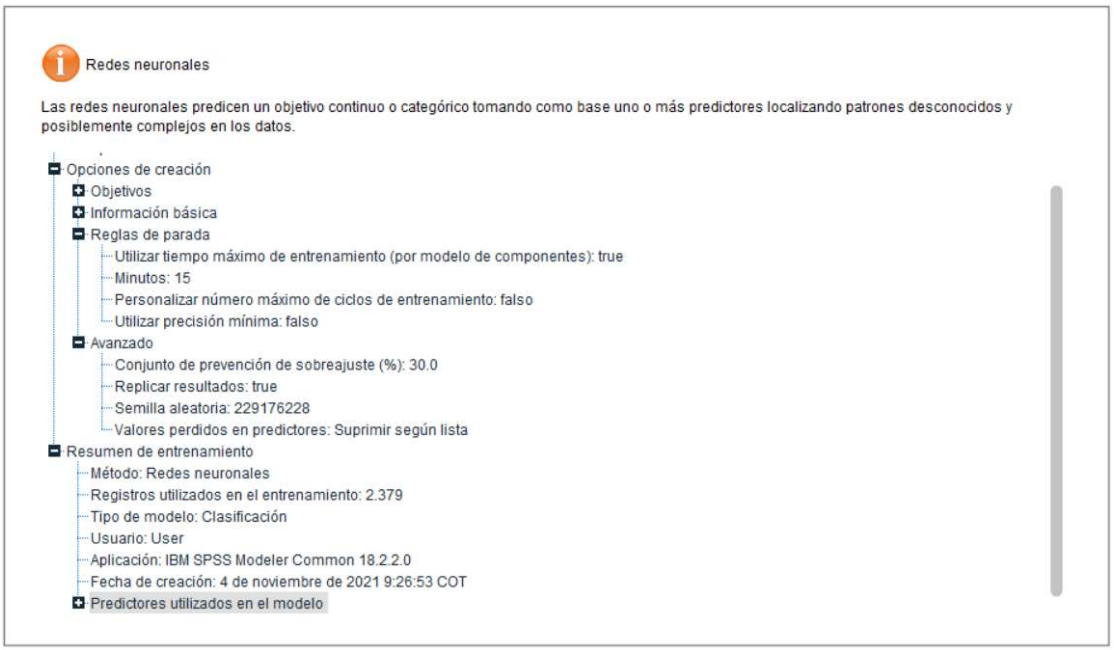
Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción de la superficie sembrada.

Asimismo, se generó el resumen estadístico del modelo generado por cada uno de los objetivos generados para el entrenamiento y comprobación de las variables objetivo (superficie cosechada, superficie sembrada y toneladas producidas) para el análisis posterior y evaluación del modelo. Para este caso en especial, el error mínimo que se aprecia en las tablas 15, 16 y 17 es muy elevado a diferencia del primer modelo generado para el cumplimiento del primer objetivo de minería de datos. Ver figura 24.

Esto se debe a que el tipo de dato escogido es uno de tipo clasificatorio, por lo que la herramienta al realizar el análisis estadístico sobre estas variables siempre tendrá un valor elevado dada a la cantidad limitada de registros en esta variable.

Figura 24

Resumen del entrenamiento del modelo para el objetivo 2



Nota: Esta tabla muestra el resumen de la generación del modelo para la predicción de las hectáreas sembradas, hectáreas cosechadas y toneladas producidas.

Descripción de los modelos

- **Modelo generado para el objetivo 1**

Este modelo ha devuelto los siguientes resultados:

-Calidad de la red neuronal: El modelo perceptrón multicapa alcanza una precisión del 57.1 % con la utilización de 16 neuronas en la capa oculta y un tiempo de ejecución de 15 minutos. Como los objetivos son continuos, la precisión del modelo se especifica como R^2.

-El error absoluto promedio: Nos muestra la media de los valores absolutos de todos los errores de los registros. Estos tienen los valores de 1,481 para el primer objetivo del modelo 1 (precio_prom), 1,702 para el segundo objetivo (precio_max) y 1,331 para la tercera variable (precio_min) luego de la ejecución del modelo sobre los datos de entrenamiento y prueba.

-El error promedio: Nos muestra el promedio de error en todos los registros procesados. Este valor nos ayuda a identificar un posible sesgo sistemático en el modelo, el cual tiene los valores que se muestran en la tabla 18.

Tabla 18

Error promedio obtenido para los objetivos del primer modelo

	Entrenamiento	Comprobación
precio_minimo	-1.07	-2.149
precio_maximo	-0.448	-3.487
precio_promedio	-1.07	-2.149

Nota: elaboración propia en base a información obtenida durante la investigación (2021).

Coeficiente de correlación del objetivo 1

Se puede ver la existencia de una fuerte relación entre el tipo de producto, el año y campaña tanto para la predicción del precio promedio como para el precio máximo. Mientras que para el precio mínimo se incluyen las variables predictoras antes mencionadas y se encuentra una fuerte relación con la cantidad de hectáreas cosechadas.

Tabla 19

Correlación de variables predictoras con la variable "precio_promedio"

predictor	coeficiente	nivel de correlación
id producto	0.339	Fuerte
año	0.007	Débil
campaña	-56	Medio
sembrado (ha)	0.014	Débil
cosechado(ha)	0.026	Débil
mes	0.04	Débil

Nota: Esta tabla muestra el nivel de correlación de las variables predictoras con el "precio promedio"

Tabla 20

Correlación de variables predictoras con la variable "precio_mínimo"

Predictor	Coeficiente	Nivel de correlación
id producto	0.317	Fuerte
año	0.050	Medio
campaña	-0.064	Fuerte
sembrado (ha)	0.027	Débil
cosechado(ha)	0.064	Fuerte
mes	0.031	Débil

Nota: Esta tabla muestra el nivel de correlación de las variables predictoras con el "precio mínimo"

Tabla 21

Correlación de variables predictoras con la variable "precio_máximo"

Predictor	Coeficiente	Nivel de correlación
id producto	0.346	Fuerte
año	-0.017	Débil
campaña	-0.052	Medio
sembrado (ha)	0.048	Débil
cosechado(ha)	0.006	Débil
mes	0.033	Débil

Nota: Esta tabla muestra el nivel de correlación de las variables predictoras con el "precio máximo"

- **Modelo generado para el objetivo 2**

Este modelo ha devuelto los siguientes resultados:

-La Calidad de la red neuronal: El modelo perceptrón multicapa para este objetivo alcanza el 81.5 % con la utilización de 16 neuronas en la capa oculta y un tiempo de ejecución de 15 minutos. Los cambios de extensión de terreno sembrada, cosechada y producida son objetivos son continuos, de tal forma que la precisión del modelo se especifica como R^2.

-El error absoluto promedio: La media de los valores absolutos de todos los errores de los registros para el objetivo cosechado(ha) es 164.993, para el segundo objetivo producido(ha) es 10193.979 y para la última variable (sembrado) es 238.464 luego de la ejecución del modelo sobre los datos de entrenamiento y prueba.

-El error promedio: Nos muestra el promedio de error en todos los registros procesados. Este valor nos ayuda a identificar un posible sesgo sistemático en el modelo.

Para este caso, los resultados obtenidos indican un error promedio fuera de lo normal y dicha observación se considerará para la parte evaluativa del modelo.

Coeficiente de correlación del objetivo 2

Se puede ver la existencia de una fuerte relación entre el año, la campaña y el mes para la predicción de la cantidad de hectáreas sembradas, ver Tabla 22. Para los otros dos objetivos se repite la variable predictora del año y para el objetivo de hectáreas cosechadas se añade el precio mínimo como el otro predictor con una fuerte correlación, ver Tabla 22. Finalmente, para la cantidad de toneladas cosechadas se tienen como variables predictoras con un alto nivel de correlación al precio máximo, el año y el tipo de producto, ver Tabla 23.

Tabla 22

Correlación de variables predictoras con la variable "sembrado"

Predictor	Coeficiente	Nivel de correlación
idproducto	0.043	Débil
preciomax	0.048	Débil
preciomin	-0.027	Débil
año	-0.274	Fuerte
campaña	-0.091	Fuerte
mes	-0.108	Fuerte
precioprom	0.014	Débil

Nota: Esta tabla muestra el nivel de correlación de las variables predictoras con la variable "sembrado ha"

Tabla 23

Correlación de variables predictoras con la variable "cosechado"

Predictor	Coeficiente	Nivel de correlación
idproducto	0.006	Débil
preciomax	0.046	Débil
preciomin	-0.064	Fuerte
año	-0.217	Fuerte
campaña	-0.032	Débil
mes	-0.022	Débil
precioprom	-0.026	Débil

Nota: Esta tabla muestra el nivel de correlación de las variables predictoras con la variable "cosechado ha"

Tabla 24

Correlación de variables predictoras con la variable "producido"

Predictor	Coeficiente	Nivel de correlación
idproducto	0.060	Fuerte

preciomax	0.070	Fuerte
preciomin	-0.020	Débil
año	-0.421	Fuerte
campaña	-0.004	Débil
mes	0.018	Débil
precioprom	0.030	Débil

Nota: Esta tabla muestra el nivel de correlación de las variables predictoras con la variable "cosechado ha"

En todos los casos existe una correlación fuerte con la variable "año". Como se puede aprecia en la tabla 24, la variable identificadora del tipo de producto y la variable de precio máximo tiene una fuerte correlación con la variable de este objetivo, pero tiene una débil correlación que, para este caso, se esperaba tener una mejor correlación con el mes o la campaña, pues son indicadores con un impacto importante en la producción de cebolla durante el año.

1.1.1. EVALUACIÓN DEL MODELO

Si bien en la siguiente parte de la investigación se realiza la evaluación de los modelos generados de acuerdo con la metodología CRISP-DM, en esta sección se realiza una evaluación que está enfocada en el cumplimiento de los objetivos de minería de datos establecidos previamente.

Para la siguiente parte de la metodología, la evaluación estará enfocada únicamente en el cumplimiento de los objetivos del negocio. Para la evaluación de la efectividad de los modos generados se utilizaron dos indicadores que son el error absoluto promedio y la desviación estándar de los errores. Asimismo, SPSS Modeler nos brinda más información de los modelos generados que podría ser de utilidad en la etapa de evaluación y generación del informe final.

Estos valores son la precisión del modelo final (calidad de la red neuronal), expresada como R^2 y la correlación lineal entre los valores reales y los predichos. El primer modelo generado para poder cumplir con el primer objetivo de minería de datos obtiene una precisión del 53.9% para el modelo de red neuronal Perceptrón Multicapa.

Sin embargo, los valores que obtiene este modelo para el error absoluto en la partición de entrenamiento (29.766) y en la participación de comprobación (34.734). De la cantidad de datos utilizados, una partición del 70% del total fue utilizada para el entrenamiento del modelo y el 30% restante para la comprobación.

La desviación estándar de los errores nos devuelve un valor del 39.18 y 43.98 para la partición de entrenamiento y comprobación respectivamente. Ambas con valores muy elevados como se puede apreciar.

A pesar de que el resultado obtenido en la precisión del modelo fue mayor al 50%, los resultados de los errores obtenidos en el modelo hacen creer que los parámetros del modelo requieren ser reajustados, pues las medidas establecidas para los campos objetivos no podrían ser las adecuadas.

Los parámetros que se configuraron para las medidas de los objetivos de este modelo y del modelo anterior fueron de tipo ordinal, pues estas variables contienen varios valores diferentes que mediante el nivel de medición ordinal permitirá crear un grupo de datos categóricos que clasifique el precio y la cantidad de cebolla cosechada de acuerdo con una clasificación natural de estos elementos.

De este modo, para el segundo modelo generado se obtiene una precisión del 76.3 % para el modelo de red neuronal Perceptrón Multicapa. Del mismo modo como se presentaron los resultados de los errores obtenidos en el modelo, tanto como para la partición de entrenamiento como para la partición de comprobación, el error absoluto tiene un valor de 19.121 y para la desviación estándar un valor de 15.887.

Los resultados obtenidos en esta primera evaluación no son los suficientes para garantizar un buen pronóstico para el precio y cosecha de cebolla en una temporada futura por los valores de los errores obtenidos.

