PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

APLICACIÓN DE TECNICAS DE MINERIA DE DATOS PARA PRONÓSTICOS DEL SECTOR AGRÍCOLA

TANIA FRANCISCA ZAMORA VILLALOBOS

INFORME FINAL DEL PROYECTO
PARA OPTAR AL TÍTULO PROFESIONAL DE
INGENIERO CIVIL EN INFORMÁTICA

Julio, 2018

Pontificia Universidad Católica de Valparaíso
Facultad de Ingeniería
Escuela de Ingeniería Informática

APLICACIÓN DE TECNICAS DE MINERIA DE DATOS PARA PRONÓSTICOS DEL SECTOR AGRÍCOLA

TANIA FRANCISCA ZAMORA VILLALOBOS

Profesor Guía: **Pamela Hermosilla Monckton**

Profesor Co-referente: **Aldo Migliaro Osorio**

Carrera: **Ingeniería Civil en Informática**

Julio, 2018

Índice

Resumen

En la actualidad empresas e instituciones requieren responder preguntas que van más allá de los datos históricos, es decir, necesitan extraer información que pueda ser útil para el futuro.

El objetivo de esta investigación es encontrar un modelo para el pronóstico de la producción y superficie sembrada de cultivos agrícolas, tales como la papa en la región de La Araucanía y Los Ríos con Los Lagos (éstas últimas en conjunto), y el trigo en la región del Biobío y La Araucanía. El análisis estará enfocado en los datos históricos obtenidos por el INE.

Este estudio se basará en la aplicación de técnicas de minería de datos, orientadas en la identificación de patrones. Una de ellas son los modelos econométricos, que generan patrones a partir de los datos recibidos, con el fin de minimizar los errores asociados. Una de las aplicaciones de estos modelos es el pronóstico de series de tiempo, donde no se busca explicar la variable de interés, sino sólo pronosticarla según su pasado comportamiento.

Finalmente, los resultados obtenidos mostrarán la eficacia del modelo y análisis de los datos, quedando estos a disposición del público.

Palabras Claves: Pronóstico, Minería de Datos, KDD, CRISP-DM.

Lista de Abreviaturas o Siglas

KDD	: Knowledge Discovery in Databases.
CRIPS-DM	: CRoss-Industry Standard Process for Data Mining.
MAE	: Mean Absolute Error.
MAPE	: Mean Absolute Percentage Error.
SEMMA	: Sample, Explore, Modify, Model, Assess.
SVM	: Support Vector Machines.
SVR	: Support Vector Regression.
RNA	: Redes Neuronales Artificiales.
WEKA	: Waikato Enviroment for Knowledge Analysis.

Lista de Figuras

Lista de Tablas

1 Introducción

La minería de datos es una nueva tecnología que está cobrando relevancia en la actualidad, su utilidad para resolver complejos problemas a los que se enfrentan las empresas e instituciones ha dado entrada a la aplicación e investigación sobre la misma. Sin embargo, esta tecnología no es una heurística cualquiera, se fundamenta en la rama de las ciencias de la computación denominada inteligencia artificial y las matemáticas mediante la estadística.

En un comienzo, las empresas sólo se preocupaban por el almacenamiento de los datos, datos históricos que permitían cálculos matemáticos simples con una finalidad, la generación de reportes. De esta manera, se buscaba responder las preguntas referentes al control del negocio. Posteriormente se profundizaron estas preguntas de control hasta llegar a la creación de un repositorio consolidado. En la actualidad las empresas e instituciones requieren responder preguntas que van más allá de los datos históricos, es decir, necesitan extraer información que pueda ser útil para el futuro. En este nuevo desafío aparece la minería de datos, la cual va inserta en un procedimiento Knowledge Discovery on Databases (KDD), puesto que para obtener información del futuro se debe estar seguro del presente.

Las técnicas de minería de datos están orientadas a resolver preguntas de negocio relacionadas con asociaciones, clasificaciones, conglomerados, pronósticos y análisis de texto. Este proyecto estará enfocado en las técnicas de minería de datos relacionadas solo con el pronóstico, los cuales sirven para estimar valores futuros de entidades. Por ejemplo, pronosticando la demanda futura de productos, un fabricante puede planear su producción.

El rubro a analizar es el de la agricultura, donde se realizarán pronósticos para la producción y superficie sembrada de papa en la región de La Araucanía y Los Ríos con Los Lagos (estas últimas en conjunto), y el trigo en la región del Biobío y La Araucanía; es en esas regiones donde se concentra la producción de tales cultivos. Estos productos forman parte de los cuatro cultivos de mayor importancia en el mundo, junto con el arroz y maíz. Los datos dispuestos para ellos datan del año agrícola 1979/80, y se cuenta con la superficie sembrada para tales productos y la producción que se obtuvo. Además con la intención de mejorar los cálculos, se cuenta con estadísticas de agua caída en las regiones antes mencionadas.

La estructura del presente documento se encuentra distribuido de la siguiente forma: en el punto 2 se plantea la descripción del problema, partiendo con una breve introducción del negocio y rubro. En el punto 3 se presenta el estado del arte de minería de datos, con la descripción de las metodologías y técnicas utilizadas para resolver un problema de pronósticos. En el punto 4 se expone el diseño de la solución detallando cada etapa del trabajo realizado. Finalizando se presentan las conclusiones y el trabajo futuro propuesto.

2 Descripción del Problema

2.1 El Negocio

La agricultura en Chile es una actividad que tiene antecedentes prehispánicos en gran parte del país. Entre los cultivos prehispánicos se cuentan el maíz, la papa, el poroto y el zapallo. Con la colonización española se introdujeron, entre otros, el manzano, el olivo, el trigo y la vid.

La papa y el trigo forman parte de los cuatro cultivos de mayor importancia en el mundo, junto con el arroz y el maíz.

Anualmente se cultivan alrededor de 50 mil hectáreas con papas, siendo el cuarto cultivo en superficie y el que tiene mayor número de agricultores: 59.606 según el VII censo agropecuario, la mayor parte de ellos pequeños. La producción es destinada casi totalmente al mercado interno y es un alimento importante en la dieta de los chilenos. Chile tiene variedades nativas de papa, lo que constituye un patrimonio genético para el país.

La superficie cultivada con cereales es de aproximadamente 514 mil hectáreas (temporada 2017/18), de las cuales un 46% corresponde a trigo, un 21% a avena y un 17% a maiz. Los demás cereales tienen participaciones inferiores a 6%. El 75% de las siembras de trigo y el 86% de las de avena se concentran en las regiones del Bío Bío y La Araucanía. El maíz se siembra principalmente en las regiones de O'Higgins y el Maule, que reúnen el 70% de la superficie.

2.2 Planteamiento del Problema

La siembra de cultivos anuales ha disminuido en Chile en las últimas décadas. Es así como las 680 mil hectáreas sembradas en la temporada 2010/11 representan una baja de 17,8% en relación a la siembra de diez años antes. Esta reducción no ha sido homogénea y ha tenido lugar principalmente en trigo, leguminosas, remolacha, raps y papas. Las menores siembras han sido parcialmente compensadas con aumentos en el área de maíz, avena, cebada y lupino.

Una evolución opuesta han tenido los rendimientos por hectárea, que han aumentado en prácticamente todos los cultivos, alcanzando en algunos de ellos niveles importantes en el concierto mundial.

Es por esta razón que el INE está realizando un proyecto de mediano plazo destinado a generar estimaciones de superficie sembrada y cosecha de cultivos anuales esenciales en el territorio comprendido entre las regiones de Coquimbo a Los Lagos, para el período intercensal 2007-2017.

2.3 Objetivos

2.3.1 Objetivo General

Desarrollar modelos de proyección sobre determinados cultivos agrícolas, entregando evidencias sobre su utilidad, para brindar información de vital importancia para los agricultores, ya que puede contribuir al desarrollo exitoso de esta actividad económica.

2.3.2 Objetivos Específicos

- Desarrollar el estado del arte en técnicas de Minería de Datos para pronósticos en series de tiempo y de las metodologías existentes para este propósito.

- Determinar las técnicas de Minería de Datos que permiten obtener la mejor proyección en la producción de determinados productos, evaluando su desempeño predictivo.

- Evaluar los modelos utilizados y comparar sus resultados para establecer qué tipo de modelo es capaz de generar un mejor pronóstico en la producción de determinados productos.

- Analizar los resultados para entender los factores que alteran el comportamiento en el tiempo.

2.4 Alcance del Estudio

Como se señala en los objetivos, este proyecto busca entregar evidencias respecto a la utilidad de las técnicas de minería de datos como herramienta para el pronóstico en el área agrícola, así como también las fortalezas que la identificación de patrones de comportamiento trae consigo.

Para tal efecto, utilizando los datos históricos de superficie sembrada y producción obtenida en papas y trigo, así como los datos de agua caída en las regiones a analizar, se construirán redes neuronales y modelos de series de tiempo con los que se generaran los pronósticos, para luego evaluar sus correspondientes errores de proyección.

Este proyecto estará enfocado en los pronósticos de producción y superficie sembrada de papa en la región de La Araucanía y Los Ríos con Los Lagos (estas últimas en conjunto), y el trigo en la región del Biobío y La Araucanía, ya que estos cultivos agrícolas forman parte de los más importantes dentro del país y en tales regiones es donde se concentra su producción.

3 Estado del Arte

En los últimos años, ha existido un gran crecimiento en nuestras capacidades de generar y colectar datos, debido básicamente al gran poder de procesamiento de las máquinas como a su bajo costo de almacenamiento. Sin embargo, dentro de estas enormes masas de datos existe una gran cantidad de información "oculta", de gran importancia estratégica, a la que no se puede acceder por las técnicas clásicas de recuperación de la información. El descubrimiento de esta información "oculta" es posible gracias a la Minería de Datos (Data Mining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés) que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados [1].

3.1 Proceso de Descubrimiento del Conocimiento

El término descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD) se utiliza frecuentemente como sinónimo de Minería de Datos, aunque existen claras diferencias entre ambos. KDD es un proceso que consta de una serie de fases, mientras que la Minería de Datos es sólo una de estas fases.

A descubrir conocimiento en bases de datos o KDD se le define como “el proceso no trivial de identificar patrones novedosos, válidos, potencialmente útiles y descifrables en el conjunto de datos” [2].

Un proceso clásico de KDD consta de 5 fases fundamentales: selección, procesamiento, transformación, Minería de Datos, evaluación e interpretación. El esquema de las etapas puede observarse en la figura 3.1.

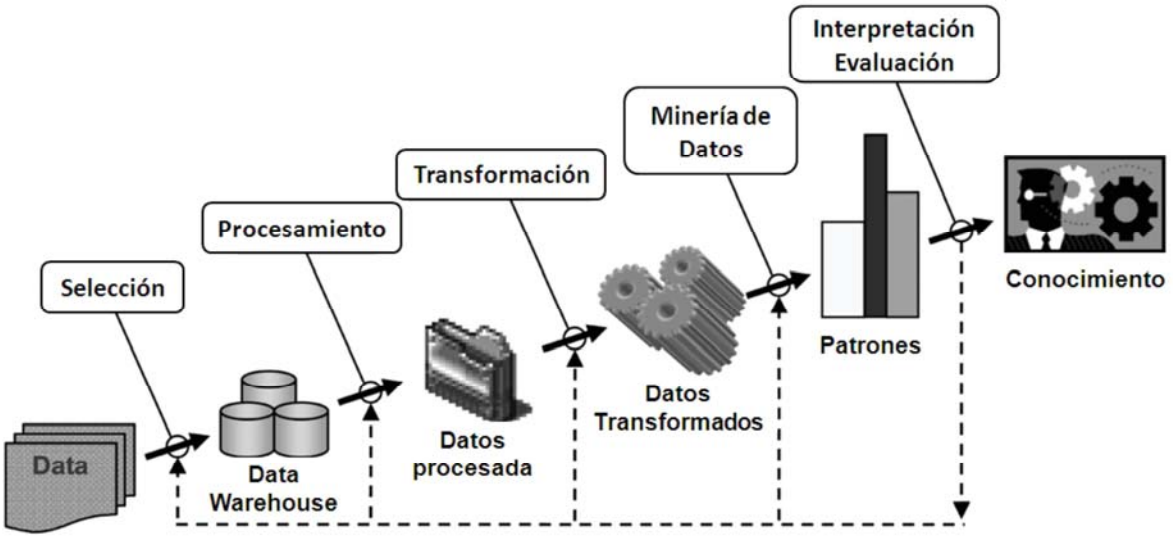

Figura 3.1 Etapas KDD [23]

3.1.1 Selección

Esta etapa tiene como función la selección de aquellos datos o muestra de datos que sean útiles y necesarios sobre los cuales se hará el descubrimiento. Esto puede abarcar tablas, atributos y datos, y con una adecuada selección, se mejora el desempeño de los algoritmos a utilizar, además se producen más rápido los resultados y facilita el entendimiento del descubrimiento del conocimiento obtenido [4].