Existe una probabilidad muy alta a que esto se deba a que los objetivos escogidos para la generación del modelo no tengan una relación directa con los demás objetivos del mismo modelo o que la medida escogida para las variables no sean las adecuadas. Por tal motivo no se descartará ninguno de los objetivos establecidos y se trabajará en la generación de modelos con las variables objetivo o "público" de forma independiente y con medidas adecuadas al tipo de resultado que se desea obtener.

CAPÍTULO II
RESULTADOS

EVALUACIÓN

Para esta parte de la metodología CRISP – DM se procederá a evaluar los modelos generados por la herramienta de minería de datos. A diferencia de la primera evaluación, esta sección busca evaluar los modelos generados desde el punto de vista de los objetivos del negocio.

Al finalizar esta evaluación se tendrá que decidir si los objetivos han sido cumplidos para poder avanzar a la fase de Implantación, caso contrario se deberá identificar cualquier falla que haya podido ocurrir y que se haya ignorado y hacer una revisión del proyecto.

EVALUACIÓN DE RESULTADOS

Desde el punto de vista de los objetivos establecidos para la investigación o los objetivos del negocio, se estableció como criterio de éxito el poder realizar pronóstico con un nivel de fiabilidad aceptable. Para que un pronóstico pueda ser catalogado como aceptable o no, dichos pronósticos deberán tener una base que sostenga este criterio, para lo cual se tendrá como base fundamental a los datos estadísticos que fueron analizados al momento de la ejecución de los modelos.

De esta forma la evaluación de los resultados de los modelos generados se basará principalmente en los indicadores generados por la herramienta SPSS Modeler y se procederá a evaluar los modelos generados y de encontrar alguno que no cumpla con los requisitos establecidos en esta parte se descartará y no será incluido en la parte de la Implementación.

- **Modelo generado para el objetivo 1:**

El modelo generado para el cumplimiento del objetivo 1 del proyecto es aceptable pues se podría realizar pronósticos de los precios mínimos, precio promedio y precio máximo que podría alcanzar la cebolla roja en temporadas futuras. Este pronóstico tendría un grado de fiabilidad del 53.9%, medida que desde el punto de vista de los objetivos del negocio es aceptable.

- **Modelo generado para el objetivo 2:**

El modelo generado para el cumplimiento del objetivo 2 del proyecto es aceptable pues se podría realizar pronósticos de la superficie de cebolla que será sembrada, la superficie que será cosechada y las toneladas de cebolla que serán producidas en temporadas de siembra y cosecha futuras. Este pronóstico tendría un grado de fiabilidad del 76.3%, medida que desde el punto de vista de los objetivos del negocio también es aceptable.

REVISIÓN DEL PROCESO

El proceso que se ha ejecutado hasta el momento se cumplió tal y como fue previsto. Ambos objetivos propuestos para el pronóstico de variables de cosecha fueron atendidos con modelos de red neuronales perceptrón multicapa (MLP) que alcanzaron un grado de precisión del 53.9% y 76.3% respectivamente. Al momento de revisar los datos de los modelos generados se obtuvieron valores defectuosos para el error absoluto promedio y la desviación estándar de los errores.

Como se mencionó anteriormente estos valores puedan deberse a que los objetivos escogidos para la generación del modelo no tengan una relación directa con los demás objetivos del mismo modelo o que la medida escogida para las variables no sean las adecuadas o podrían jugar un rol como predictoras y no como otros objetivos del mismo modelo.

DETERMINAR LOS PRÓXIMOS PASOS

El siguiente paso para realizar es la generación de los modelos para el cumplimiento de los objetivos 1 y 2. Los parámetros para la creación de los modelos serán modificados con el fin de reducir los errores generados por cada modelo. Esta modificación implicará específicamente la generación de un modelo de red neuronal por cada objetivo, es decir, por cada variable objetivo-establecida, se espera generar un único modelo que aumente el nivel de correlación de las variables, reduzca la cantidad de errores y mejorar la precisión del modelo.

De esta forma se repite el procedimiento desde la configuración de los parámetros antes de la generación de los modelos estableciendo 3 variables objetivo por cada objetivo de minería de datos. Se configuró nuevamente la medida o el tipo de dato de cada una de las variables utilizadas para la generación del modelo.

En la figura 25 se aprecia como se modificó la medida de los campos de precio promedio, precio mínimo y precio máximo para que tomen un valor **Nominal** que servirá para describir los datos

almacenados en estas variables y que deben ser tratados como un pequeño conjunto de datos (International Business Machines Corporation, 2021).

De igual forma, para los valores de cantidad de hectáreas sembradas, cantidad de hectáreas cosechadas y producción generada se cambió el tipo de medida a **Ordinal** debido al orden de clasificación de números enteros que tendrán estas variables y que servirán para la definición de este conjunto de datos categóricos para la visualización y generación de modelos (IBM, 2021).

Figura 25

Campos y medidas modificadas para la generación de modelos para los objetivos de minería de datos

Campo	Medida	Valores	No se enc...	Comprobar	Rol
precioprom	Nominal	<Leer>		Ninguno	Entrada
preciomax	Nominal	1.5,1.6,1.7,...		Ninguno	Entrada
preciomin	Nominal	0.16,0.2,0....		Ninguno	Entrada
idproducto	Categórico	<Leer>		Ninguno	Entrada
año	Continuo	[2011.0,20...		Ninguno	Entrada
mes	Continuo	[1.0,12.0]		Ninguno	Entrada
campaña	Continuo	[20111.0,2...		Ninguno	Entrada
sembrado(ha)	Ordinal	7418.0,876...		Ninguno	Entrada
cosechado(ha)	Ordinal	8622.0,872...		Ninguno	Destino
producido(t)	Ordinal	369670.0,3...		Ninguno	Entrada

Nota: Esta tabla muestra los campos y medidas escogidas para el cumplimiento de los objetivos de minería de datos.

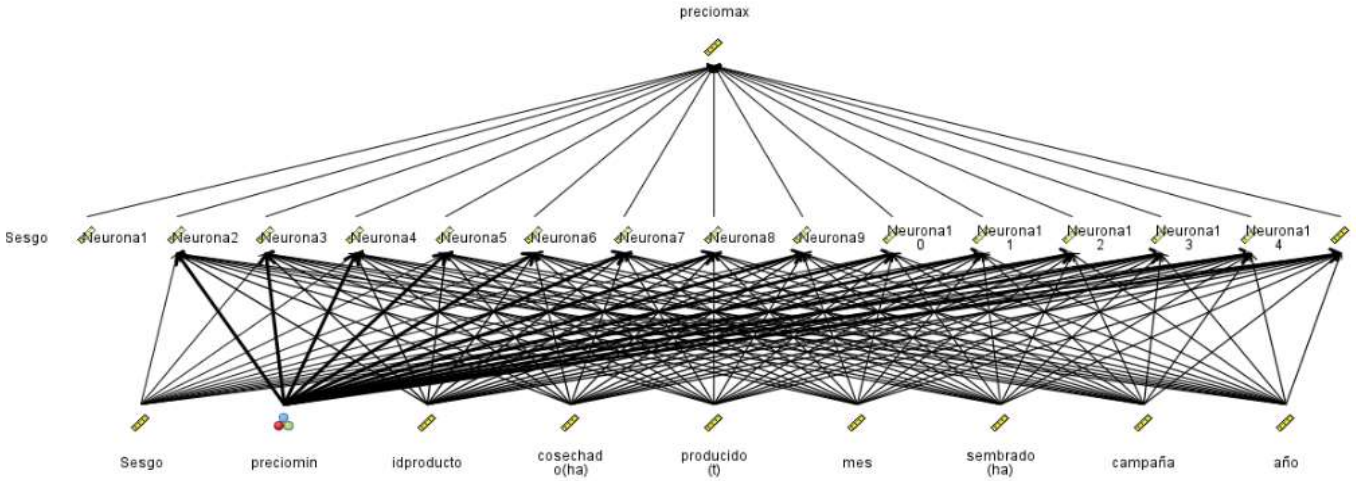
De igual forma, para los valores de cantidad de hectáreas sembradas, cantidad de hectáreas cosechadas y producción generada se cambió el tipo de medida a **ordinal** debido al orden de clasificación de números enteros que tendrán estas variables y que servirán para la definición de este conjunto de datos categóricos para la visualización y generación de modelos.

Modelo para el objetivo 1- Precio Máximo:

El modelo de red neuronal que se generó para el primer objetivo – precio máximo se muestra en la Figura 26. Este fue generado con la misma configuración que anteriores modelos, pero solo se seleccionó la variable de Precio Máximo de la fuente de datos como la variable objetivo a predecir.

Figura 26

Modelo de red neuronal generado para la predicción del precio máximo

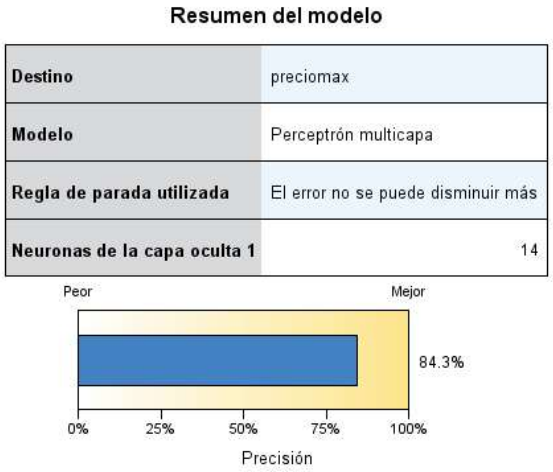


Nota: Esta figura muestra el modelo de red neuronal generado por la herramienta SPSS Modeler para la predicción del precio máximo.

-Calidad de la red neuronal: El modelo perceptrón multicapa alcanza una precisión del 84.3 % con la utilización de 14 neuronas en la capa oculta y un tiempo de ejecución de 15 minutos. Ver Figura 27. A diferencia del modelo generado para el primer objetivo, la calidad de la red neuronal aumenta al considerar únicamente una variable objetivo. Esto debido a que no hubo ningún estudio previo para poder encontrar alguna relación entre las variables objetivo-escogidas y solamente fueron consideradas por la hipótesis que se manejaba.

Figura 27

Resumen del modelo generado para el primer objetivo (Precio Máximo)



Nota: Esta tabla muestra los campos y medidas escogidas para el cumplimiento de los objetivos de minería de datos.

Tabla 25

Resumen de las estadísticas encontradas para el primer objetivo (Precio Máximo)

Estadísticas	Valor
Recuento	0.060
Media	92.005
Mín.	1.500
Máx.	350.000
Rango.	348.500
Varianza	57.211
Desviación estándar	1.702
Modo	100.000

Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción del precio máximo.

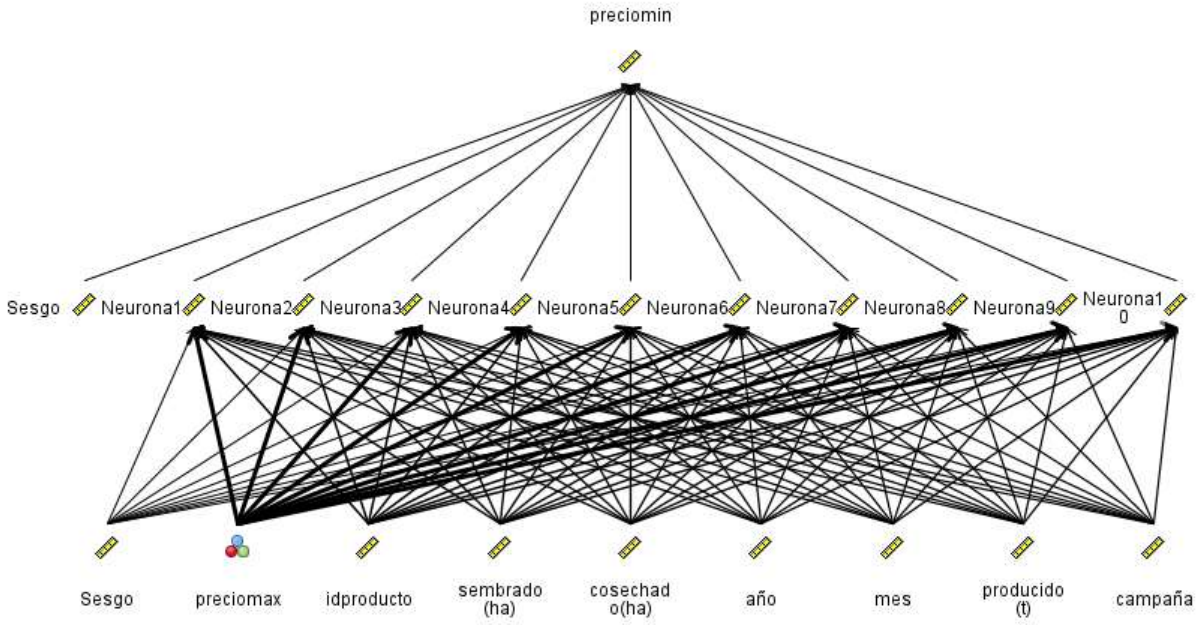
-El error estándar de la media: Nos muestra la incertidumbre encontrada al estimar la media de la variable en todos los registros. Este tiene un valor de 1.702. Ver tabla 27.