3.1.2 Procesamiento

En esta etapa, se realiza el procesamiento y limpieza de los datos, que consiste en eliminar datos incoherentes, registros duplicados y basura en los datos producidos por las migraciones que se desarrollan en ellos, y si es necesario y apropiado se eliminan datos que no sean útiles en el proceso. Es común encontrar datos que contienen errores pero en este caso es preferible mantenerlos debido a que la data no es muy extensa y eliminarlos podría causar más un daño que un beneficio.

El procesamiento de los datos, es la fase donde se ocupa la mayor parte del tiempo para analizar y preparar los datos y por tal motivo esto es muy importante ya que de la preparación de los datos dependen en gran medida los resultados.

La información que se va formando del proceso de limpieza y el procesamiento de los datos van generando información o datos muy valiosos que pasando por cada proceso los resultados esperados llevaran a tomar excelentes decisiones [4].

3.1.3 Transformación

La transformación de los datos se basa en que es necesario que los datos sean enriquecidos con otras fuentes de información ya sean internas o externas para reducir los datos y el número de variables y transformarlos en un formato que ayude el mejoramiento de la información para llegar al conocimiento requerido.

La transformación de los datos es muy importante ya que es aquí donde se toma el formato requerido para vincularlo a la minería de datos y extraer el conocimiento que se debe tomar al filtrar los datos que se van obteniendo a través del proceso. En este punto se incluyen operaciones básicas sobre los datos, en el cual consiste el filtrado de los datos para reducirlos y decidir qué hacer con los datos que puedan resultar faltantes [4].

3.1.4 Minería de Datos

Este paso en el proceso de KDD, consiste en la aplicación de análisis de datos para descubrir un algoritmo ad-hoc que “bajo limitaciones computacionales aceptables, produzca una particular enumeración de patrones” [2]. En esta etapa se selecciona el modelo a ocupar, bajo los supuestos que mantienen los objetivos primarios del estudio. Además, es en esta etapa en donde los algoritmos “aprenden” a partir de los datos, por lo que se ejecuta múltiples veces el “entrenamiento” del modelo [3].

3.1.5 Interpretación y Evaluación

En esta fase se evalúan los patrones y se analizan por expertos, y si es necesario, se vuelve a las fases anteriores para una nueva iteración. Para realizar la evaluación se entrena el modelo con una parte de los datos y luego se valida con los restantes. Dependiendo de la tarea de Minería de Datos existen diferentes medidas de evaluación de los modelos. Para la interpretación por parte de los usuarios del conocimiento que aportan los modelos aprendidos, se pueden aplicar técnicas como la visualización de modelos, o visualización posterior.

3.2 Principales Metodologías del KDD

El desarrollo de un proceso de KDD no es trivial y la existencia de una guía para llevarlo a cabo permite organizar los recursos materiales y humanos de forma eficiente.

Entre las metodologías más utilizadas internacionalmente se encuentran CRISP-DM y SEMMA, aunque en ocasiones los desarrolladores emplean metodologías propias [5].

3.2.1 CRISP-DM

La metodología CRISP-DM (CRoss-Industry Standard Process for Data Mining) es una de las más difundidas y utilizadas. Está descrita como "un modelo de procesos jerárquicos, consistente en un conjunto de tareas descritas en 4 niveles de abstracción" [6], desde el general hasta el específico: fase, tareas generales, tareas específicas e instancias de proceso.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de KDD en seis fases (ver figura 3.2), que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto.

A continuación detalla cada una de las etapas de la metodología CRISP-DM [6].

a) En primer lugar se encuentra la etapa de Análisis del Problema, en donde se contextualiza la problemática y la interpretación que se quiere dar, trabajo a realizar en conjunto con el experto del negocio. En esta etapa se plantea una planificación de trabajo, lineamientos y objetivos a lograr para la investigación.

b) Conjuntamente a la etapa anterior, se puede desarrollar el Análisis de los datos que serán recolectados y, posteriormente utilizados para la obtención de conocimiento relevante. Esta etapa se debe complementar con el análisis práctico de cada variable involucrada, modelo de datos, documentación asociada y la interpretación del experto del negocio. Tal como se mencionó, puede existir una constante retroalimentación con el entendimiento del negocio para una mejor comprensión general.

c) En tercer lugar se procede a la Preparación de los datos, en la cual bajo ciertos criterios fundamentados se realiza la selección, limpieza y formateo de los datos. Esta etapa es bastante importante para obtener resultados significativos en

función de los datos seleccionados y del filtrado de las variables más representativas al problema. Puede producirse la transformación de ciertos atributos, ya sea a través de categorización de variables cualitativas o la discretización de variables numéricas.

d) La etapa de preparación de los datos se encuentra muy relacionada con la siguiente etapa de Modelado, debido a que depende de las técnicas de modelado que serán utilizadas sobre los datos. Por lo tanto la preparación y modelado interactúan de forma sistemática.

e) En la etapa de Evaluación se evalúa la aplicación del modelo de análisis adoptado, no desde el punto de vista de los datos sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso seguido, teniendo en cuenta los resultados obtenidos para poder volver a repetir alguna etapa o paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido competer algunos errores o imprecisiones indeseadas.

f) Si el modelo definido en la etapa anterior es válido y competente en función de los criterios de éxito establecidos desde las etapas iniciales en el entendimiento del problema, se procede a la Explotación; normalmente los proyectos de minería de datos no terminan en la implantación directa del modelo sino que se debe documentar y presentar los resultados de manera comprensible al experto del negocio. Una fase de explotación debe asegurar el mantenimiento de la aplicación del modelo de análisis y la difusión de los resultados, lo que en ciertas oportunidades puede implicar comenzar nuevamente el ciclo con el nuevo conocimiento aprendido en el ciclo anterior.

Las flechas indicadas en la figura 3.2, indican las relaciones más habituales entre las fases, mientras que el círculo exterior simboliza la naturaleza cíclica de la minería de datos, ya que la solución final obtenida puede conducir al planteamiento de nuevas hipótesis que den origen a otros proyectos.

Figura 3.2 Fases CRISP-DM [23]

3.2.2 SEMMA

El proceso SEMMA (Sample, Explore, Modify, Model, Assess) se basa en una serie de etapas para conducir proyectos de minería de datos. El SAS Institute lo ha desarrollado, considerando un ciclo compuesto de cinco etapas [7]:

a) Sample consiste en obtener una muestra de datos a través de la extracción de una porción suficientemente grande como para contener la información significativa, pero no tan elevada para que sea fácil de manipular.

b) Explore, es la etapa para explorar los datos buscando, de forma anticipada, tendencias y anomalías para lograr el entendimiento e ideas sobre los datos.

c) Modify se centra en la transformación o modificación de los datos a través de la creación, selección y tratamiento de variables orientados a la selección de un modelo.

d) Model es la etapa para el modelamiento de los datos a través de la aplicación de distintos algoritmos, buscando combinaciones de datos que sean útiles y confiables para predecir resultados esperados.

e) Assess consiste en la evaluación de los datos mediante la valoración de la utilidad y confiabilidad del conocimiento descubierto a partir del proceso de minería de datos.

3.3 Minería de Datos Para Predicción

El objetivo de la minería de datos es producir nuevos conocimientos con los que el usuario pueda operar. Para ello es necesaria la construcción de un modelo basado en datos recogidos a partir de diversas fuentes: operaciones corporativas, historial de clientes, información demográfica, datos de control de procesos y bases de datos externas (por ejemplo, información de la oficina de crédito o datos meteorológicos). El resultado del modelo es una descripción de los patrones y relaciones existentes en los datos, que pueden ser utilizados con cierto grado de confianza en el proceso de predicción [8,9].

3.3.1 Series de Tiempo

Este tipo de modelos predice valores futuros que son desconocidos, basados en la historia acumulada de una o más variables. A diferencia de las regresiones, en una serie de tiempo el orden de las observaciones es importante porque se trata de valores que ocurren en un orden determinado. Los modelos más tradicionales para analizar series de tiempo son los promedios móviles, modelos ARIMA, X11 y Suavización Exponencial. El modelo obtenido en este caso también es utilizado para predecir el valor de la variable objetivo en los meses, días, años o períodos para los cuales sea necesario [10].

3.3.1.1 Promedios Móviles

La utilización de esta técnica supone que la serie de tiempo es estable, es decir, los datos que la componen se generan sin variaciones importantes entre un dato y otro (error aleatorio=0), lo cual significa que el comportamiento de los datos, aunque muestre un crecimiento o un decrecimiento, lo hará con una tendencia constante.

Al usar el Método de Promedios Móviles [11] se supone que todas las observaciones de la serie de tiempo son igualmente importantes para la estimación del parámetro a pronosticar. De esta manera, se utiliza como pronóstico para el siguiente periodo el promedio de los n valores pertenecientes a los datos más recientes de la serie de tiempo. Utilizando una expresión matemática, se obtiene:

$$Promedio\ Móvil = \frac{\sum(n\ valores\ más\ recientes\ de\ la\ serie\ de\ tiempo)}{n} \tag{3.3.1}$$

En la Ecuación (3.3.1), el término n indica que, conforme se tiene una nueva observación de la serie de tiempo, se reemplaza la más antigua de la ecuación y se calcula un nuevo promedio. El resultado es un desplazamiento del promedio (un periodo en el futuro), y en la medida que se obtienen nuevos datos, se sustituyen en la fórmula y generan una modificación del valor promedio.

No existe una regla específica que indique cómo seleccionar la base del promedio móvil. Si la variable a pronosticar no presenta variaciones considerables, esto es, si su comportamiento es relativamente estable en el tiempo, se recomienda que el valor de n sea grande. Por el contrario, es aconsejable un valor pequeño de n si la variable muestra patrones cambiantes. En la práctica, los valores de n oscilan entre 2 y 10 [2].

El método de promedios móviles es muy útil cuando se tiene información no desagregada, y cuando no se conoce otro método más sofisticado que permita predecir con mayor confianza.

3.3.1.2 Suavización Exponencial

Otro método para realizar un pronóstico es el método de suavización exponencial [11]. A diferencia de los promedios móviles, este método predice otorgando una ponderación a los datos dependiendo del peso que tengan dentro del cálculo. Esta ponderación se lleva a cabo otorgando un valor a la constante de suavización α, y puede ser mayor que cero y menor a uno.

El método de Suavización Exponencial supone que el proceso es constante, al igual que el método de Promedios Móviles. Sin embargo, es un promedio ponderado de los valores reales y los valores pronosticados, a diferencia del otro método, en donde los datos para calcular el promedio tienen la misma ponderación. De manera particular, esta técnica considera que las observaciones recientes tienen mayor valor y le otorga mayor peso dentro del promedio.