Modelo para el objetivo 1- Precio Mínimo:

El modelo de red neuronal que se generó para la predicción del precio mínimo se muestra en la Figura 28. Este fue generado con la misma configuración que anteriores modelos, pero a diferencia de otros modelos, este fue generado con solamente 10 neuronas en la capa oculta. Esto con la finalidad de mejorar la precisión del modelo

Figura 28

Modelo de red neuronal generado para la predicción del precio mínimo



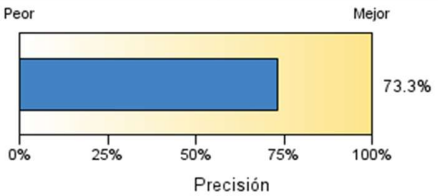
Nota: Esta figura muestra el modelo de red neuronal generado por la herramienta SPSS Modeler para la predicción del precio mínimo.

Figura 29

Resumen del modelo generado para el primer objetivo (Precio Mínimo)

Resumen del modelo

Destino	preciomin
Modelo	Perceptrón multicapa
Regla de parada utilizada	El error no se puede disminuir más
Neuronas de la capa oculta 1	14



Nota: Esta tabla muestra el resumen de la creación de un modelo de red neuronal para la predicción del precio mínimo.

-Calidad de la red neuronal: El modelo perceptrón multicapa alcanza una precisión del 73.3 % con la utilización de 14 neuronas en la capa oculta y un tiempo de ejecución de 15 minutos. Ver Figura 29.

-El error estándar de la media: Nos muestra la incertidumbre encontrada al estimar la media de la variable en todos los registros. Este tiene un valor de 1.331. Ver Tabla 26.

Tabla 26

Resumen de las estadísticas encontradas para el primer objetivo (Precio Mínimo)

Estadísticas	Valor
Recuento	0.060
Media	92.005
Mín.	1.500
Máx.	350.000
Rango.	348.500
Varianza	57.211
Desviación estándar	1.702
Modo	100.000

Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción del precio mínimo.

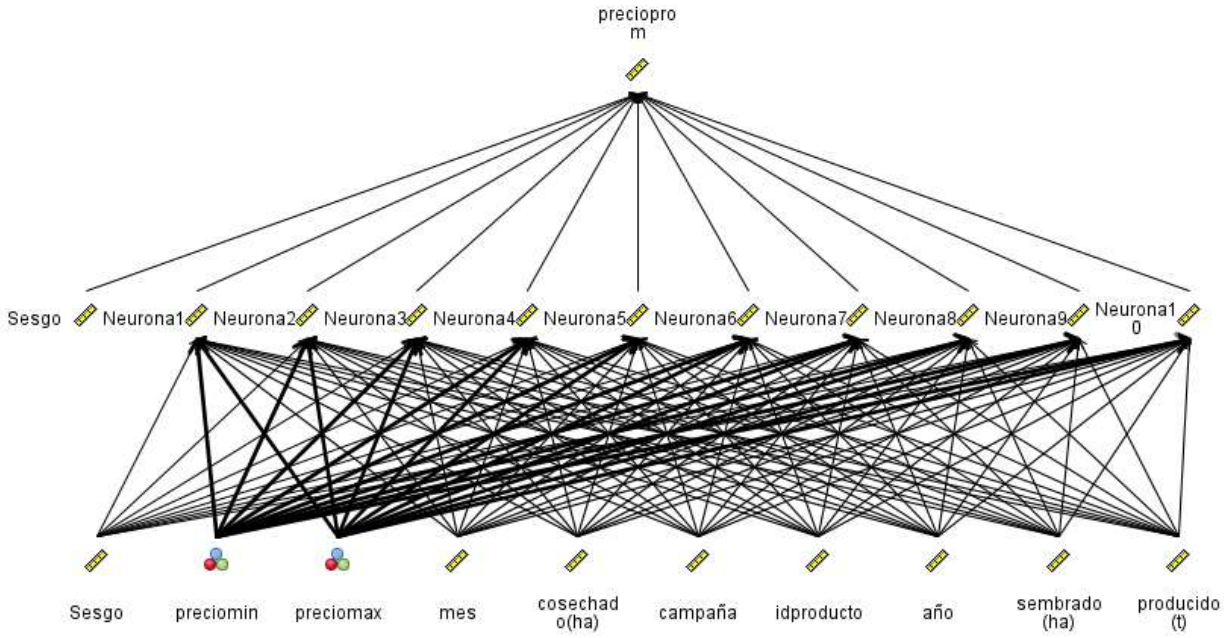
Modelo para el objetivo 1- Precio Promedio:

El modelo de red neuronal que se generó para la predicción del precio **promedio** se muestra en la Figura 27. Se configuró para que en el modelo tenga únicamente 10 neuronas en la capa oculta ante la mejora de la calidad del modelo de red neuronal generado anteriormente. Este modelo ha devuelto los siguientes resultados:

-Calidad de la red neuronal: El modelo perceptrón multicapa alcanza una precisión del 73.3 % con la utilización de 14 neuronas en la capa oculta y un tiempo de ejecución de 15 minutos. Como los objetivos son continuos, la precisión del modelo se especifica como R^2. Ver Figura 30.

Figura 30

Modelo de red neuronal generado para la predicción del precio promedio



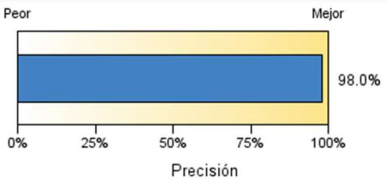
Nota: Esta figura muestra el modelo de red neuronal generado por la herramienta SPSS Modeler para la predicción del precio promedio.

Figura 31

Resumen del modelo generado para el primer objetivo (Precio Promedio)

Resumen del modelo

Destino	precioprom
Modelo	Perceptrón multicapa
Regla de parada utilizada	El error no se puede disminuir más
Neuronas de la capa oculta 1	13



Nota: Esta tabla muestra el resumen de la creación de un modelo de red neuronal para la predicción del precio promedio.

-El error estándar de la media: Nos muestra la incertidumbre encontrada al estimar la media de la variable en todos los registros. Este tiene un valor de 1.481. Ver tabla 27.

Tabla 27

Resumen de las estadísticas encontradas para el primer objetivo (Precio Promedio)

Estadísticas	Valor
Recuento	1130
Media	80.687
Mín.	0.720
Máx.	266.980
Rango.	2478.042
Varianza	49.780
Desviación estándar	1.481
Modo	100.000

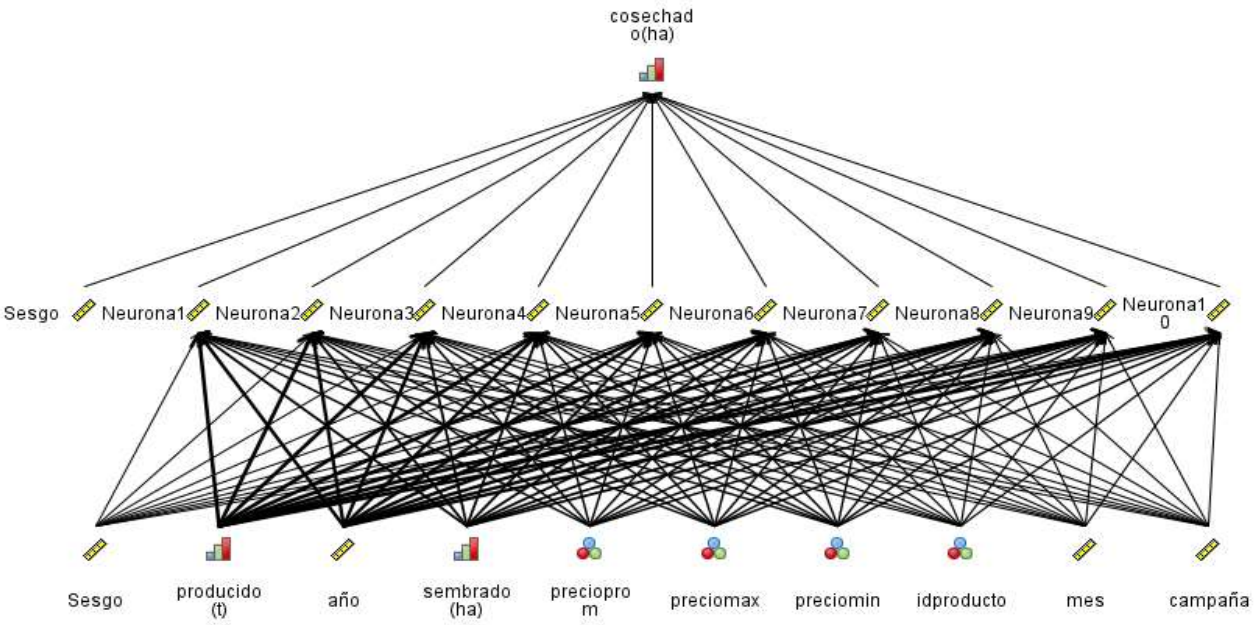
Nota: Esta tabla muestra el resumen de las estadísticas del modelo generado para la predicción del precio promedio.

Modelo para el objetivo 2- Hectáreas cosechadas:

El modelo de red neuronal que se generó para la predicción de las hectáreas cosechadas se muestra en la figura 32. Al igual que el modelo anterior, este modelo incluye 10 neuronas en la capa oculta.

Figura 32

Modelo de red neuronal generado para la predicción de hectáreas cosechadas

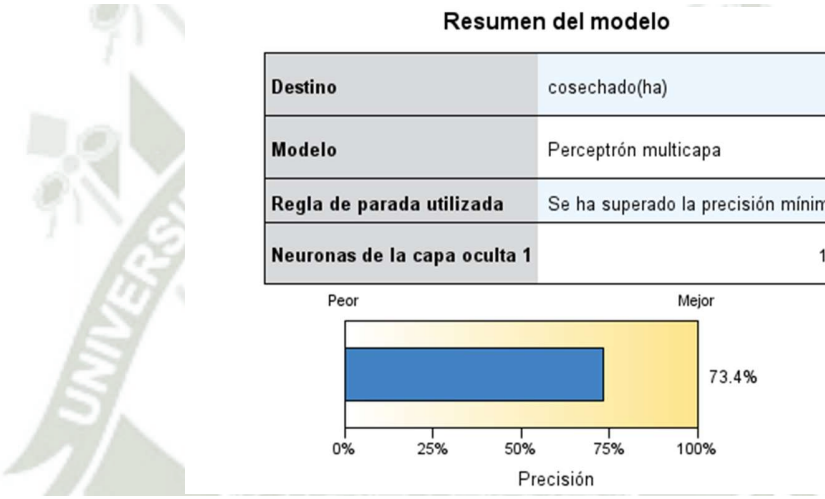


Nota: Esta figura muestra el modelo de red neuronal generado por la herramienta SPSS Modeler para la predicción de las hectáreas cosechadas.

-Calidad de la red neuronal: El modelo perceptrón multicapa alcanza una precisión del 73.4 % con la utilización de 14 neuronas en la capa oculta y un tiempo de ejecución de 15 minutos. Ver Figura 33.

Figura 33

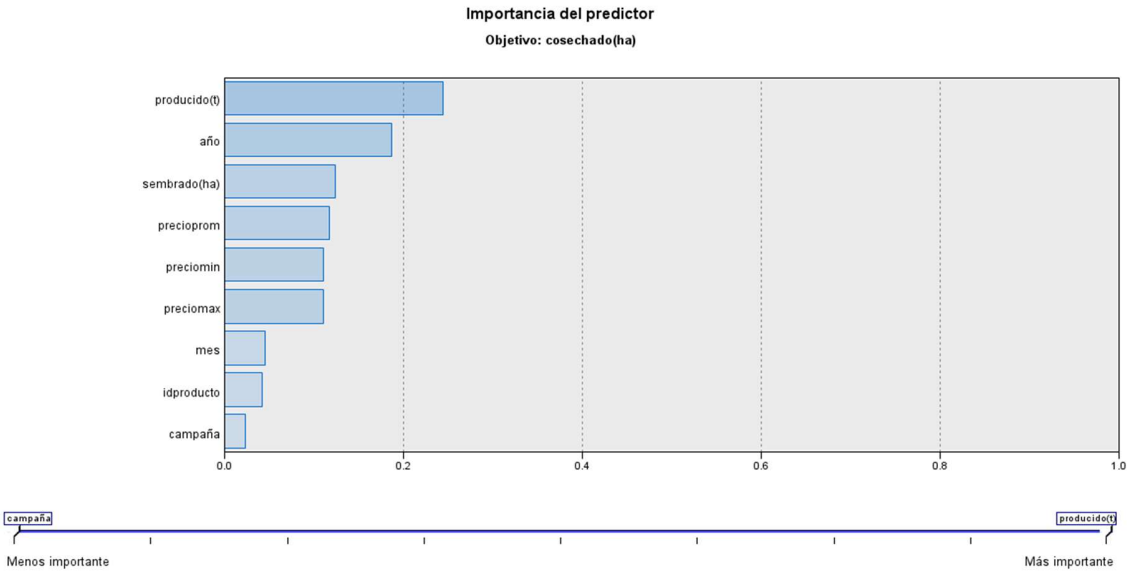
Resumen del modelo generado para el segundo objetivo (hectáreas cosechadas)



Nota: Esta tabla muestra el resumen de la creación de un modelo de red neuronal para la predicción de las hectáreas cosechadas.

Figura 34

Relación entre variables predictoras y variables objetivo – Hectáreas cosechadas



Nota: Esta figura muestra el resumen de la importancia de las variables predictoras y la variable de hectáreas cosechadas.

Como el tipo de valor que tenemos en este caso para el objetivo no es un valor continuo, el error estándar para esta variable seguiría siendo un valor muy grande, por lo que únicamente se muestra el nivel de correlación de la variable objetivo con las demás variables predictoras. Ver figura 34.

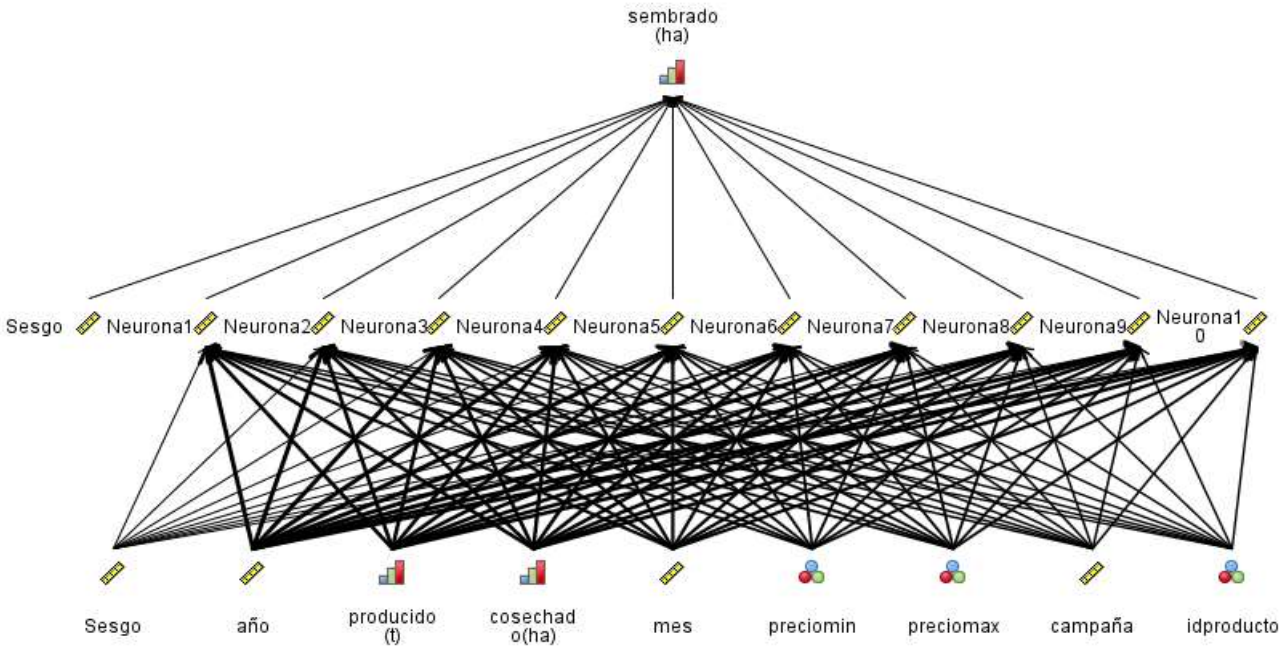
Como se puede apreciar en la Figura 34, existe una mayor influencia de la variable de producción respecto a las demás variables sobre el objetivo (toneladas cosechadas). Esto nos indica una relación positiva entre el rendimiento de las hectáreas cosechadas y la producción generada. De igual forma, la otra variable objetivo-escogida para el cumplimiento del segundo objetivo de minería de datos tiene una importancia considerable en esta variable.

Modelo para el objetivo 2- Hectáreas sembradas:

El modelo de red neuronal que se generó para la predicción de las hectáreas sembradas modelo tiene 10 neuronas en la capa oculta, vera Figura 35.

Figura 35

Modelo de red neuronal generado para la predicción de hectáreas sembradas

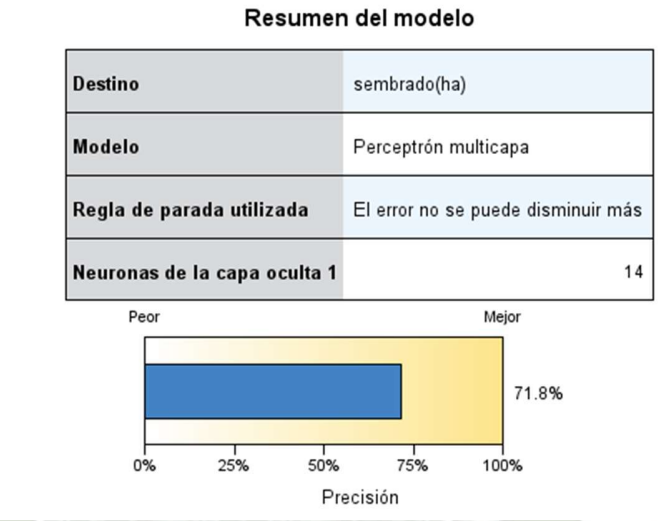


Nota: Esta figura muestra el resumen de la importancia de las variables predictoras y la variable de hectáreas cosechadas.

-Calidad de la red neuronal: El modelo perceptrón multicapa alcanza una precisión del 71.8 % con la utilización de 14 neuronas en la capa oculta y un tiempo de ejecución de 15 minutos. Ver Figura 36.

Figura 36

Resumen del modelo generado para el segundo objetivo (hectáreas sembradas)

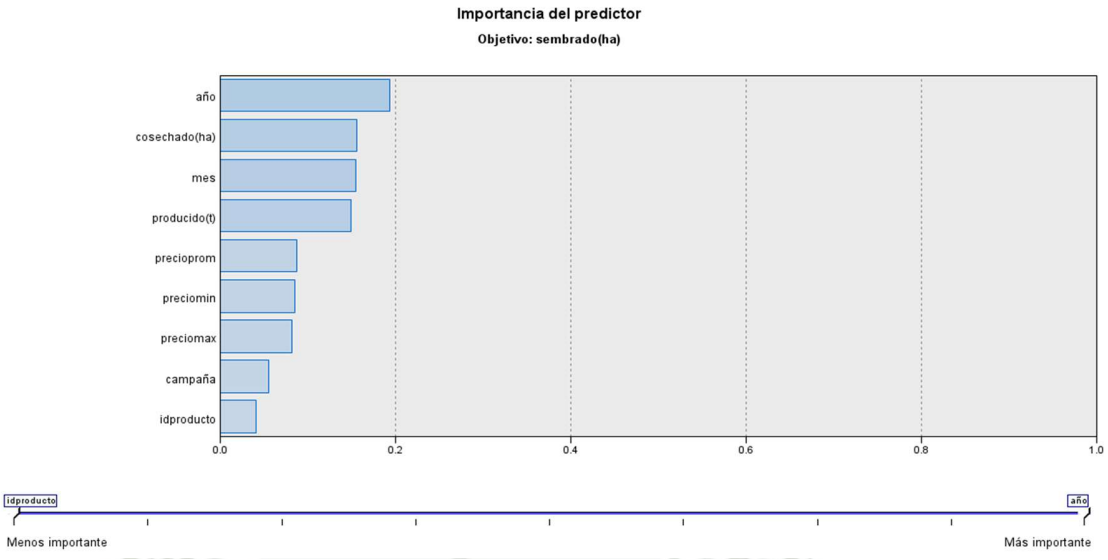


Nota: Esta tabla muestra el resumen de la creación de un modelo de red neuronal para la predicción de las hectáreas sembradas.

Como se puede expresar en la Figura 37, existe una mayor influencia de la variable de año respecto a las demás variables sobre el objetivo (hectáreas sembradas). Esto se debe a que en los últimos años la cantidad de hectáreas sembradas ha disminuido considerablemente. Como otras variables influyentes para esta variable objetivo, tenemos la cantidad de hectáreas producidas y la cantidad de hectáreas cosechadas.

Figura 37

Relación entre variables predictoras y objetivo – Hectáreas sembradas



Nota: Esta figura muestra el resumen de la importancia de las variables predictoras y la variable de hectáreas sembradas.

Modelo para el objetivo 2- Toneladas producidas:

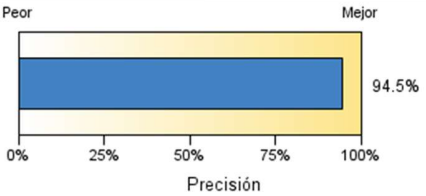
-Calidad de la red neuronal: El modelo perceptrón multicapa alcanza una precisión del 94.5 % y un tiempo de ejecución de 15 minutos. Ver Figura 38.

Figura 38

Resumen del modelo generado para el segundo objetivo (toneladas producidas)

Resumen del modelo

Destino	producido(t)
Modelo	Perceptrón multicapa
Regla de parada utilizada	Se ha superado la precisión mínima
Neuronas de la capa oculta 1	14

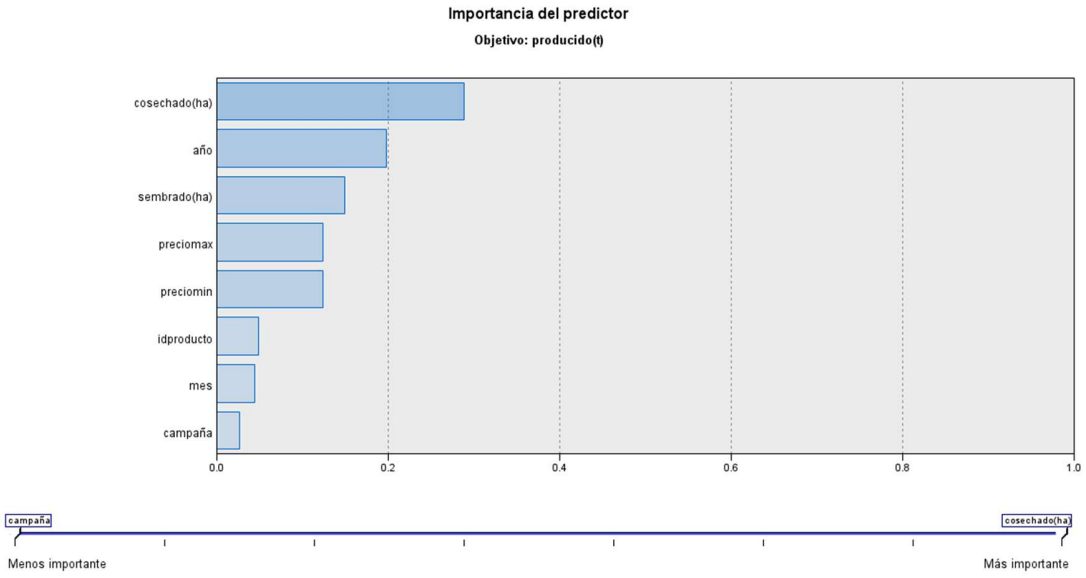


Nota: Esta tabla muestra el resumen de la creación de un modelo de red neuronal para la predicción de las toneladas producidas.