La Suavización Exponencial utiliza un promedio móvil ponderado de los datos históricos de la serie de tiempo como pronóstico; es un caso especial de Promedio Móvil en donde se selecciona un solo valor de ponderación. El modelo básico de suavización exponencial se presenta la ecuación (3.3.2).

$$F_0 = X_0$$
$$F_t = \alpha X_{t-1} + (1-\alpha)F_{t-1} \tag{3.3.2}$$

Donde:

$F_t =$ Pronóstico de la serie de tiempo para el periodo t.
$X_{t-1} =$ Valor real del periodo $t-1$.
$F_{t-1} =$ Pronóstico del periodo $t-1$.
$\alpha =$ Constante de suavización $(0 \leq \alpha \leq 1)$.

La utilización de esta ecuación implica algunas especificaciones. El cálculo de F_t está ligado con los dos periodos anteriores. En otras palabras, el pronóstico de suavización exponencial en determinado periodo (F_t) es igual al valor real de la serie de tiempo en el periodo anterior (X_{t-1}), multiplicado por la constante de suavización (α), sumado $(1-\alpha)$ multiplicado por el pronóstico del periodo anterior (F_{t-1}).

A pesar de que la suavización exponencial entrega un pronóstico que es un promedio ponderado de todas las operaciones anteriores, no es necesario guardar todos los datos del pasado a fin de calcular el pronóstico para el periodo siguiente. De hecho, una vez seleccionada la constante de suavización α, sólo se requiere de dos elementos de información para calcular el pronóstico. La Ecuación (3.3.2) muestra que con un α dado se puede calcular el pronóstico para el periodo t, simplemente conociendo los valores reales y pronosticados de la serie de tiempo para el periodo t, es decir, X_{t-1} y F_{t-1}.

La elección de la constante de suavización α es crucial en la estimación de pronósticos futuros. Si la serie de tiempo contiene una variabilidad aleatoria sustancial, se preferirá un valor pequeño como constante de suavización. La razón de esta aseveración es que gran parte del error de pronóstico es provocado por la variabilidad aleatoria, por lo que un valor pequeño de α permite un pronóstico mejor. Por el contrario, para una serie de tiempo con una variabilidad aleatoria relativamente pequeña, valores más elevados de la constante de suavización tienen la ventaja de ajustar con rapidez los pronósticos cuando ocurren errores de pronóstico, y permiten, por lo tanto, que el pronóstico reaccione con mayor rapidez a las condiciones cambiantes. En la práctica, el valor de α está entre 0.01 y 0.90 [2].

3.3.2 Regresión

También llamada predicción o estimación. Consiste en aprender una función real que asigna a cada instancia un valor real. El objetivo es minimizar el error (generalmente el error cuadrático medio) entre el valor predicho y el valor real.

3.3.2.1 Regresión Lineal

El modelo de regresión lineal es el más utilizado a la hora de predecir los valores de una variable cuantitativa a partir de los valores de otra variable explicativa también cuantitativa (modelo de regresión lineal simple). Una generalización de este modelo, el de regresión lineal múltiple, permite considerar más de una variable explicativa cuantitativa. Por otra parte, es también posible incluir variables explicativas categóricas en un modelo de regresión lineal si se sigue una determinada estrategia en la codificación de los datos conocida como codificación ficticia.

En concreto, según el modelo de regresión lineal simple, las puntuaciones de los sujetos en dos variables, una de ellas considerada como variable predictor (X) y la otra como variable de respuesta (Y), vienen representadas (modeladas) por la ecuación (3.3.3).

$$Y = \beta_0 + \beta_1 X_1 \tag{3.3.3}$$

Cuando hay más de una variable explicativa (modelo de regresión lineal múltiple), se utiliza un subíndice para cada una de ella, por ejemplo para el caso de dos variables explicativas se muestran en la ecuación (3.3.4).

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{3.3.4}$$

En cualquier caso, las relaciones del tipo anterior raramente son exactas, más bien se trata de aproximaciones en las que se han omitido muchas variables de importancia secundaria, por esta razón se debe incluir un término de perturbación aleatoria u_t que refleje todos los factores (distintos de X) con influencia sobre la variable endógena, pues ninguno de ellos es relevante individualmente. Con ello, la relación obtenida se muestra en la ecuación (3.3.5).

$$Y = \beta_0 + \beta_1 X_1 + u_t \tag{3.3.5}$$

El objetivo principal de la regresión es la determinación de β_0 y β_1 a partir de la información contenida en las observaciones que se dispone. Esta estimación se puede llevar a cabo mediante diversos procedimientos [16].

3.3.2.2 Regresión Polinomial

En muchas ocasiones los datos no muestran una dependencia lineal. Esto es lo que sucede si, por ejemplo, la variable respuesta depende de las variables independientes según una función polinómica, dando lugar a una regresión polinómica que puede planearse agregando las condiciones polinómicas al modelo lineal básico. De esta forma y aplicando ciertas transformaciones a las variables, se puede convertir el modelo no lineal en uno lineal que puede resolverse entonces por el método de mínimos cuadrados. Por ejemplo considérese una relación polinómica cúbica dada por:

$$y = a + b_1 x + b_2 x^2 + b_3 x^3 \tag{3.3.6}$$

Para convertir esta ecuación a la forma lineal, se definen las nuevas variables:

$$x_1 = x \qquad x_2 = x^2 \qquad x_3 = x^3 \tag{3.3.7}$$

Con lo que la ecuación anterior puede convertirse entonces a la forma lineal aplicando los cambios de variables, y resultando la ecuación (3.3.8), que es resoluble por el método de mínimos cuadrados.

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 \tag{3.3.8}$$

No obstante, algunos modelos son especialmente no lineales como, por ejemplo, la suma de términos exponenciales y no pueden convertirse a un modelo lineal. Para estos casos, puede ser posible obtener las estimaciones del mínimo cuadrado a través de cálculos extensos en fórmulas más complejas [15].

En la tabla se pueden observar las principales transformaciones a las variables para convertirlas a un modelo lineal.

Tabla 3.1 Modelos de Regresión Lineal

Modelo	**Ecuación**	**Ecuación Alternativa**
Logarítmica	$Y = \beta_0 + \beta_1 \ln(x)$	
Cuadrática	$Y = \beta_0 + \beta_1 X + \beta_2 X_2$	
Cúbica	$Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_3$	
Potencia	$Y = \beta_0 X^{\beta_1}$	$\ln(Y) = \ln(\beta_0) + \beta_1 \ln(X)$
Compuesto	$Y = \beta_0 \beta_1 X$	$\ln(Y) = \ln(\beta_0) + X ln(\beta_1)$
Curva-S	$Y = exp^{\beta_0} + \beta_1 / t$	$\ln(Y) = \beta_0 + \beta_1 / t$
Crecimiento	$Y = exp^{(\beta_0 + \beta_1 X)}$	$\ln(Y) = \beta_0 + \beta_1 X$
Exponencial	$Y = \beta_0 exp^{\beta_1} X$	$\ln(Y) = \ln(\beta_0) + \beta_1 X$

3.3.3 Redes Neuronales

El interés por comprender los procesos cognitivos del cerebro humano fue, en definitiva lo que permitió el surgimiento de las Redes Neuronales Artificiales (RNA), debido a que el cerebro es un sistema de procesamiento de la información extremadamente complejo, cuyo modo de funcionamiento es eminentemente paralelo u cuyo comportamiento no puede describirse por medio de modelos sencillos como lo son los lineales. Se buscó modelarlo con la esperanza de que se pudieran crear sistemas pensantes que tuvieran mejores resultados en tareas como clasificación, problemas de decisión, predicciones y sistemas de control adaptables que un sistema computacional convencional.

Las RNA están inspiradas de la estructura del cerebro, y fueron concebidas para resolver cierto tipo de problemas especialmente mal resueltos por las técnicas de programación tradicionales. Formalmente, una red neuronal es un modelo computacional con un conjunto de propiedades específicas, como son la habilidad de adaptarse o aprender, generalizar u organizar la información, todo esto basado en un procesamiento eminentemente paralelo [17].

Cada neurona recibe inputs desde otras neuronas y genera un resultado que depende sólo de la información localmente disponible, ya sea almacenada internamente o plasmada en los ponderadores de las conexiones. El output generado por la neurona servirá de input para otras neuronas.

Mediante la adecuada modificación de los ponderadores de la red, en un proceso denominado aprendizaje, la red mejorará su desempeño en el desarrollo de la tarea para la cual fue construida.

Las Redes Neuronales tienen el potencial de implementar funciones complejas. Se puede demostrar que una Red Neuronal suficientemente grande, con una estructura y ponderadores adecuados, es capaz de aproximar cualquier función con el nivel de precisión que se desee [18].

Los elementos básicos que posee un modelo de red neuronal son los siguientes [19,20]:

a) Conexiones de entrada: Se representan como los datos de entrada, es decir, $x_1, \dots, x_n$ siendo n la cantidad de instancias. Los pesos asignados a estas conexiones se denotan como $w_1, \dots, w_n$, sin embargo, existe una conexión de entrada que es constante, llamada tendencia y se escribe x_0.

b) Función de entrada o propagación: Es aquella encargada de efectuar el cálculo acumulado de la red, por ende, tiende a tener la forma de una sumatoria, es decir $u = u(x, w) = \sum_{i=1}^{n} x_i w_i$.

c) Función de activación: También denominada señal, se encarga de generar el nivel de activación de la neurona, su notación es $a = a(u)$. Nótese que el argumento de esta función es la función de entrada de la neurona, la cual considera las observaciones y los pesos. Hay una infinidad de funciones para ser utilizadas como función de activación en una red neuronal artificial, pero se pueden distinguir tres grandes clases: tipo escalón, lineal y sigmoide. Es más frecuente utilizar funciones sigmoidales como la de la ecuación (3.3.9), puesto que éstas y sus derivadas son continuas.

$$Y_k = F_k(S_k) = \frac{1}{1+e^{S_k}} \tag{3.3.9}$$

d) Función de salida: Esta función entrega el resultado de la neurona, su notación será $o = o(u)$, sin embargo, se tiende a asumir que es igual a la función de activación, o sea $o = a$.

e) Tasa de aprendizaje: Es la tasa que se utiliza para ajustar la cantidad de modificación de los pesos en cada iteración del entrenamiento.

f) Parámetro de corte θ: Es aquel parámetro bajo el cual la neurona decide (mediante la función de activación) cuando una neurona es activada o no.

3.3.4 Support Vector Regression

Este modelo, también llamado SVRs por sus siglas en inglés, se basa en el concepto de planos de decisiones que definen los límites de decisión; utiliza Support Vector Machines (SVM) para realizar una regresión. Normalmente los SVM son utilizados como clasificadores, es decir, para separar elementos de una clase de elementos pertenecientes a otra. No obstante, las SVMs pueden emplearse para efectuar regresiones, ocasión en las que se les llama SVR.

El objetivo primordial al usar SVR es buscar y optimizar los límites dados para la regresión. Se fundamenta en el concepto de una función de pérdida que ignora el error, el cual usualmente es colocado a una cierta distancia de las observaciones reales. Hay situaciones en que el modelo lineal no es el más adecuado. Las SVRs permiten, de este modo, pasar un hiperplano que no es lineal para clasificar los elementos de mejor manera.

Para esto, el método consiste en realizar un re-ordenamiento de los elementos del conjunto inicial, utilizando diferentes tipos de funciones, llamadas funciones de Kernel [30]. Es relevante destacar que al efectuar esta transformación, los elementos se vuelven linealmente separables en el nuevo espacio.