Al igual que el modelo generado para la cantidad de hectáreas cosechadas como la variable objetivo, se ve el mismo grado de correlación entre la variable de cosecha, de siembra y de año. Al igual que el modelo mostrado en la Figura 28, la influencia de las variables de mes, tipo de producto (id) y campaña es mínima, teniendo una importancia menor al 20% entre estas tres. Ver Figura 39.

Figura 39

Relación entre variables predictoras y objetivo – Toneladas producidas

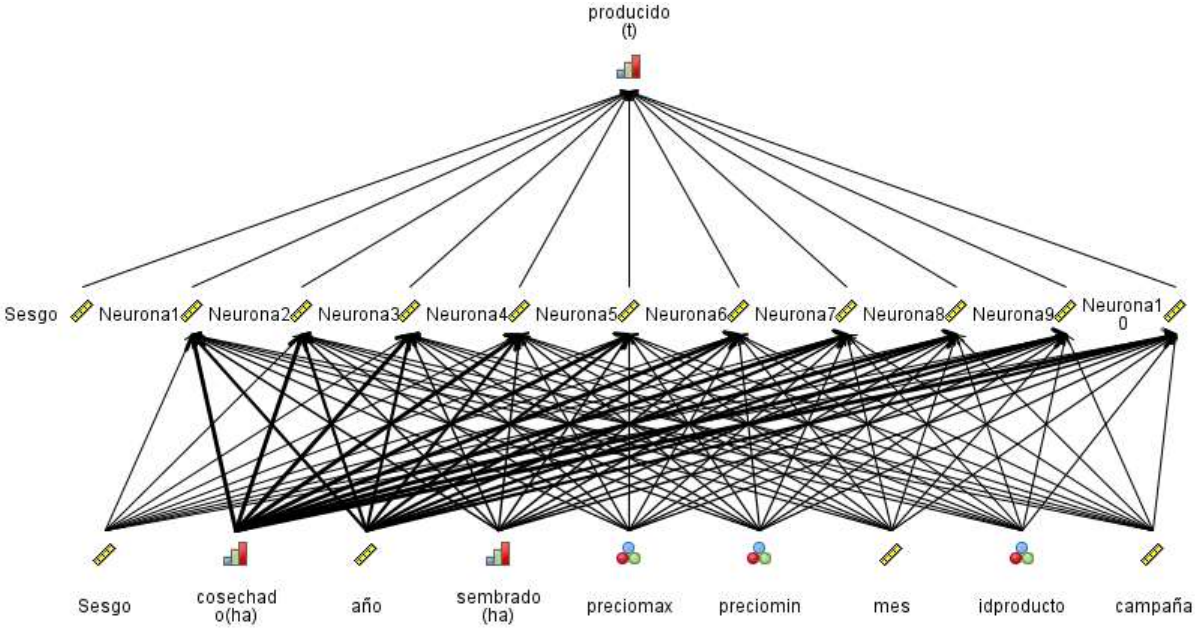


Nota: Esta tabla muestra el resumen de la creación de un modelo de red neuronal para la predicción de las toneladas producidas.

El modelo de red neuronal que se generó para la predicción de toneladas producidas tiene 10 neuronas en la capa oculta, vera Figura 40. Este es el último modelo que junto con los anteriores buscaba corregir el problema de la calidad de los modelos generados inicialmente. Con la información obtenida y el nivel de relación entre las diferentes variables es posible generar nuevos modelos con múltiples objetivos que en la prueba hayan tenido un nivel adecuado de importancia con respecto a otros predictores.

Figura 40

Modelo de red neuronal generado para la predicción de producción de cebolla



Nota: Esta figura muestra el resumen de la importancia de las variables predictoras y la variable de producción de cebolla.

Estudiando los resultados de los modelos generados, se encontró que las variables de entrada de precios tenían mayor relevancia para la predicción de otras variables de precio. Es el mismo caso para las variables de producción: cn el caso de la variable que indica la cantidad de hectáreas sembradas en una campaña específica, las variables predictoras con un mayor nivel de correlación son la variable de cantidad de hectáreas cosechadas, toneladas producidas y el campo de fecha.

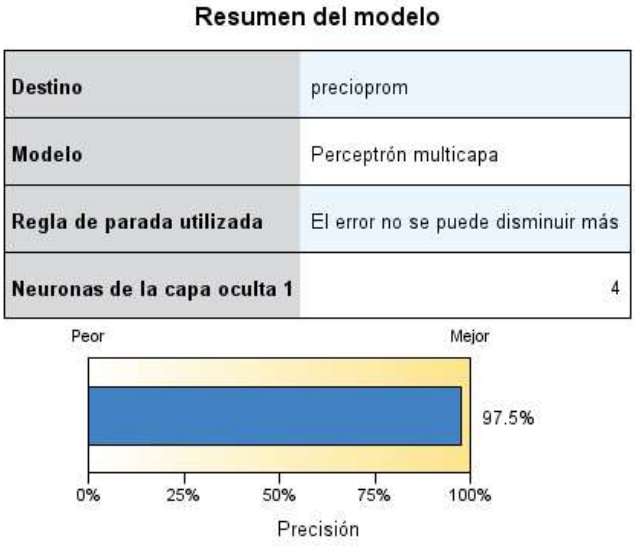
Para este caso se generó un modelo adicional teniendo como variables objetivo las variables definidas como objetivos de minería de datos, pero tomando como variables predictoras aquellas que tuvieron un mejor nivel de correlación en la evaluación de los modelos generados anteriormente. Posteriormente se evaluaron todos los modelos generados con la finalidad de encontrar el mejor modelo generado por objetivos definidos en función del error absoluto promedio, el error porcentual y el nivel de correlación con las variables predictoras.

Para el caso del primer objetivo de minería de datos, para la predicción del precio máximo, se utilizaron únicamente las variables de precio promedio, precio mínimo, el identificador del producto, el año y el mes. Todas estas variables tenían un nivel de correlación superior al de las variables de producción, que, al ser descartadas, se presume podrían mejorar la eficiencia de la red neuronal. Para el caso del cumplimiento del primer objetivo de minería de datos se utilizaron las variables de precios como predictoras para la predicción de los objetivos de precios.

Para la predicción de la variable precio máximo obtuvo una precisión del 95.7%, para la predicción del precio promedio 97.5% y para el precio mínimo un 92. 6%. Para el caso de la predicción del precio promedio y precio mínimo la cantidad de neuronas utilizadas fueron 4 y o respectivamente, ver figura 41 y 42.

Figura 41

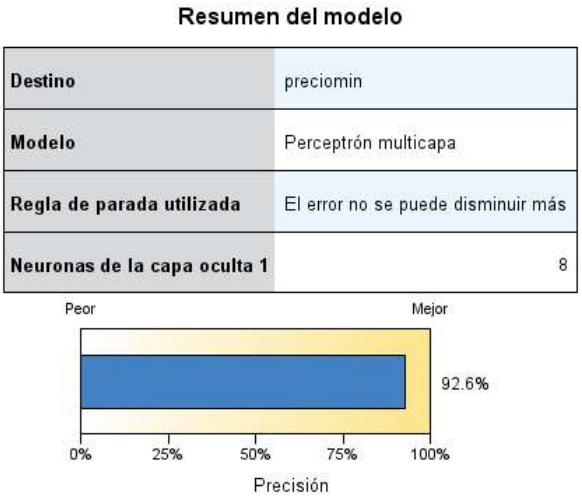
Modelo de red neuronal generado para la predicción de precio promedio con 5 variables de entrada



Nota: Esta figura muestra el resumen del modelo de red neuronal generada para la predicción del precio promedio con 5 variables de entrada

Figura 43

Modelo de red neuronal generado para la predicción de precio mínimo con 5 variables de entrada



Nota: Esta figura muestra el resumen del modelo de red neuronal generada para la predicción del precio mínimo con 5 variables de entrada

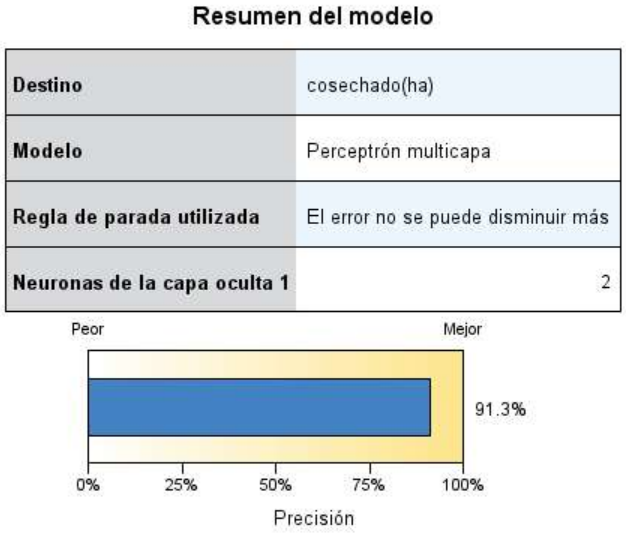
Para la predicción del precio máximo y precio mínimo se observa un nivel de correlación alto con respecto a la variable de precio promedio. En ambos modelos esta variable tiene una correlación del 0.85 y 0.89 para el valor del precio máximo y precio mínimo respectivamente.

Para el caso del segundo objetivo de minería de datos, para la predicción de las variables de cosecha, se utilizaron únicamente las mismas variables de cantidad de toneladas producidas, hectáreas sembradas y hectáreas cosechada, descartando las variables de precio mínimo, precio máximo, precio promedio y el identificador del producto, cuyo nivel de correlación no era muy bueno. El nivel de precisión del segundo modelo para el cumplimiento del segundo objetivo de minería de datos nos devuelve un 92.7% para las toneladas producidas, 71.2% para el objetivo de hectáreas sembradas y 91.3% para la cantidad de hectáreas cosechadas.

Para este modelo se utilizaron 1 nodo en la capa oculta para la predicción de las toneladas producidas y 2 neuronas en la capa oculta para la predicción de las hectáreas cosechadas. Ver figura 44 y 45.

Figura 44

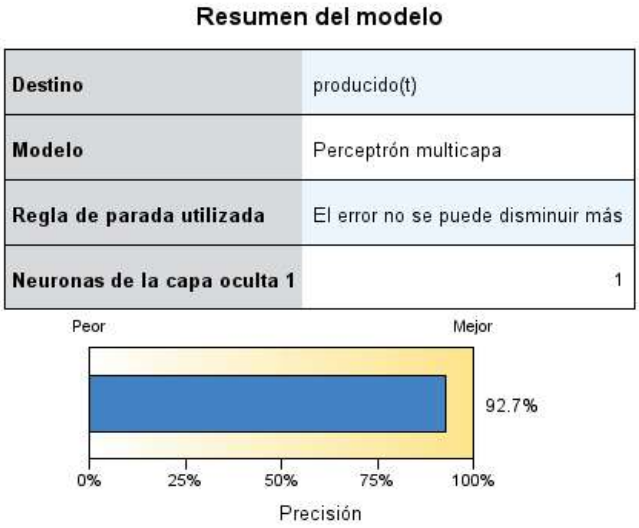
Modelo de red neuronal generado para la predicción de hectáreas cosechadas con 5 variables de entrada



Nota: Esta figura muestra el resumen del modelo de red neuronal generada para la predicción del precio mínimo con 5 variables de entrada

Figura 44

Modelo de red neuronal generado para la predicción de hectáreas cosechadas con 5 variables de entrada



Nota: Esta figura muestra el resumen del modelo de red neuronal generada para la predicción del precio mínimo con 5 variables de entrada

En cualquiera de los dos casos, la cantidad de errores generada debe ser baja en comparación de los modelos generados anteriormente, ya que las variables descartadas no generarían algún tipo de ruido al momento del entrenamiento y comprobación. De esta forma se generaron hasta 3 estructuras para el mejor modelo de red neuronal para la investigación. En la tabla se hace una comparativa del modelo generado con múltiples objetivos de minería de datos y el modelo con variables objetivo-individuales.

Por cada variable utilizada en el modelo, la herramienta de análisis nos ofreció algunos datos estadísticos de su desempeño con el resto de las variables predictoras y su nivel de error con los registros utilizados en dicho modelo de predicción. En la tabla 28 se realiza la comparación de las variables utilizadas en el modelo con 3 nodos de salida (precios y producción) y el modelo que utiliza individualmente como nodo de salida una única variable.

El error absoluto promedio (EAM) representa el valor promedio de la diferencia entre todos los valores reales y los predichos. El error porcentual ($E\%$) es la diferencia de los valores predichos y los exactos representados con un porcentaje. En el caso del primer modelo observamos un error absoluto promedio para los variables de precios, pero su valor es mayor en las variables de producción. Sin embargo, en el segundo modelo observamos que la configuración del modelo cambiando los nodos de salida ayudan a la reducción del nivel de error encontrado en las variables de precios y las variables de producción. La mejora se hace más notoria en las variables de precio máximo, precio mínimo y precio promedio donde la reducción del error es de 20 unidades en promedio.

Para las variables de producción la reducción del error para los campos de cantidad de hectáreas cosechadas y cantidad de hectáreas sembradas el resultado no reduce mucho a comparación de las variables de precios. Para el caso de las toneladas producidas si se observa una reducción considerable del error promedio. Nótese que en la partición de validación del modelo el nivel de error tiende a ser mayor a comparación de las demás particiones, pero para la variable de toneladas producidas, el valor obtenido es menor al de prueba, lo que nos indica una buena predicción para este objetivo. La mejor arquitectura para el modelo de predicción dadas las estadísticas de la tabla es del modelo con un único nodo de salida y 7 nodos de entrada. Para este, la variable con un alto nivel de precisión en la predicción fue la de precio promedio.

Tabla 28

Comparación del error absoluto promedio, error porcentual y el grado de correlación de RN con 3 nodos de salida 5 nodos de entrada y RN con 1 nodo de salida y 7 nodos de entrada

		RN con 3 nodos de salida y 5 nodos de entrada			RN con 1 nodo de salida y 7 nodos de entrada		
	Variable	**Entrenamiento**	**Comprobación**	**Validación**	**Entrenamiento**	**Comprobación**	**Validación**
EAM	precio_prom	22.948	28.953	28.524	4.197	4.328	4.716
	precio_max	27.882	33.336	31.498	7.07	7.764	6.568
	precio_min	27.968	35.722	36.605	6.048	7.261	6.294
	cosechado(ha)	81.259	87.107	101.074	84.269	86.858	94.953
	sembrado(ha)	133.86	139.988	150.317	135.091	188.584	189.413
	producido(t)	4359.694	8459.222	7816.021	1232.695	1642.389	1637.003
E%	precio_prom	0.0289	0.1275	0.2593	0.0053	0.0191	0.0429
	precio_max	0.034	0.12	0.25	0.005	0.019	0.042
	precio_min	0.0352	0.1469	0.2863	0.0089	0.0342	0.0597
	cosechado(ha)	0.0353	0.1574	0.3328	0.0076	0.0320	0.0572
	sembrado(ha)	0.1025	0.3837	0.2321	0.1063	0.3826	0.8632
	producido(t)	0.1688	0.6167	0.3665	0.1704	0.8308	0.7219
r_k	precio_prom	0.771	0.663	0.665	0.989	0.988	0.976
	precio_max	0.728	0.638	0.667	0.976	0.972	0.976
	precio_min	0.768	0.656	0.662	0.971	0.955	0.979
	cosechado(ha)	0.967	0.961	0.953	0.978	0.981	0.967
	sembrado(ha)	0.958	0.96	0.927	0.958	0.957	0.932
	producido(t)	0.974	0.965	0.971	0.998	0.995	0.995

Nota: Esta figura muestra la comparativa del error absoluto promedio, error porcentual y nivel de correlación entre el modelo generado con 3 nodos de salida y 5 nodos de entrada y el modelo con 1 nodo de salida y 7 nodos de entrada.

Con la finalidad de encontrar el mejor modelo, se realizó la misma comparación con el último modelo generado que solo incluía como variables de entrada las variables con el mejor nivel de correlación y se descartaban las variables cuya importancia no era tan alta para la predicción de las variables objetivo. Ver tabla 29.

Tabla 29

Comparación del error absoluto promedio, error porcentual y el grado de correlación de RN con 1 nodos de salida 4-5 nodos de entrada más relevantes y RN con 1 nodo de salida y 7 nodos de entrada

		RN con 1 nodos de salida y nodos más relevantes de entrada			RN con 1 nodo de salida y 7 nodos de entrada		
	Variable	**Entrenamiento**	**Comprobación**	**Validación**	**Entrenamiento**	**Comprobación**	**Validación**
EAM	precio_prom	3.852	4.112	2.979	3.852	4.328	4.716
	precio_max	6.6	7.091	6.415	6.6	7.764	6.568
	precio_min	6.28	6.418	5.504	6.28	7.261	6.294
	cosechado(ha)	30.259	37.944	31.61	30.259	86.858	94.953
	sembrado(ha)	239.352	257.644	196.777	239.352	188.584	189.413
	producido(t)	6540.732	6621.487	7079.651	6540.732	1642.389	1637.003
E%	precio_prom	0.0049	0.0181	0.0271	0.0053	0.0191	0.0429
	precio_max	0.0083	0.0312	0.0583	0.005	0.019	0.042
	precio_min	0.0079	0.0283	0.0500	0.0089	0.0342	0.0597
	cosechado(ha)	0.0382	0.1672	0.2874	0.0076	0.0320	0.0572
	sembrado(ha)	0.3018	0.1350	0.5089	0.1063	0.3826	0.4332
	producido(t)	0.2481	0.4595	0.545	0.1704	0.3308	0.4619
r_k	precio_prom	0.987	0.985	0.996	0.989	0.988	0.976
	precio_max	0.978	0.974	0.975	0.976	0.972	0.976
	precio_min	0.962	0.962	0.98	0.971	0.955	0.979
	cosechado(ha)	0.967	0.961	0.953	0.978	0.981	0.967
	sembrado(ha)	0.958	0.96	0.927	0.958	0.957	0.932
	producido(t)	0.963	0.963	0.962	0.998	0.995	0.995

Nota: Esta figura muestra la comparativa del error absoluto promedio, error porcentual y nivel de correlación entre el modelo generado con 1 nodos de salida y las variables de entrada más relevantes y el modelo con 1 nodo de salida y 7 nodos de entrada.

En la tabla 29 se hace la comparación entre el mejor modelo generado contra el modelo generado con nodos de entrado que incluían únicamente las variables con mejor nivel de correlación e importancia para la predicción del primer modelo. En las 3 particiones de los datos (entrenamiento, prueba y validación) se obtienen datos similares: el último modelo de RN generado tiene un menor nivel de error para las variables de precios y la variable de hectáreas cosechadas.

Sin embargo, para las demás variables de producción, la cantidad de errores en las particiones de aprendizaje y prueba son mayores a las del mejor modelo escogido. Esto se debe al alto nivel de correlación existente entre las variables de precio para el último modelo es muy fuerte y la precisión de predicción del modelo puede acercarse mucho a los valores reales. Asimismo, las variables que no fueron consideradas estarían disminuyendo el nivel de precisión del modelo y aumentando los errores en los registros al no ser valores importantes o con un nivel de correlación superior al 70%.

Es todo lo contrario para las variables de producción, que parecen tomar valores parecidos al primer modelo generado. Esto puede deberse a la medida escogida para estas variables de entrada y que perdieron su nivel de confianza con las demás variables de precios, que, aunque no tener un alto nivel de correlación, disminuía el error en las particiones de registros establecida.

El último modelo generado demuestra una mejor capacidad de predicción para las variables de precios con respecto al segundo modelo, pero un débil desempeño para la predicción de variables de producción con la generación de errores alejados a los valores reales en la mayoría de sus registros. Existe un alto nivel de correlación en la estructura escogida para los dos últimos modelos de la tabla 29, pero el modelo con múltiples variables de entrada de la tabla 28 tiene menor precisión por el rol que tienen las variables de entrada en el modelo. Los dos últimos modelos de la tabla 29 son útiles para la predicción de precios, pero para la obtención de mejores resultados se recomienda incluir más variables de entrada para la predicción de producción de cebolla y aumentar la cantidad de registros para el entrenamiento, prueba y validación.

DESPLIGUE

En esta última fase de la metodología CRISP es el de explicar al cliente o las partes interesadas del proyecto, como utilizar la información generada por los modelos que fueron construidos en las fases anteriores.

De igual forma se pretende explicar los resultados que se obtuvieron de forma que sea comprensible y fácil de entender por cualquier otra persona que no tenga un alto conocimiento en IA o Machine Learning.

Asimismo, en esta fase también se busca crear un plan para el mantenimiento de este proyecto y la generación de un informe final en donde se comenten algunas mejoras y recomendaciones para futuras investigaciones y de las dificultades que se encontraron para la elaboración de este proyecto.

DESLPEGAR EL PLAN

Para la implementación de este proyecto se requiriere tener acceso a una base de datos real del negocio, en este caso, la fuente donde se encontrarán los datos relacionados a las variables de cosecha seleccionadas para esta investigación. Desde ese punto, los pasos a seguir en el proceso de implementación desde la fase de comprensión del negocio hasta la parte del despliegue se seguirán tal y como se ha realizado en esta investigación.

Es necesario recordar que la base de datos generada para esa investigación y los registros con la que fue alimentada fueron obtenidos directamente de los sistemas de información del Minagri y para la ejecución de consultas se deberá actualizar mensual o anualmente con las variables de cosecha escogidas y si es necesario, incluir otras que ayuden a la mejora de la precisión del modelo o que no se hayan considerado.

De esta forma los datos que alimentarán la base de datos tendrán que ser ingresados manualmente y desde distintas fuentes.

En este caso se recomienda la implementación de una base de datos en Oracle o SQL Server principalmente, ya que la mayoría de las herramientas de minería de datos tienen compatibilidad con estos dos gestores de bases de datos. En cualquiera de los dos casos se tendrá que realizar un preprocesamiento de los datos, tal y como se realizó en la investigación. Esto con el fin de exportar los datos a una base de datos de Oracle o SQL Server.

En el caso que se opte por una base de datos de Oracle, el mismo gestor incluye la extensión Oracle Data Miner extensión que permite el análisis de datos y la creación de múltiples modelos de aprendizaje automático, comparación de modelos e implementación.

Solo se plantearon 2 objetivos de minería de datos y la generación de más objetivos dependerá de la cantidad de datos disponibles y las variables objetivo que se quieran predecir y que podría dar un mayor valor al modelo y que pueda contribuir al cumplimiento de los objetivos del negocio.

MONITOREAR Y MANTENER

Este paso es uno de los más importantes en la fase de despliegue del proyecto, ya que los datos de la fuente de datos serán procesados constantemente y estos podrían ser manipulados por algún responsable del negocio. La modificación de los datos no debería ser una tarea que se ejecute con frecuencia a menos que se pueda validar algún error al momento de su codificación en la base de datos.

El proceso de extracción de información debe realizarse en periodos de 5 meses, ya que coincide con el inicio de la temporada de siembra de cebolla roja en la región Arequipa entre los meses de julio y setiembre.

Se recomienda realizar el análisis de los datos antes de estos meses para tener un adecuado plan de siembra y cosecha. Asimismo, es de vital importancia alimentar mensualmente la base de datos con los precios de venta de la cebolla roja en los diferentes mercados mayoristas del país con ayuda de las herramientas del Minagri.

Como parte del plan de monitoreo y mantenimiento se establecen las siguientes actividades:

- Cada 5 meses se realizará la recolección, extracción y almacenamiento de los datos de variables de cosecha de la última campaña de cebolla en la base de datos creada.
- Almacenamiento de los archivos de la explotación de datos generados en algún medio escogido por las partes interesadas y de la misma forma organizarlo por campaña de cosecha o algún otro criterio escogido por los responsables de la información.
- Los resultados obtenidos en cada proceso de explotación de los datos deberán ser llevados a algún formato para la generación de gráficas y/o alguna herramienta de visualización para una mejor interpretación de la información en cada campaña de siembra y cosecha.