Debido a que éste es uno de los algoritmos más nuevos y ampliamente usados en el área de pronósticos y regresiones, también será utilizado para los diversos experimentos de esta investigación.

En la construcción de un hiperplano óptimo, una SVR usa un algoritmo iterativo de entrenamiento con el fin de minimizar la función de error [30]. De acuerdo a la forma de la función de error, los modelos SVR se pueden clasificar en cuatro grupos distintos [21]:

a) Clasificación SVM tipo 1 (conocido como la clasificación C-SVM).
b) Clasificación SVM tipo 2 (conocido como la clasificación de nu-SVM).
c) Regresión SVM tipo 1 (conocido como la regresión epsilon-SVM).
d) Regresión SVM tipo 2 (conocido como la regresión nu-SVM).

3.3.4.1 Regresión Usando SVMs Tipo 1

Para ejecutar una regresión los datos deben cumplir la ecuación (3.3.10)

$$y = f(x) + error \tag{3.3.10}$$

La tarea consiste, entonces, en encontrar una forma funcional para f que pueda predecir correctamente los nuevos casos que la SVM no ha presentado en la anterior. Esto se puede lograr mediante el entrenamiento del modelo de SVM sobre un conjunto de muestras, es decir, un conjunto de ejercicio, proceso que implica, al igual que la clasificación, la optimización secuencial de una función de error. Dependiendo de la definición de esta función, dos tipos de modelos SVM se pueden reconocer:

$$error = \frac{1}{2} w^t w + C \sum_{i=1}^{N} \xi_i + C \sum_{i=1}^{N} \xi'_i \tag{3.3.11}$$

El cual se debe minimizar de acuerdo a varias restricciones:

$$w^t\phi(x_i) + b - y_i \leq \varepsilon + \xi'_i \tag{3.3.12}$$

$$y - w^t\phi(x_i) - b_i \leq \varepsilon + \xi_i \tag{3.3.13}$$

$$\xi_i, \xi'_i \geq 0;\ i = 1, \ldots, N \tag{3.3.14}$$

3.3.4.2 Regresión Usando SVMs Tipo 2

En este caso, también se utiliza la ecuación (3.3.10), pero la expresión para denotar el error del modelo es diferente.

$$error = \frac{1}{2}w^t w - C(v\varepsilon + \frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi'_i) \tag{3.3.15}$$

$$(w^t\phi(x_i) + b) - y_i \leq \varepsilon + \xi_i \tag{3.3.16}$$

$$y_i - (w^t\phi(x_i) + b) \leq \varepsilon + \xi'_i \tag{3.3.17}$$

$$\varepsilon, \xi_i, \xi'_i \geq 0;\ i = 1, \ldots, N \tag{3.3.18}$$

3.3.4.3 Funciones de Kernel

Las funciones kernel que se indican a continuación, son las funciones estándar para transformar el espacio [22].

$$Lineal: \phi = x_i x_j \tag{3.3.19}$$

$$Polinomial: \phi = (x_i x_j + coef)^{\alpha} \tag{3.3.20}$$

$$Gaussiana: \phi = exp\left(-\frac{\|x_i - x_j\|^2}{y}\right) \tag{3.3.21}$$

$$Sigmoidal: \phi = \tanh(y(x_1 x_j) + coef) \tag{3.3.22}$$

3.4 Herramientas Para Realizar Análisis de los Datos

En esta sección se presentarán algunas herramientas que pueden ser utilizadas para el problema señalado, empleando una diversidad de algoritmos y modelos que traen integradas cada una de ellas, los cuales serán analizados y estudiados, de tal forma de realizar la selección de la herramienta y algoritmo(s) lo más óptima posible para lograr dar solución al problema.

3.4.1 Clementine

Es una herramienta visual desarrollada por ISL (Integral Solutions Limited) y comercializada por SPSS que constituye uno de los sistemas más populares en el mercado. Entre sus características más significativas se destaca el hecho de que a diferencia de otras herramientas que se centran en el modelado, ella apoya el ciclo completo de KDD y está diseñada bajo la metodología CRISP-DM. Posee una arquitectura distribuida (cliente/servidor). Permite el uso de técnicas de aprendizaje tales como: redes neuronales, árboles de decisión (C5.0 y CART), agrupamiento (K-Medias), reglas de asociación (GRI, A priori, etc.), regresión lineal y regresión logística, entre otras. Posee un potente soporte gráfico que permite al usuario tener una visión global de todo el proceso y que comprende gráficos estadísticos, gráficos 3-D y animados, visualizadores interactivos de las diferentes tareas que realiza el experto y navegadores para árboles de decisión, reglas de asociación, redes neuronales de Kohonen, agrupamientos, etc. Permite trabajar con datos estructurados (tabulares) en diferentes formatos de bases de datos, archivos de texto y hojas de cálculo de Excel. También permite hacer minería de texto y minería Web. SPSS Clementine es un sistema multiplataforma. La aplicación está disponible para sistemas Windows, Sun Solaris, HP-UX AIX y OS/400 [12].

3.4.2 SAS

Es una herramienta comercial proporcionada por SAS. Su diseño está inspirado por la metodología SEMMA. Entre sus características más significativas se encuentra el hecho de que posee una arquitectura distribuida y una potente interfaz gráfica de usuario. La herramienta tiene soporte para almacenes de datos y permite trabajar con archivos en un formato propio de SAS y de sistemas de bases de datos comerciales. Incluye técnicas para ayudar al pre-procesado de datos. Además implementa algoritmos que proveen modelos predictivos y descriptivos, tales como árboles de decisión, redes neuronales, asociación, agrupamiento, entre otros. Permite la visualización y representación de los resultados mediante información en lenguaje natural, gráficos en dos o tres dimensiones y un generador automático de reportes que resume los resultados en un informe HTML. Tanto el programa cliente como servidor de SAS Enterprise Manager, puede ser trasladado a diferentes plataformas: Windows, Linux, Solaris, HP-UX, Digital Unix, etc [13].

3.4.3 RapidMiner

Fue implementado en Java por la Universidad de Dortmund para la realización de experimentos de aprendizaje automático. Funciona en los sistemas operativos Windows y Linux. Es un software de código abierto y de libre distribución. Se retroalimenta de las librerías de funciones de WEKA en su entorno de aprendizaje. En julio del 2007 cambió su nombre por RapidMiner. Permite la entrada de datos en formato Microsoft Excel y SPSS. Incluye operadores para el aprendizaje automático (máquina de vectores soporte, árboles de decisión, agrupamiento y algoritmos genéticos). Desde la perspectiva de la visualización ofrece representaciones de datos en dispersión en 2D y 3D; coordenadas paralelas y grandes posibilidades de transformar las visualizaciones de los datos [13].

3.4.4 WEKA

Acrónimo de Waikato Enviroment for Knowledge Analysis, es una de las aplicaciones de minería más populares, desarrollada por un equipo de investigadores de la Universidad de Waikato (Nueva Zelanda). Una de las ventajas fundamentales de esta herramienta es que su desarrollo sobre el lenguaje java la hace multiplataforma. Además, el hecho de ser de código abierto unido a su prestigio, hace que se encuentre en constante evolución por parte de la comunidad internacional. El formato de entrada de los datos es un archivo plano organizado en filas y columnas (formato ARFF). Incluye una gran cantidad de filtros para el pre procesado de los datos. Está formado por una serie de paquetes de código abierto con diferentes implementaciones de las técnicas de clasificación, asociación, agrupamiento y visualización de datos. Posee una interfaz gráfica de usuario compuesta de cuatro entornos que permiten diferentes funcionalidades y formas de análisis [13].

3.4.5 KNIME

Es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual el cual está construido bajo la plataforma de Eclipse y programado esencialmente en Java.

Knime contiene una serie de nodos (que encapsulan distintos tipos de algoritmos) y flechas (que representan flujos de datos) que se despliegan y combinan de manera gráfica e interactiva. Los nodos implementan distintos tipos de acciones que pueden ejecutarse sobre una tabla de datos:

- Manipulación de filas, columnas, etc., muestreos, transformaciones, agrupaciones, etc.
- Visualización (histogramas, etc.).
- Creación de modelos estadísticos y de minería de datos, como árboles de decisión, máquinas de vector soporte, regresiones, etc.
- Validación de modelos, como curvas ROC, etc.
- Scoring o aplicación de dichos modelos sobre conjuntos nuevos de datos.
- Creación de informes a medida gracias a su integración con BIRT.

El carácter abierto de la herramienta hace posible su extensión mediante la creación de nuevos nodos que implementen algoritmos a la medida del usuario. Además, existe la posibilidad de llamar directa y transparentemente a Weka y/o de incorporar de manera sencilla código desarrollado en R o python/jython [14].

3.5 Evaluación de los Pronósticos

Para realizar una correcta medición en la calidad de los pronósticos, se aplicarán distintos cálculos estadísticos como lo son la desviación absoluta, el error absoluto medio, el error absoluto porcentual promedio y la correlación.

En estadística la desviación absoluta de un elemento en una colección de datos, corresponde a la diferencia absoluta entre ese elemento y un punto dado de la colección [24]. Comúnmente, el punto desde el cual se mide la desviación es un punto que calcula la tendencia central de la muestra, es decir, la mediana, la moda o la media.

La ecuación (3.5.1) muestra la expresión general de la desviación absoluta, donde se tiene un set de n observaciones $\{x_1, x_2, x_3, \dots, x_n\}$ y, $m(x)$ es una medida de tendencia central.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - m(x)| \tag{3.5.1}$$

En particular, si se toma la media para el cálculo se estaría hablando de Mean Absolute Deviation o MAD, que es una medida comúnmente utilizada para calcular la calidad de un pronóstico en series de tiempo.

Otra medida relacionada al MAD, es el Mean Absolute Error o MAE, que indica la diferencia absoluta entre el valor real y el valor pronosticado. Este indicador se enfoca mucho más en la salida de los modelos de pronóstico. La expresión del MAE se encuentra en la ecuación (3.5.2) y la misma es posible escribirla en términos porcentuales, lo cual da origen al MAPE o Mean Absolute Percentage Error, cuya expresión se encuentra en la ecuación (3.5.3).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \tag{3.5.2}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{f_t - y_t}{f_t}\right| \tag{3.5.3}$$

Finalmente, en este trabajo se empleará el MAPE en la medición de la calidad global de los pronósticos calculados, dado que es posible estimar este indicador para todos los métodos utilizados; también se realizarán comparación de resultados entre dos herramientas y así lograr una mejor estimación.

4 Propuesta de Solución

En este capítulo se detalla la propuesta realizada para el pronóstico de superficie y producción en la agricultura, en función de los objetivos planteados desde un comienzo. Esta propuesta se basa en la metodología CRISP-DM descrita en el punto 3.2.1 y se detallará bajo un análisis exploratorio de los datos que ayuden en la aplicación de algoritmos tradicionales, lo cual se realizará en la etapa de Modelado.

4.1 Metodología y Herramientas Utilizadas

Se realizará el proyecto en la herramienta WEKA 3.8 al ser OpenSource y gratuita, esto permitirá realizar modificaciones al código fuente de ser necesario para obtener mejores resultados. Además se pueden obtener reportes gráficos lo que ayudara a una mejor evaluación de los resultados obtenidos.

A continuación se presentan las etapas del modelo CRISP-DM en el contexto de la problemática expuesta. Prestando especial atención a las etapas de modelado y evaluación, ya que son el foco de este proyecto.

4.1.1 Análisis del Problema

Es la etapa en la cual se define claramente cuál es la problemática existente y qué se debe resolver. Para este caso se pretende contextualizar e interpretar el problema centrado en la búsqueda de un modelo que genere conocimiento relevante en cuanto al pronóstico. El análisis del problema está realizado en la primera etapa de este informe.

4.1.2 Análisis de los Datos

Paralelo al entendimiento del problema se recolectan los datos necesarios de todas las fuentes que se disponen y que van a formar parte del estudio. Por lo que en etapa comienza la familiarización con la terminología asociada, los atributos que constituyen a cada registro y la documentación existente.

Los datos con los que se disponen provienen del Instituto Nacional de Estadísticas (INE) y corresponden a los obtenidos en el informe anual agropecuario desde el año 1980 a la fecha. Cabe aclarar que el año agrícola comienza el 1 de mayo y termina el 30 de abril del año siguiente. Los datos con los que se cuentan corresponden a la superficie sembrada, producción y rendimiento de los cultivos. Estos datos provienen de las encuestas realizadas entre octubre y diciembre de cada año, donde se consulta por la superficie sembrada, y en el mes de mayo, donde se consulta por la producción y rendimiento obtenidos, para posteriormente emitir el informe anual entre octubre y noviembre.

Las encuestas levantadas por el INE se aplica a una muestra probabilística de un tamaño muestral que asegura un coeficiente de error relativo de 5%, con un nivel de confianza de 95%. La población objetivo está constituida por todas las explotaciones de una o más

hectáreas, que declararon tener cultivos anuales esenciales en el año agrícola (2006-2007) que corresponde al VII Censo Agropecuario y Forestal 2007.

Además se cuenta con datos correspondientes a los mm de agua caída. Se consideraron las estaciones meteorológicas más importantes para cada región. Estos datos fueron obtenidos a través del sitio de meteorología de Chile desde 1980 en adelante.

El pronóstico se realizará para los casos del trigo, el cual es el más importante dentro de la categoría de los cereales, y para la papa, el cual resalta en la categoría leguminosas y tubérculos, seleccionando para cada caso las dos regiones más importantes.

Los gráficos de los datos correspondientes a la estimación de superficie sembrada, y producción para el trigo se encuentran en las figuras 4.1 y 4.2. Para más información ver tabla A4.1 en Anexo A.

Figura 4.1 Superficie y producción de trigo en El Biobío.

Figura 4.2 Superficie y producción de trigo en La Araucanía.

Los gráficos de los datos correspondientes a la estimación de superficie sembrada, y producción para la papa se encuentran en las figuras 4.3 y 4.4. Para más información ver tabla A4.2 en Anexo A.

Figura 4.3 Superficie y producción de papa en Los Ríos y Los Lagos.

Figura 4.4 Superficie y producción de papa en La Araucanía.

Los gráficos de los datos correspondientes al agua caída se encuentran en las figuras 4.5, 4.6 y 4.7. Para más información ver tabla A4.3 en Anexo A.

Figura 4.5 Agua caída en El Biobío.

Figura 4.6 Agua caída en La Araucanía.

Figura 4.7 Agua caída en Los Ríos y Los Lagos.

En las figuras correspondientes a superficie sembrada, se puede observar una tendencia a la baja en el caso del trigo y una estabilidad en el caso de la papa. En cuanto a las figuras correspondientes a la producción, en el caso de la papa se observa un aumento de forma paulatina y en el trigo se observa una mayor inestabilidad.

4.1.3 Preparación de los Datos

Una vez obtenida la información histórica, se procede a la importación, transformación y selección de los datos según los objetivos planteados.

En un primer paso se realiza la importación de los datos, los cuales provienen de un archivo PDF, por lo que manualmente fueron copiados en un archivo .xslx para manipularlos fácilmente, luego se creó un archivo .arff el cual es el tipo de archivo con que trabaja WEKA.

Como se puede apreciar en los gráficos correspondientes a producción (Figura 4.1, 4.2, 4.3 y 4.4), hay un dato faltante, el cual corresponde a la temporada 1994/95, esto es debido a que no se realizó la encuesta correspondiente para obtenerlos. Para tales casos es necesario realizar una transformación y buscar un valor estimado de aquellos datos. Usando una interpolación lineal en la papa y el trigo, debido a que los tramos donde está el dato faltante la tendencia es a la baja, el resultado fue de 314.193 Tn y 350.643 Tn para el trigo en la región

del Biobío y La Araucanía respectivamente, y 179.942 Tn y 262.435 Tn para la papa en la región de La Araucanía y Los Ríos con Los Lagos respectivamente.

En cuanto a los datos de agua caída, se seleccionaron las estaciones meteorológicas mas importantes de cada región analizada. Se obtuvieron los datos desde el sitio web de climatología de meteorología de Chile. Estos datos solo se encontraban disponibles en la web, por lo que fue necesario digitarlos en el formato requerido uno a uno. Luego, al estar disponibles de forma mensual, pudimos agruparlos de forma que calzaran con el año agrícola, y así no tener desfase con los datos y generar resultados erróneos.

4.1.4 Modelado

Para el desarrollo del modelo predictivo, es posible utilizar distintas técnicas de minería de datos que permiten modelar un problema de pronóstico tal como se explica en el Estado del Arte.

En esta etapa se presentarán tareas orientadas al descubrimiento del conocimiento, por lo que se mostrarán los distintos resultados que arrojaran todas las técnicas antes mencionadas.

Los siguientes gráficos corresponden a análisis realizados con una red neuronal con el algoritmo backpropagation, ya que este modelo fue el que arrojó mejores resultados al momento de evaluar la superficie sembrada en función del tiempo y el agua caída, y la producción en función del tiempo, agua caída y superficie sembrada (ver Anexo B para otros pronósticos). Se pronosticó la serie para dos periodos futuros.

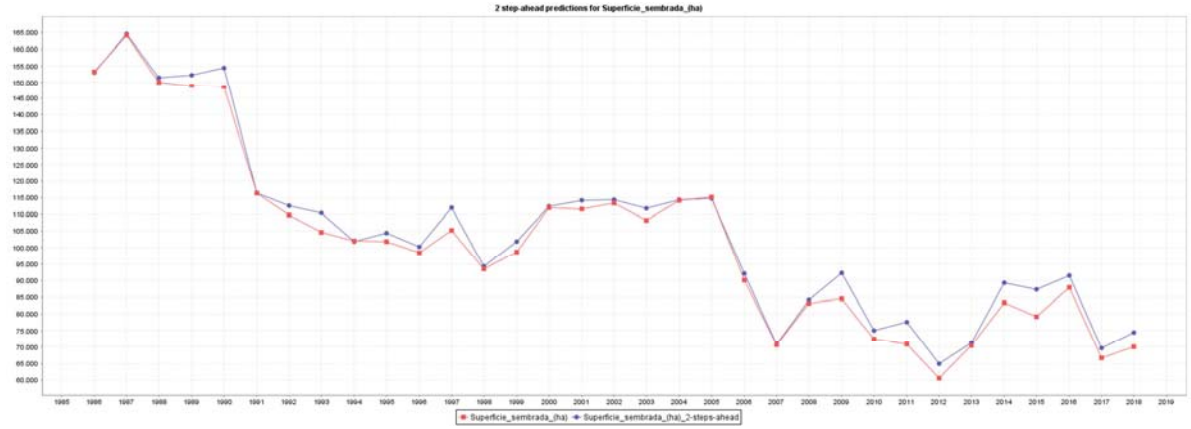

Figura 4.8 Red neuronal para siembra de trigo en Biobío (en función del tiempo y agua caída).

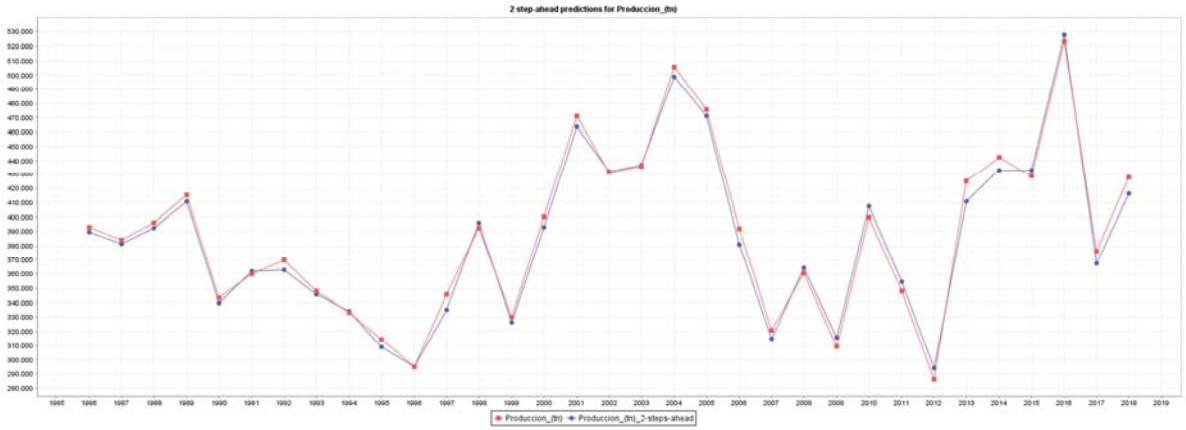

Figura 4.9 Red neuronal para producción de trigo en Biobío (en función del tiempo, agua caída y superficie sembrada).

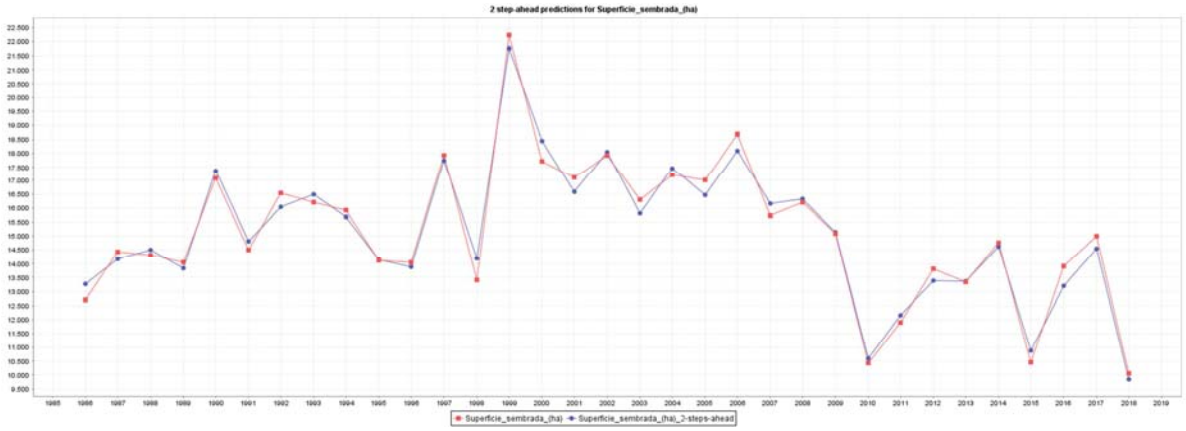

Figura 4.10 Red neuronal para siembra de papa en Los Ríos y Los Lagos (en función del tiempo y agua caída).

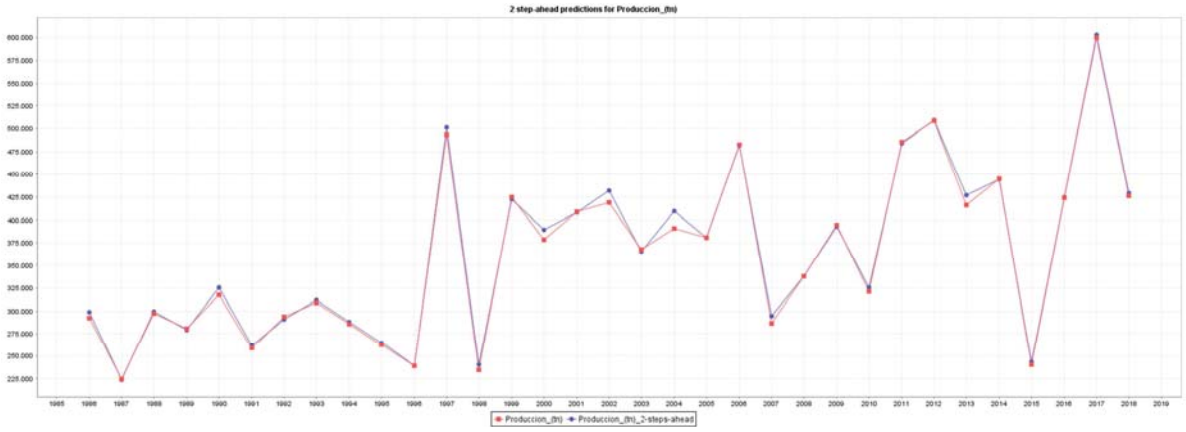

Figura 4.11 Red neuronal para producción de papa en Los Ríos y Los Lagos (en función del tiempo, agua caída y superficie sembrada).