- Retroalimentación y adición de nuevas variables predictoras y variables objetivos los modelos de acuerdo con algún posible nuevo requerimiento para el cumplimiento de las necesidades del negocio.

GENERAR INFORME FINAL

Para este penúltimo paso se presentará un informe donde se resumirá las partes más importantes del proyecto y el conocimiento que se adquirió durante el proceso de desarrollo. El público al que está dirigido este informe es a los agricultores dedicados al cultivo de cebolla y a los diferentes gremios cebolleros que venden sus productos en diferentes mercados mayoristas y minoristas del país, con la finalidad de poder estudiar la situación actual y de acuerdo con el correcto uso de la información disponible poder tomar decisiones correctivas para la mejora de la rentabilidad del cultivo de cebolla roja en Arequipa.

Al haber usado la metodología CRISP-DM en este proyecto de investigación, se ha logrado descubrir gran cantidad de sistemas de información que reúnen datos agrícolas, climatológicos y económicos que no son muy conocidos y que su utilización podría significar la mejora o incluso la creación de metodologías de cultivo que existen actualmente.

De esta forma es que se logró encontrar una conducta predictiva en los modelos de redes neuronales generados al momento de estimar el precio y la producción de cebolla roja cultivada en Arequipa. Se ha seguido un detallado plan de extracción y codificación de estos datos para la ejecución de procesos de minería de datos cada 5 meses para el aprovechamiento de la información en cada campaña de siembra y cosecha.

Se definieron inicialmente 3 objetivos de minería de datos, sin embargo, uno de estos fue descartado al momento de la definición por la poca cantidad de datos para las variables predictoras y las variables objetivos que no permitirían el cumplimiento de este objetivo. Los otros dos objetivos se lograron cumplir con un nivel aceptable de precisión y reduciendo lo más que se pudo los errores generados por ambos modelos.

A parte del cumplimiento de estos objetivos, a lo largo de todo el proceso de familiarización de los datos, se ha podido sacar otras conclusiones, específicamente de la relación que existen entre las variables de cosecha y su influencia con las demás variables predictoras y variables objetivos que pudiesen tener un impacto con los precios de venta de la cebolla roja. La primera etapa del proyecto incluyó la parte de la creación de una base de datos desde cero, al no contar una de la que partir.

Previamente se hizo un análisis de todas las variables que se necesitarían para el cumplimiento de los objetivos y que estas también tuvieran relación con la problemática que se desea enfrentar, pues se deseaba tener una simulación que sea lo más real posible, sin tener que generar datos adicionales con la herramienta de minería de datos utilizada y que estos datos que se obtendrían puedan cumplir todos los requerimientos antes mencionados. Afortunadamente, la recolección de información y la selección de las variables de cosecha hicieron que estos datos puedan ser reunidos de fuentes confiables y posteriormente cargados a una base de datos.

El crear una base de datos desde cero y con las variables recomendadas por los mismos expertos del área ha facilitado enormemente la primera etapa de preparación de los datos, ya que el proceso de limpieza, conversión y formateo de los datos fue más sencillo. Esto queda reflejado en el tiempo de duración de la etapa 3 del proyecto.

Posteriormente, se escogieron los parámetros adecuados para el modelo de red neuronal que se utilizaría para la predicción de variables de cosecha. La ejecución de las técnicas de modelado con los datos escogidos fue realizada en la herramienta de minería de datos de IBM (SPSS Modeler). Esta herramienta permitió la aplicación de las técnicas de modelado escogidas y mostrar la calidad de cada uno de los modelos generados para el cumplimiento de los objetivos de minería de datos.

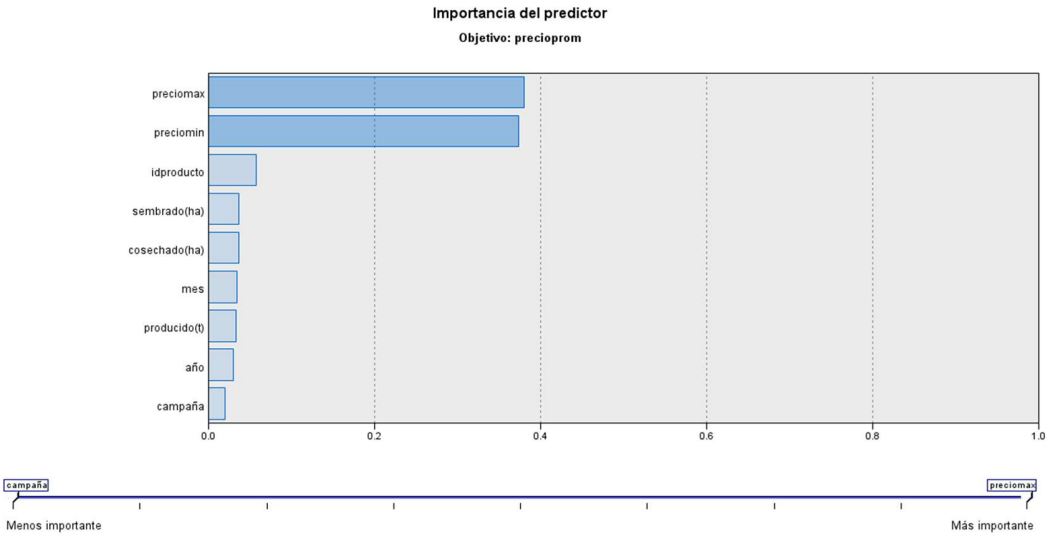
Luego de la obtención de los modelos estos fueron evaluados para determinar el nivel de adecuación a los objetivos planteados inicialmente. Para este caso se evaluarán los modelos para los dos objetivos de minería de datos. Inicialmente se descartaron los modelos generados por tener múltiples objetivos: 3 objetivos por cada modelo. Esto con el fin de reducir los errores generados en el modelo y la posibilidad de que las variables objetivas no tuvieran relación entre ellas mismas. Al realizar este cambio se logró reducir considerablemente los errores de los modelos y se pudo tener un grado de precisión del modelo aceptable. Luego de realizar estos procedimientos, a continuación, se presentan los resultados obtenidos a las partes interesadas y el objetivo de este apartado.

Tomando como referencia los modelos generados con variables individuales, se puede apreciar de una mejor forma el comportamiento de cada variable predictora con respecto a la variable objetivo-escogida. De esta forma, para el primer objetivo propuesto se puede apreciar el mismo nivel de importancia entre las variables objetivo, es decir, para los 3 modelos generados existe un alto nivel de importancia de predicción de la variable de precio promedio, precio mínimo y precio máximo, siendo estas mismas, las variables objetivo o las que deben ser predichas.

Existe un muy bajo nivel de correlación respecto a la cantidad de hectáreas sembradas, las cosechadas y la producción total, pero existe un nivel de correlación alto con respecto al precio mínimo, el precio máximo, el precio promedio y el tipo de producto, ver Figura 41. Por lo que la teoría expuesta por Bermudes (2019) sobre que el precio final de la cebolla depende mucho de la producción generada, no tendría mucho valor ahora.

Figura 41

Relación entre variables predictoras de precio de cebolla



Nota: elaboración propia en base a información obtenida durante la investigación (2021).

La relación existente entre el precio máximo, el precio mínimo y el precio promedio tiene un alto nivel de correlación pues existente una relación entre los precios a lo largo de los años y existe una dependencia entre estos. La sobre producción o falta de producción de cebolla roja no puede considerarse como el principal motivo de los precios cambiantes de la hortaliza. Por otra parte, respecto al cumplimiento del segundo objetivo relacionado a la predicción de la cantidad de

hectáreas sembradas, hectáreas cosechadas y toneladas producidas se muestra un comportamiento similar al caso de las variables objetivo y las predictoras de los modelos del primer objetivo.

Sin embargo, se observa una fuerte correlación con el precio mínimo, algo que quizás se esperaba en los primeros modelos al esperar encontrar una relación entre el precio y otra variable de producción. Pero para este segundo objetivo, se ve una relación más importante entre las variables de tiempo, como lo son el año, el mes y la campaña y las variables de producción.

Es importante recalcar esto pues la relación entre los datos históricos acerca de la producción y los últimos 10 años parecen tener un comportamiento descendente conforme avanza el tiempo.

No sería raro ver que en los próximos años la cantidad de hectáreas sembradas y que la cantidad de hectáreas cosechadas disminuyan aún más. A pesar de esto, se nota que el rendimiento por hectárea de las cosechas de cebolla tiende a elevarse a pesar de la disminución de la superficie sembrada y cosechada. Esto es un buen indicador que habla mucho de la calidad del producto que es cosechado.

CONCLUSIONES

1. Si fue posible crear un modelo de red neuronal para encontrar un comportamiento entre la producción de cebolla y el establecimiento de los precios de venta de esta hortaliza. La información histórica que fue utilizada para el modelo fue reunida del sistema de información del MIDAGRI. Esta fue correctamente procesada para que pueda ser utilizada para el entrenamiento de los modelos de predicción. Gracias a esto, fue posible generar varios modelos de predicción para su posterior análisis y selección del mejor modelo que cumpliera con los objetivos de minería de datos. Los resultados fueron validados con precios reales de chacra de la zona agrícola el Cural.
2. Fue posible predecir los precios de cebolla para campañas con los datos obtenidos producto del entrenamiento y validación. Asimismo, se realizó un estudio de rentabilidad manejando el escenario donde se utilizaba el precio promedio de los últimos 10 años y donde se utiliza un precio promedio estimado en base al conocimiento generado del análisis de la información. Sin embargo, este análisis de la rentabilidad no ha considerado las variables recomendadas de oferta y demanda del producto, por lo que se basa únicamente en la teoría que sostiene que los precios de la cebolla dependen de la sobreproducción o falta de producción de cebolla roja. Este análisis nos indica el precio de la cebolla podría llegar a costar entre S/ 1.00 /Kg y S/ 1.50 /Kg para la campaña del 2022. Asimismo, se espera que la cantidad de hectáreas sembradas en Arequipa aumente en relación con el año 2020, para superar las 9300 hectáreas sembradas y obtener una producción entre 365000 y 377500 toneladas. Sin embargo, el rendimiento de los terrenos podría no aumentar como se ha observar en campañas de cosecha pasadas.
3. No fue posible utilizar variables climatológicas para la predicción de la cantidad de hectáreas que se ven afectadas por fenómenos naturales y que impactan directamente en el déficit de producción de cebolla roja por campaña. La información que sería utilizada para la predicción para el cumplimiento de este objetivo fue extraída de la plataforma del SENAMHI, pero esta estaba incompleta: los datos de las precipitaciones acumuladas, temperatura mínima y temperatura máxima tenían registros hasta cierto año y al ser datos de más de 5 años los que se tendrían que generar para el entrenamiento del modelo, se optó por omitir estas variables que pueden tener un impacto directo en la producción de cebolla. Sin

embargo, en la etapa de comprensión de los datos se realiza un análisis preliminar de los datos climatológicos obtenidos y su relación con datos de cosecha y precios en fechas específicas. En algunos años donde se observa un nivel de precipitación acumulada superior se visualiza una reducción del rendimiento de los terrenos cosechados, lo que impacta también en la calidad del producto y en el establecimiento del precio final. Es posible y altamente recomendada la utilización de variables climatológicas de mejor calidad para la predicción de variables de producción.

4. El estudio de los resultados obtenidos del modelo nos permite descubrir el motivo de los precios bajos de venta de la cebolla roja en los mercados mayoristas del país, que inicialmente se creía se debía a la sobreproducción o baja producción de cebolla en la región. Luego del análisis podemos concluir que el precio final no se define por la producción total de cebolla, hectáreas sembradas o hectáreas cosechadas. Existen otros factores externos que intervienen en el precio final de venta que no se consideraron para las variables de entrada del modelo por la hipótesis que se manejaba. Ante este escenario, se recomendó incluir en este tipo de modelos la oferta y demanda del producto, ya que tienen mayor influencia en el establecimiento de los precios de venta y que mejorarían la precisión del modelo, específicamente en la predicción del precio de venta de cebolla roja en los mercados mayoristas del país.
5. El mejor modelo de predicción obtenido para el cumplimiento del primer objetivo es un modelo de red neuronal perceptrón multicapa (MLP) que tiene un promedio de precisión del 85.2% mientras que, para el segundo objetivo, se obtuvo un promedio de 79.9% de precisión luego de modificar los objetivos múltiples a individuales, lo que redujo el porcentaje de error.