Posteriormente se realizaron los pronósticos con cada una de las técnicas para la producción, pero esta vez considerando la variable superficie sembrada como dependiente. De esta forma poder observar si aporta información relevante para el pronóstico. Dichos gráficos no serán adjuntados debido a que no es posible apreciar ambas tendencias al tener valores muy lejanos uno del otro. Los resultados de estos pronósticos pueden observarse en el punto 4.1.5.

4.1.5 Evaluación

En las siguientes tablas, se encuentran los resultados para los distintos modelamientos realizados; ya sea regresión lineal, red neuronal y SVR. Lo anterior se realizó para cada variable de forma independiente, solo en función del tiempo.

Tabla 4.1 Rendimiento de regresión lineal para superficie sembrada.

	Superficie Sembrada con trigo (Biobío)		**Superficie Sembrada con trigo (Araucanía)**		**Superficie sembrada con papa (Araucanía)**		**Superficie sembrada con papa (Los Ríos y Los Lagos)**	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	6.460	6.823	8.124	9.734	1.406	1.552	1.938	1.978
MAPE	6,883	7,2359	6,4929	7,914	10,4616	11,5323	13,6562	14,565
Pronóstico	68.324	41.661	96.957	93.561	11.455	10.646	12.376	13.834

Tabla 4.2 Rendimiento red neuronal para superficie sembrada.

	Superficie Sembrada con trigo (Biobío)		**Superficie Sembrada con trigo (Araucanía)**		**Superficie sembrada con papa (Araucanía)**		**Superficie sembrada con papa (Los Ríos y Los Lagos)**	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	2.563	2.843	2.368	2.464	132	159	324	339
MAPE	2,7082	3,1692	1,8798	2,0065	0,9628	1,2066	2,1624	2,2615
Pronóstico	61.603	52.806	70.050	106.492	14.294	13.610	17.040	7.932

Tabla 4.3 Rendimiento red SVR para superficie sembrada.

	Superficie Sembrada con trigo (Biobío)		**Superficie Sembrada con trigo (Araucanía)**		**Superficie sembrada con papa (Araucanía)**		**Superficie sembrada con papa (Los Ríos y Los Lagos)**	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	4.395	6.695	8.296	11.208	1.073	1.289	919	1.138
MAPE	4,7456	7,3842	6,5985	9,0837	7,9126	9,401	6,58	8,0628
Pronóstico	58.814	54.731	94.389	88.319	11.991	12.917	13.424	8.188

Tabla 4.4 Rendimiento de regresión lineal para producción.

	Producción de trigo (Biobío)		**Producción trigo (Araucanía)**		**Producción de papa (Araucanía)**		**Producción de papa (Los Ríos y Los Lagos)**	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	31.182	61.698	50.594	62.228	50.807	81.749	44.486	42.5198
MAPE	8,1883	8,3217	9,9472	12,0898	20,639	20,8152	13,5128	12,9932
Pronóstico	420.825	388.648	600.740	556.581	334.877	338.863	565.060	361.860

Tabla 4.5 Rendimiento de red neuronal para producción.

	Producción de trigo (Biobío)		Producción trigo (Araucanía)		Producción de papa (Araucanía)		Producción de papa (Los Ríos y Los Lagos)	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	3.223	4.729	14.308	16.688	17.193	16.918	6.162	6.195
MAPE	0,8616	1,2388	3,0809	3,4926	8,4872	7,9074	1,7415	1,7615
Pronóstico	540.136	395.409	597.061	642.176	261.266	314.088	655.876	88.216

Tabla 4.6 Rendimiento de SVR para producción.

	Producción de trigo (Biobío)		Producción trigo (Araucanía)		Producción de papa (Araucanía)		Producción de papa (Los Ríos y Los Lagos)	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	26.660	30.911	52.013	70.795	43.334	45.244	30.042	37.264
MAPE	7,3297	8,2218	9,9461	12,8856	15,0814	15,7016	9,3654	11,3938
Pronóstico	456.708	398.160	631.867	677.927	379.920	342.732	499.826	500.159

En las siguientes tablas, se encuentran los resultados para los distintos modelamientos realizados; ya sea regresión lineal, red neuronal y SVR. Lo anterior realizado para la variable producción en función de la superficie sembrada, el agua caída y el tiempo.

Tabla 4.7 Rendimiento de regresión lineal para producción.

	Producción de trigo (Biobío)		Producción trigo (Araucanía)		Producción de papa (Araucanía)		Producción de papa (Los Ríos y Los Lagos)	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	33.117	33.988	38.461	47.977	48.422	53.606	33.759	38.849
MAPE	8,8111	8,8421	7,3917	9,2704	18,6401	20,586	10,0776	12,1469
Pronóstico	467.506	415.441	491.489	599.077	400.301	406.664	502.329	453.732

Tabla 4.8 Rendimiento de red neuronal para producción.

	Producción de trigo (Biobío)		Producción trigo (Araucanía)		Producción de papa (Araucanía)		Producción de papa (Los Ríos y Los Lagos)	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	3.679	5.532	51.820	34.205	14.538	12.947	1.176	3.885
MAPE	0,985	1,4376	11,801	6,7875	7,239	6,2657	0,3363	1,1139
Pronóstico	382.917	352.960	444.620	627.395	240.405	504.430	567.232	170.437

Tabla 4.9 Rendimiento de SVR para producción.

	Producción de trigo (Biobío)		Producción trigo (Araucanía)		Producción de papa (Araucanía)		Producción de papa (Los Ríos y Los Lagos)	
	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20	2018/19	2019/20
MAE	23.330	28.352	26.831	44.231	30.774	34.779	21.883	30.478
MAPE	6,2561	7,5354	4,9677	8,6727	10,8474	12,3398	6,5864	9,1204
Pronóstico	430.281	374.577	541.748	566.432	400.687	383.961	556.159	403.528

Analizando estos resultados, se puede observar que los mejores rendimientos fueron utilizando redes neuronales, esto puede deberse a que en general los datos son no lineales.

Por lo general en temas referentes a pronósticos, lo principal es detectar la tendencia que seguirá en periodos futuros. En el caso de la superficie sembrada con trigo en la región del Biobío y con papa en la región de la Araucanía principalmente, la tendencia está claramente marcada para ambos periodos futuros. Sin embargo en los otros pronósticos, en particular los de producción, estos generan confusión en algunas de las regiones.

Con el fin de conocer otro método de comparación, se realizó un pronóstico para los años 2016/17 y 2017/18, luego se comparó con los ya conocidos. Los resultados para la superficie sembrada arrojaron los mejores resultados y se muestran en las tablas 4.10 y 4.11 y 4.12. Para el caso de producción los resultados arrojaron gran margen de error y se encuentran en las tablas A4.4 a A4.8 del Anexo A.

Tabla 4.10 Superficie sembrada (ha) para periodo 2016/2017 y 2017/2018.

	Superficie Sembrada con trigo (Biobío)		**Superficie Sembrada con trigo (Araucanía)**		**Superficie sembrada con papa (Araucanía)**		**Superficie sembrada con papa (Los Ríos y Los Lagos)**	
	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18
Superficie sembrada	66.742	70.066	94.526	107.529	13.886	12.486	15.001	10.067

Tabla 4.11 Pronóstico de superficie sembrada (ha) para periodo 2016/2017 y 2017/2018.

	Producción de trigo (Biobío)		**Producción trigo (Araucanía)**		**Producción de papa (Araucanía)**		**Producción de papa (Los Ríos y Los Lagos)**	
	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18
Regresión lineal	107.740	96.731	113.527	122.305	14.173	11.063	14.977	15.638
Red neuronal	83.126	64.757	95.092	95.875	10.501	12.049	12.293	8.910
SVR	87.087	83.215	105.027	106.081	15.037	13.114	10.803	12.597

Tabla 4.12 Error en pronóstico de superficie sembrada (ha) para periodo 2016/2017 y 2017/2018.

	Producción de trigo (Biobío)		**Producción trigo (Araucanía)**		**Producción de papa (Araucanía)**		**Producción de papa (Los Ríos y Los Lagos)**	
	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18
Regresión lineal	38,05%	27,56%	16,73%	12,08%	2,02%	12,86%	0,16%	35,62%
Red neuronal	19,70%	8,19%	0,59%	12,15%	32,2%	3,62%	22,02%	12,98%
SVR	23,36%	15,80%	9,99%	1,36%	7,65%	4,78%	38,85%	20,08%

5 Conclusiones

Con los antecedentes recopilados a través de este informe se espera, en primera instancia, se haya logrado entender, y al mismo tiempo, comprender la problemática expuesta, referente al tema de pronósticos en el área de la agricultura.

En segundo lugar por medio del estado de arte, se quiso plasmar el término de proceso de descubrimiento del conocimiento, mejor conocido como KDD. Además de las metodologías más conocidas en este tema. Por otro lado, cuales son las técnicas y herramientas existentes para resolver el problema planteado.

También, es importante destacar que el rubro donde se aplica este proyecto es complejo, y a su vez, interesante; ya que al tratarse de agricultura puede influir un factor que no es menor, el climático, ya que mientras el año sea más lluvioso, la producción puede aumentar o disminuir considerablemente, independiente de la cantidad de hectáreas que fueron plantadas.

En cuanto a los resultados es preciso aclarar que el valor correspondiente al MAPE, es calculado en base a los datos de training (datos conocidos), y estos muestran que el mejor modelamiento para la data conocida vendría siendo con redes neuronales. Es de suponer que esto es debido a que se está trabajando con datos que no son lineales. Por otro lado, al crear pronósticos para datos ya conocidos y así poder comparar, se puede apreciar que estos arrojan buenos resultados solo para pronosticar la superficie sembrada, por lo que es de suponer que pronosticar la producción es muy complejo debido a los factores que pueden estar involucrados, predominantemente climáticos. En el estudio solo se consideró la variable de agua caída.

Finalmente, en base a la solución propuesta se puede concluir que los modelos analizados no son adecuados para pronosticar series altamente complejas como las presentadas en este proyecto, por lo que sería necesario crear un modelo en el que fuera posible mezclar distintas técnicas e involucrar otras variables, para así llegar a mejores resultados.

6 Referencias

[1] Valcárcel V., "*Data Mining y el descubrimiento del conocimiento"*, Industrial Data, vol. 7, no. 2, pp. 83-86, Perú, 2004.

[2] Fayyad U., Piatetsky-shapiro G. y Smyth P., *"Data Mining to Knowledge Discovery in Databases",* AI Magazine, Vol. 17, pp. 37-54, 1996.

[3] Witten I. H., Frank E., "*Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann Publishers, second edition, San Francisco, CA, 2005.

[4] García H., "*Avances en Informática y Sistemas Computacionales*", CONAIS, tomo II, primera edición, pp. 44-47, México, 2007.