RECOMENDACIONES Y TRABAJOS FUTUROS

Al finalizar la investigación, se proponen las siguientes recomendaciones:

- En vista que solo se incluyeron variables de cosecha en el modelo generado y no se logró encontrar una relación directa entre las variables predictoras y las variables objetivo, se recomienda el uso de variables adicionales para complementar las actuales o para la creación de modelos nuevos, como las variables de oferta y demanda, por ejemplo, que podrían tener afectar el precio final o la producción de cebolla roja. En un futuro se recomendaría incrementar la cantidad de variables predictoras para poder encontrar más variables causales que tengan una correlación directa con las variables objetivo-escogidas.
- Para la aplicación de técnicas de aprendizaje automático aplicado al pronóstico de cosechas se recomienda el uso de las herramientas del Minagri, que contiene información agrícola sobre el abastecimiento, precios, mercados y productos agrícolas a nivel nacional y que se encuentra disponible para del público en general.
- El SENAMHI administra toda la información meteorológica de las diferentes zonas del país y es posible solicitar información de carácter público con fines investigativos desde su plataforma virtual. El uso de esta plataforma sería de gran utilidad para la aplicación de técnicas de aprendizaje automático, pero se recomienda complementar esta información mediante solicitudes especiales a esta entidad, puesto que la información publicada en la plataforma está incompleta o es muy antigua, lo que perjudicaría la calidad del modelo si se decide incluir dentro de las variables de predicción.
- El entrenamiento de modelos con el uso de reconocimiento de imágenes es otro posible escenario en este campo. Para esto se recomienda el uso de Geoportal, una plataforma del Sistema Integrado de Estadística Agraria para la visualización de información georreferenciada. Asimismo, la plataforma posee herramientas satelitales para poder analizar imágenes satelitales en tiempo real de siembras y cosechas, anomalías en las precipitaciones, heladas, humedad en los suelos y sequias, entre otros.

REFERENCIAS

Agraria.pe. (2017). "Producción de cebolla en Arequipa por campaña". Recuperado el 12 de julio de 2020 de: https://agraria.pe/noticias/arequipa-produce-3325-mil-toneladas-de- cebolla-por-campana-14612

Amaya, J.E., y Méndez, E.F. (2013). Respuesta de niveles crecientes de NK en la producción de cebolla (Allium cepa L.) var. "Roja Arequipeña". Scientia Agropecuaria, 4, 15-25

Basogin Olabe, X., Redes Neuronales Artificiales y sus Aplicaciones, 2005, Escuela Superior de Ingeniería de Bilbao. pág. 79.

Bermudes, T.F. (2019) Estructura Productiva-Económica, Comercial Externa y nivel de Competitividad de la Cebolla Fresca de Bulbo: 1990-2015. Tesis de Licenciatura no publicada., Universidad Nacional Agraria La Molina, Lima, Perú.

Bellido Anicama Alfredo, Schwarz Diaz Max, (Abril 2019), Redes neuronales para predecir el comportamiento del conjunto de activos financieros más líquidos del mercado de valores peruano, *Revista Científica de la UCSA*, 6,49-64

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., (2000) Step-by-step data mining guide, SPSS.

D´ Negri, C., De Vito, E., (2006). Introducción al razonamiento aproximado: lógica difusa. Revista Argentina de Medicina Respiratoria, 4, 126-136

De Sena Júnior D.G., Pinto F., De Queiroz F., Mantovani E. (2001). Algoritmo para classificação de plantas de milho atacadas pela lagarta do cartucho. Revista Brasileira de Engenharia Agrícola e Ambiental. 502-509.

Fernando Villada, Nicolás Muñoz y Edwin García-Quintero (Octubre 2016), Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro*, Información Tecnológica, 27, 143-150*

Figueredo Ávila, G., Ballesteros Ricaurte J. (2015). Identificación del estado de madurez de las frutas con redes neuronales artificiales, una revisión. Ciencias y Agricultura, 13, 117-132.

H. Mhaskar and T. Poggio, “Deep vs. Shallow Networks: an Approximation Theory Perspective,” no. 54, pp. 1–16, 2016.

International Business Machines Corporation, (2021), Guía de CRISP-DM de IBM SPSS Modeler de IBM Documentation, Extraído de: https://www.ibm.com/docs/es/spss-modeler

J. Martinez. (2020), ¿Clasificación o Regresión? Recuperado de: https://www.iartificial.net/clasificacion-o-regresion/

J. Hernández, Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones. (2018), Archivos Venezolanos de Farmacología y Terapéutica (37), 5.

London Bullion Market Association,LBMA. London Bullion Market Association. Consultado el 5 de enero de 2015. Extraido de http://www.lbma.org.uk/pages/index.html.

López Miranda W., Mamani Coila E., Oda Ortiz M. P., Rubina Cárdenas P. E., (2018) Planeamiento estratégico de la cebolla en el Perú: Periodo 2013 – 2021 [Trabajo de maestría]. Pontificia Universidad Católica del Perú.

Martínez-Casasnovas J., Bordes Aymerich X. (Setiembre 2015). Viticultura de precisión: Predicción de cosecha a partir de variables del cultivo e índices de vegetación, XI Congreso Nacional de Teledetección, Puerto de la Cruz, Tenerife

Medina, R.L., (2012) Análisis de la rentabilidad de la cebolla roja de Ilabaya. Tesis de titulación, Universidad Nacional Jorge Basadre Grohmann, Tacna, Perú

Mendoza Vallejos, J., Burgos Chinchay L., (2018) Análisis sectorial de la cebolla roja en el Perú. Trabajo de Investigación para Máster en Dirección de Empresas., Universidad de Piura, Piura, Perú.

Ministerio de Agricultura y Riego (2020), “Sistema Integrado de Abastecimiento y Precios”. Recuperado de http://sistemas.minagri.gob.pe/sisap/portal2/mayorista/

Ministerio de Agricultura y Riego, Sistema Integrado de Estadística Agraria. (2020). “Calendario de Siembras y Cosechas” (formato HTML). Recuperado el 12 de julio del 2020 de: http://siea.minagri.gob.pe/calendario3/

Monzón J, E., Pisarello M, I., (2004). Identificación de latidos cardíacos anómalos con redes neuronales difusas. Comunicaciones Científicas y Tecnológicas.

Ochoa Martínez C., Ayala Aponte A (2007). Prediction of mass transfer kinetics during os- motic dehydration of apples using neural ne-tworks. LWT - Food Sci. Technol. 40 (4): 638- 645.

Peláez Toledo A. (2019)., Redes neuronales artificiales, una herramienta útil para el procesamiento de datos de cosecha [Trabajo de diploma]. Universidad Central Marta Abreu de Las Villas.

Rouhiainen L. (2018), Inteligencia Artificial 101 cosas que debes saber hoy de nuestro futuro. Barcelona, España: Editorial Planeta

Rojas Naccha, Julio; Vásquez Villalobos, Víctor (Agosto 2012), Predicción mediante Redes Neuronales Artificiales (RNA) de la difusividad, masa, humedad, volumen y sólidos en yacón (Smallantus sonchifolius) deshidratado osmóticamente, *Scientia Agropecuaria*, 3,201-214

R. Lasse, "Inteligencia Artificial, 101 cosas que debes saber hoy sobre nuestro futuro". (2018), Editorial Planeta S.A, pp. 16–17, 2018.

Sanches Rodriguez E. (2016). La importancia de la I+D+i en el sector agrícola: una apuesta segura. Recuperado de: http://www.qcom.es/alimentacion/diciembre-2016/la-importancia-de-la-idi-en-el-sector-agricola-una-apuesta segura_31030_2879_34660_0_1_in.html#:~:text=Para%20ello%2C%20la%20investigaci%C3%B3n%20en,mundial%20de%20alimentos%20desde%201960.com

Serna M.E, (2017), Desarrollo e Innovación en Ingeniería. Medellín, Antioquía: Editorial Instituto Antioqueño de Investigación

Toxqui Toxqui R., (2003)., Redes neuronales difusas dinámicas para identificación y control adaptable [Trabajo de diploma]. Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional.

Villazón Bustíos D., Rubio Arias H., Ochoa Rivero J.M., y De la Mora C. (2017). Pronóstico Productivo de la avena forrajera de temporal por efecto del cambio climático en el noroeste de Chihuahua, México. Nova Scientia, 19, 551-567

W. Rivas, B. Mazón, Redes neuronales artificiales aplicadas al reconocimiento de patrones. (2018), Editorial UTMACH S.A, pp. 12–14, 2018.

Oxford Internet Institute. (2020). Guía básica de la IA. Recuperado de: https://atozofai.withgoogle.com/intl/es/predictions/

W. Rivas, B. Mazón, Redes neuronales artificiales aplicadas al reconocimiento de patrones. (2018), Editorial UTMACH S.A, pp. 12–14, 2018.

ANEXOS

ANEXO A: GLOSARIO DE TÉRMINOS

- **Quickbird 2:** El Quickbird 2 fue un satélite de los Estados Unidos que capturaba imágenes de la Tierra y fue colocado en órbita en 2001 y culminó sus operaciones en el 2015.
- **WEKA:** Es un software que contiene una colección de algoritmos para ejecutar tareas de preparación, clasificación, regresión, clustering, reglas de asociación y visualización de datos.
- **Trips (plaga):** Hace referencia a un grupo de insectos del orden *Thysantopera* que son de un tamaño pequeño y de forma alargada y plana. Son portadores de virus que dañan gravemente las cosechas.
- **Botrytis:** Enfermedad que ataca a diferentes especies agrícolas y se propaga con gran facilidad. El Botrytis se trata de un hongo que atacan las flores, frutos, hojas y bulbos de una especie de cultivos, generalmente hortícolas.
- **Raiz Rosada**: Es una de las enfermedades más comunes de la cebolla que es cultivada en climas cálidos. Las plantas contagiadas crecen con más lentitud, sus hojas mueren y la producción de sus bulbos son más pequeños.
- **Aprendizaje automático:** Campo de la inteligencia artificial y su objetivo principal es el de desarrollar técnicas mediante la observación de datos y construcción de modelos en base a los mismos.
- **Aprendizaje supervisado:** Es un conjunto de técnicas para el análisis de datos que utiliza algoritmos que aprenden de los datos de entrenamiento para que los computadores puedan encontrar información oculta de forma iterativa.
- **Inteligencia Artificial:** Capacidad para que los computadores puedan realizar actividades que normalmente necesitan del razonamiento humana. Esta capacidad es posible gracias al uso de algoritmos que le permiten aprender de los datos tal y como lo haría un ser humano.
- **CRISP-DM:** Siglas de “Cross-Industry Standard Process for Data Mining” que describe 6 fases de un modelo de proceso y describe perfectamente el ciclo de vida de minería de datos.
- **SISAP:** Es una aplicación web del Ministerio de Agricultura y Riego que puede ser accedida por cualquier persona natural para la visualización en tiempo real de información referida a volúmenes, precios y productos agropecuarios en diferentes mercados nacionales.
- **SENAMHI:** Organismo público que tiene como misión la generación y publicación de información meteorológica, hidrológica y climatológica para la población peruana.
- **SPSS IBM MODELER:** Es un software de minería de datos de IBM con herramientas para el análisis de datos, construcción de modelos predictivos y otras tareas analíticas.

- **IBM:** Reconocida empresa multinacional de tecnología dedicada a la comercialización de hardware y software además de otros servicios relacionados al área de la tecnología.
- **Error estándar:** Término estadístico utilizado para referirse a un valor estimado de la desviación estándar de una muestra utilizada para el cálculo de las estadísticas.
- **Desviación Estándar:** La desviación estándar es un valor mayor o igual que cero y que indica la dispersión media de una variable.
- **Varianza**: La varianza es una medida de dispersión, y en estadística, la varianza se define como el valor medio de los cuadrados de las desviaciones de la media, la cual siempre es cero.
- **Error porcentual:** Es la representación de un error relativo en términos porcentuales. Se calcula dividiendo el error absoluto entre el resultado esperado para luego multiplicarlo por 100.