[5] KDnuggets, "*KDnuggets: Polls: Data Mining Methodology*", 2007

[6] Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C. y Wirth R., *"CRISP-DM 1.0 step-by-step data mining guide"*, Technical report, The CRISP-DM consortium, pp. 10-29, 2000.

[7] Azevedo A. y Santos M., "*KDD, SEMMA and CRISP-DM: a parallel overview"*, IADIS European Conference Data Mining, pp. 182-185, 2008.

[8] Berry M. y Linoff G., *"Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management"*, Wiley Publishing, second edition, 2004.

[9] Two Crows Corporation, *"Introduction to Data Mining and Knowledge Discovery"*, third edition, pp. 9-10, 2005.

[10] Yaffee R. y McGee M., *"Introduction to Time Series Analysis and Forecasting: with applications of SAS and SPSS"*, Academic Press, Inc., 2000.

[11] Instituto Tecnológico y de Estudios Superiores de Monterrey, *"Métodos Estadísticos para la Estimación de Ingresos"*, México, 2006.

[12] Norris D., *"Clementine data mining workbench: from SPSS"*, Bloor Research, pp. 3-7, 2005.

[13] Gómez N., González R. y Rosete A., *"Predicción de complicaciones cardiacas utilizando Minería de Datos: Estado del Arte"*, 14 Convención Científica de Ingeniería y Arquitectura, 2008.

[14] Cuevas A., *"Web Usage Mining aplicado al estudio del comportamiento del los usuarios en el sistema de biblioteca de la PUCV"*, Informe Final del Proyecto para Optar al Título Profesional de Ingeniero Civil en Informática, pp. 40, 2010.

[15] Molina J. y García J., *"Técnicas de Análisis de Datos: Aplicaciones Prácticas Utilizando Microsoft Excel y WEKA"*, Madrid, 2006.

[16] Molina G. y Rodrigo M. F., *"Estadística descriptiva en Psicologia"*, Universidad de Valencia, pp. 1-3, 2009.

[17] McCulloch W. S. y Pitts W. *"A logical calculus of the ideas immanent in nervous activity"*, Bulletin of Mathematical iophysics, Vol. 5, pp. 115-133, 1943.

[18] Reed R. y Marks II R., *"Neural Smithing"*, The MIT Press, 1999.

[19] Fausett L., *"Fundamentals of neural networks: architectures, algorithms, and applications"*, Prentice-Hall Inc., Upper Saddle River, NJ, USA, 1994.

[20] Kasabov N., *"Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering"*, The MIT Press, first edition, Cambridge, MA, USA, 1996.

[21] Gárnica D., *"Pronostico a corto plazo de afluencia de pasajeros utilizando técnicas de data mining: Metro S. A."*, Tesis para optar al grado de magister en gestión de operaciones, 2011.

[22] Cristianini N. y Shawe-Taylor J., *"An Introduction to Support Vector Machines and Other Kernel-based Learning Methods"*, first edition, Cambridge University Press, 2000.

[23] Aránguiz I., *"Análisis de Accidentes de Tránsito en Zonas Urbanas y Rurales Usando Minería de Datos Difusa"*, Tesis de Grado Magíster en Ingeniería Informática, 2012.

[24] Larose D., *"Discovering Knowledge in Data"*, Wiley, New Jersey, USA, 2005.

Anexos

A: Tablas

Tabla A4.1 Datos pertenecientes al trigo.

	Región del Biobío			Región de La Araucanía		
Año agrícola	**Superficie Sembrada (ha)**	**Producción (tn)**	**Rendimiento (tn/ha)**	**Superficie Sembrada (ha)**	**Producción (tn)**	**Rendimiento (tn/ha)**
1979/80	160.850	249.113	1,55	175.600	313.773	1,79
1980/81	127.270	175.583	1,38	140.010	201.347	1,44
1981/82	123.340	196.111	1,59	105.990	151.566	1,43
1982/83	113.700	184.653	1,62	116.490	168.542	1,45
1983/84	154.610	320.632	2,07	133.810	243.721	1,82
1984/85	136.510	314.758	2,31	142.450	269.718	1,89
1985/86	153.230	392.878	2,56	156.380	358.437	2,29
1986/87	164.220	384.081	2,34	189.400	437.199	2,31
1987/88	150.090	395.911	2,64	171.250	479.766	2,80
1988/89	148.970	415.350	2,79	160.160	492.685	3,08
1989/90	148.590	343.630	2,31	177.620	422.759	2,38
1990/91	116.400	359.848	3,09	148.690	422.250	2,84
1991/92	109.870	369.714	3,37	144.730	433.720	3,00
1992/93	104.640	347.974	3,33	132.180	403.096	3,05
1993/94	102.150	333.123	3,26	108.590	360.503	3,32
1994/95	101.807	-	-	132.304	-	-
1995/96	98.470	295.263	3,00	130.210	340.793	2,62
1996/97	105.316	345.569	3,28	140.863	510.702	3,63
1997/98	93.592	392.291	4,19	137.694	628.633	4,57
1998/99	98.667	330.003	3,34	144.611	514.911	3,56
1999/00	112.006	400.326	3,57	160.663	595.282	3,71
2000/01	111.600	471.001	4,22	166.970	668.217	4,00
2001/02	113.330	431.516	3,81	168.660	698.162	4,14
2002/03	108.250	436.297	4,03	161.000	693.628	4,31
2003/04	114.100	505.463	4,43	158.900	726.173	4,57
2004/05	115.200	475.776	4,13	160.910	699.959	4,35
2005/06	90.070	391.805	4,35	122.000	551.440	4,52
2006/07	70.735	320.776	4,53	94.394	453.136	4,80
2007/08	83.100	360.654	4,34	109.700	482.680	4,40
2008/09	84.519	309.340	3,66	107.431	431.873	4,02
2009/10	72.479	399.635	5,51	115.665	655.806	5,67
2010/11	70.966	348.046	4,90	115.572	680.719	5,89

2011/12	60.641	286.464	4,72	106.791	504.054	4,72
2012/13	70.438	425.112	6,04	105.528	578.305	5,48
2013/14	83.331	442.685	5,31	99.224	527.871	5,32
2014/15	79.240	428.661	5,41	107.869	598.673	5,55
2015/16	87.901	523.299	5,95	107.388	626.383	5,83
2016/17	66.742	376.305	5,64	94.526	597.835	6,32
2017/18	70.066	427.752	6,10	107.529	663.517	6,17

Tabla A4.2 Datos pertenecientes a la papa.

	Región de La Araucanía			**Región de Los Ríos y Los Lagos**		
Año agrícola	**Superficie Sembrada (ha)**	**Producción (tn)**	**Rendimiento (tn/ha)**	**Superficie Sembrada (ha)**	**Producción (tn)**	**Rendimiento (tn/ha)**
1979/80	14.270	99.741	6,99	31.020	402.777	12,98
1980/81	14.730	109.020	7,40	27.600	384.118	13,92
1981/82	10.410	75.582	7,26	24.480	306.734	12,53
1982/83	9.880	50.317	5,09	22.480	292.424	13,01
1983/84	12.410	97.359	7,85	25.540	418.439	16,38
1984/85	11.260	133.003	11,81	17.300	341.304	19,73
1985/86	9.900	126.861	12,81	12.710	292.780	23,04
1986/87	9.080	115.667	12,74	14.420	224.878	15,59
1987/88	9.280	111.006	11,96	14.310	297.660	20,80
1988/89	9.540	95.489	10,01	14.050	280.565	19,97
1989/90	9.400	104.788	11,15	17.080	317.712	18,60
1990/91	12.980	146.738	11,30	14.500	258.800	17,85
1991/92	13.730	195.067	14,21	16.540	293.641	17,75
1992/93	13.790	191.133	13,86	16.200	309.349	19,10
1993/94	12.020	168.955	14,06	15.920	286.061	17,97
1994/95	16.325	-	-	14.127	-	-
1995/96	16.721	190.928	11,42	14.063	238.808	16,98
1996/97	19.243	332.425	17,28	17.943	493.272	27,49
1997/98	15.396	175.007	11,37	13.418	234.599	17,48
1998/99	15.448	231.950	15,01	22.241	424.694	19,10
1999/00	15.299	190.332	12,44	17.713	378.801	21,39
2000/01	18.510	342.053	18,48	17.110	408.561	23,88
2001/02	18.030	420.347	23,31	17.930	419.319	23,39
2002/03	15.000	297.629	19,84	16.310	367.637	22,54
2003/04	16.800	302.400	18,00	17.200	390.784	22,72
2004/05	15.620	321.303	20,57	17.010	380.684	22,38
2005/06	17.980	446.084	24,81	18.700	482.834	25,82
2006/07	14.124	212.545	15,05	15.742	286.408	18,19
2007/08	14.800	220.224	14,88	16.200	338.142	20,87

2008/09	13.473	265.553	19,71	15.089	394.244	26,13
2009/10	16.756	315.519	18,83	10.439	321.712	30,82
2010/11	17.757	615.990	34,69	11.902	485.201	40,77
2011/12	10.383	272.035	26,20	13.812	508.640	36,83
2012/13	14.459	314.852	21,78	13.346	416.255	31,19
2013/14	13.054	272.045	20,84	14.765	444.884	30,13
2014/15	16.788	305.710	18,21	10.458	240.774	23,02
2015/16	14.976	338.757	22,62	13.913	424.179	30,49
2016/17	13.886	369.923	26,64	15.001	599.820	39,99
2017/18	12.486	396.541	31,76	10.067	426.324	42,35

Tabla A4.3 Datos de agua caída en mm.

	Región del Biobío		**Región de La Araucanía**	**Región de Los Ríos y Los Lagos**		
Año agrícola	**Estación O'Higgins**	**Estación Carriel Sur**	**Estación Maquehue**	**Estación Pichoy**	**Estación Cañal Bajo**	**Estación El Tepual**
1979/80	1.359,7	1.336,1	1.473,4	2.087,1	1.734,8	2.103,9
1980/81	1.256,5	2.233,8	1.439,7	2.041,8	1.432,3	1.674,8
1981/82	1.066,4	1.104,4	1.103,0	1.547,5	1.323,5	1.831,8
1982/83	1.795,3	1.329,5	1.175,7	1.939,8	1.407,4	1.853,4
1983/84	1.020,3	862,0	851,7	1.365,9	999,2	1.454,9
1984/85	1.360,7	1.334,5	1.410,3	1.966,4	1.353,0	1.578,1
1985/86	1.050,8	998,9	1.240,9	1.837,7	1.338,0	1.827,2
1986/87	1.314,8	1.177,1	1.041,1	1.698,3	1.098,9	1.449,6
1987/88	1.166,1	1.123,9	1.036,4	1.612,5	1.056,0	1.375,6
1988/89	912,2	989,2	788,2	1.115,2	758,2	1.366,3
1989/90	962,7	1.076,5	1.027,9	1.638,5	1.070,6	1.527,1
1990/91	724,9	768,0	1.039,4	1.609,3	1.170,3	1.630,5
1991/92	1.240,7	1.184,7	1.226,8	1.833,0	1.466,2	1.988,3
1992/93	1.527,6	1.393,1	1.449,4	1.982,3	1.572,9	1.913,8
1993/94	1.284,4	1.164,3	1.388,9	2.074,2	1.249,9	1.630,4
1994/95	955,1	867,2	1.274,9	1.790,7	1.471,7	2.010,6
1995/96	935,6	885,6	1.205,2	1.714,6	1.300,3	1.453,9
1996/97	817,4	829,0	826,7	1.446,8	1.211,0	1.520,3
1997/98	1.127,0	1.318,9	1.307,5	1.976,9	1.349,1	1.611,9
1998/99	457,4	627,9	664,1	1.021,5	846,3	1.130,3
1999/00	1.143,6	1.168,6	1.116,5	1.702,7	1.233,3	1.486,2
2000/01	1.056,8	1.406,9	1.344,6	1.990,8	1.354,6	1.704,5
2001/02	1.363,4	1.358,6	1.225,9	1.958,3	1.303,0	1.580,3
2002/03	999,6	1.245,7	1.261,9	2.061,2	1.519,2	1.907,6
2003/04	837,0	1.025,0	1.163,0	1.981,6	1.339,7	1.561,0
2004/05	811,9	970,4	1.112,7	1.636,5	1.207,2	1.482,1

2005/06	1.289,4	1.511,5	1.659,9	2.292,8	1.501,6	1.983,2
2006/07	1.221,4	1.356,1	1.276,6	1.873,1	1.484,1	1.686,0
2007/08	608,7	639,6	906,0	1.176,5	801,4	1.274,5
2008/09	941,5	1.114,3	1.066,7	1.983,5	1.069,2	1.643,3
2009/10	951,5	954,1	1.254,3	1.789,4	1.383,2	1.591,1
2010/11	764,0	842,8	976,4	1.630,7	1.191,8	1.412,9
2011/12	818,5	686,6	942,0	1.561,6	852,9	1.549,0
2012/13	786,4	713,8	919,9	1.748,6	1.252,1	1.581,1
2013/14	832,5	707,0	888,1	1.509,6	1.009,0	1.615,6
2014/15	845,8	867,2	938,5	1.658,3	1.279,0	1.281,4
2015/16	1.035,2	767,6	1.228,7	2.004,3	1.253,4	1.465,4
2016/17	497,2	657,4	945,2	1.342,0	941,6	1.132,6
2017/18	903,7	957,4	1.262,9	1.878,3	1.420,1	1.760,4

Tabla A4.4 Producción (tn) para periodo 2016/2017 y 2017/2018.

	Superficie Sembrada con trigo (Biobío)		**Superficie Sembrada con trigo (Araucanía)**		**Superficie sembrada con papa (Araucanía)**		**Superficie sembrada con papa (Los Ríos y Los Lagos)**	
	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18
Producción	376.305	427.752	597.835	663.517	369.923	396.541	599.820	426.324

Tabla A0.5 Pronóstico de producción (tn) para periodo 2016/2017 y 2017/2018.

	Producción de trigo (Biobío)		**Producción trigo (Araucanía)**		**Producción de papa (Araucanía)**		**Producción de papa (Los Ríos y Los Lagos)**	
	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18
Regresión lineal	569.991	581.222	641.475	672.148	307.369	311.939	372.205	426.602
Red neuronal	619.044	578.601	430.669	509.244	207.664	356.212	385.913	537.877
SVR	589.787	554.878	663.136	679.557	322.867	299.740	374.742	457.010

Tabla A0.6 Error en pronóstico de producción (tn) para periodo 2016/2017 y 2017/2018.

	Producción de trigo (Biobío)		**Producción trigo (Araucanía)**		**Producción de papa (Araucanía)**		**Producción de papa (Los Ríos y Los Lagos)**	
	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18
Regresión lineal	33,98%	26,40%	6,80%	1,28%	20,35%	27,12%	61,15%	0,06%
Red neuronal	39,21%	26,07%	38,81%	30,2%	78,13%	11,32%	55,42%	20,73%
SVR	36,19%	22,91%	9,84%	2,36%	14,57%	32,29%	60,06%	6,71%

Tabla A0.7 Pronóstico de producción (tn) (en función de la superficie sembrada) para periodo 2016/2017 y 2017/2018.

	Producción de trigo (Biobío)		**Producción trigo (Araucanía)**		**Producción de papa (Araucanía)**		**Producción de papa (Los Ríos y Los Lagos)**	
	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18
Regresión lineal	529.174	529.814	589.011	589.277	336.547	371.621	415.550	375.846
Red neuronal	644.440	623.101	513.603	614.060	0	317.869	317.021	481.127
SVR	588.236	558.840	604.131	637.134	334.529	360.344	330.437	415.079

Tabla A0.8 Error en pronóstico de producción (tn) (en función de la superficie sembrada) para periodo 2016/2017 y 2017/2018.

	Producción de trigo (Biobío)		**Producción trigo (Araucanía)**		**Producción de papa (Araucanía)**		**Producción de papa (Los Ríos y Los Lagos)**	
	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18	2016/17	2017/18
Regresión lineal	28,88%	19,26%	1,49%	12,59%	9,81%	6,70%	44,34%	13,43%
Red neuronal	41,60%	31,35%	16,40%	8,05%	100%	24,74%	89,20%	11,39%
SVR	36,02%	23,45%	1,04%	4,14%	10,58%	10,04%	81,52%	2,70%

B: Figuras

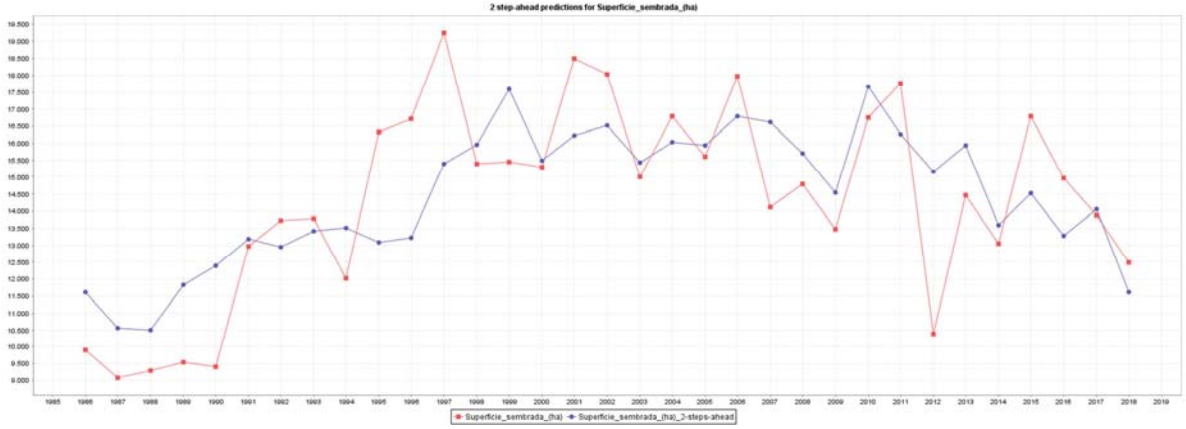

Figura B4.1 Regresión lineal para siembra de papa en La Araucanía (en función del tiempo y agua caída).

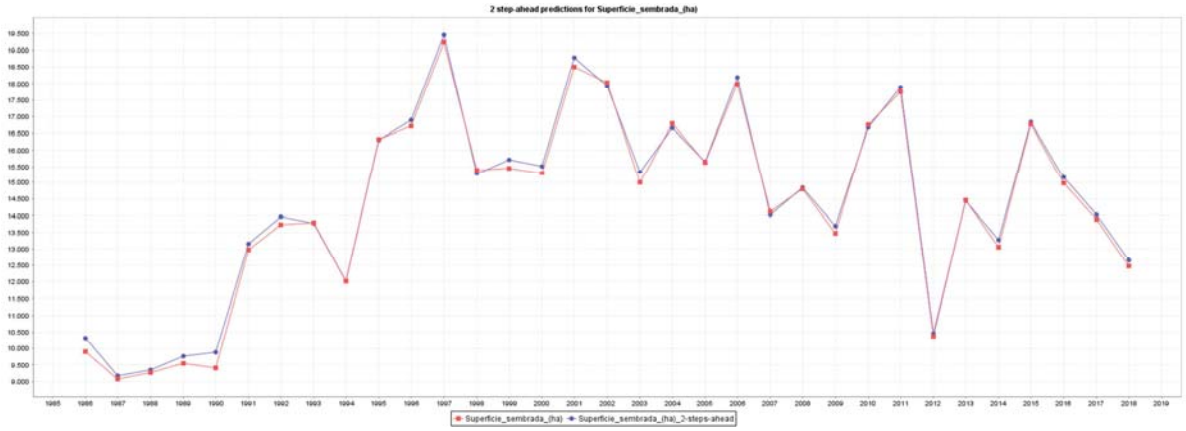

Figura B4.2 Red neuronal para siembra de papa en La Araucanía (en función del tiempo y agua caída).

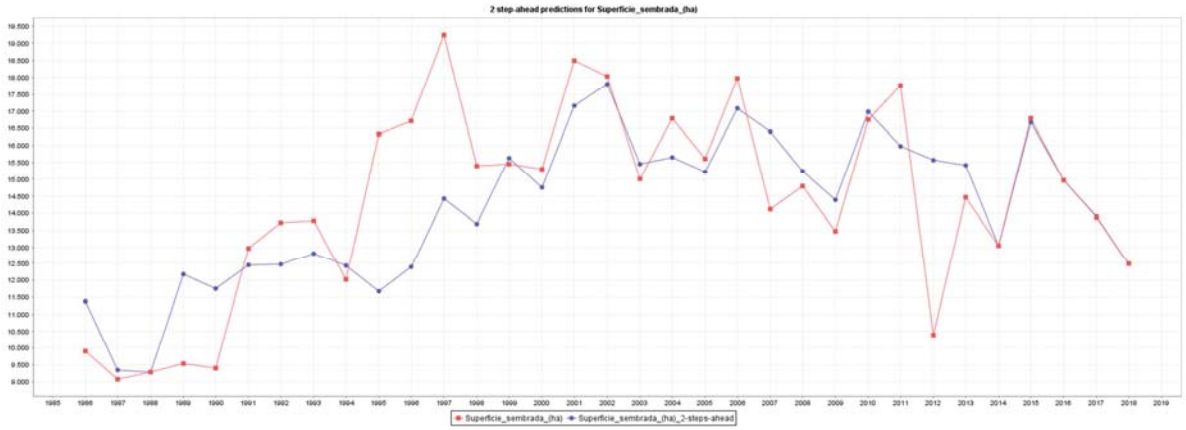

Figura B4.3 SVR para siembra de papa en La Araucanía (en función del tiempo y agua caída).

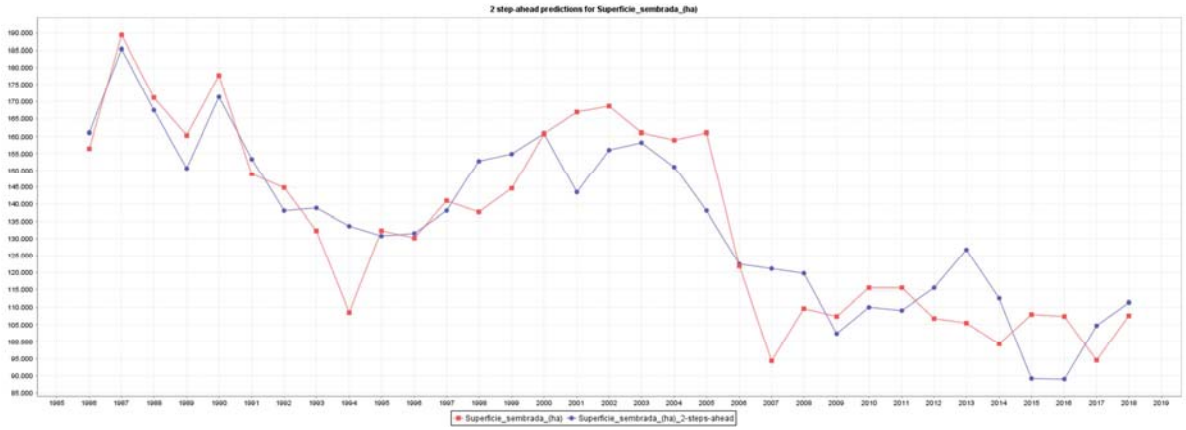

Figura B4.4 Regresión lineal para siembra de trigo en La Araucanía (en función del tiempo y agua caída).

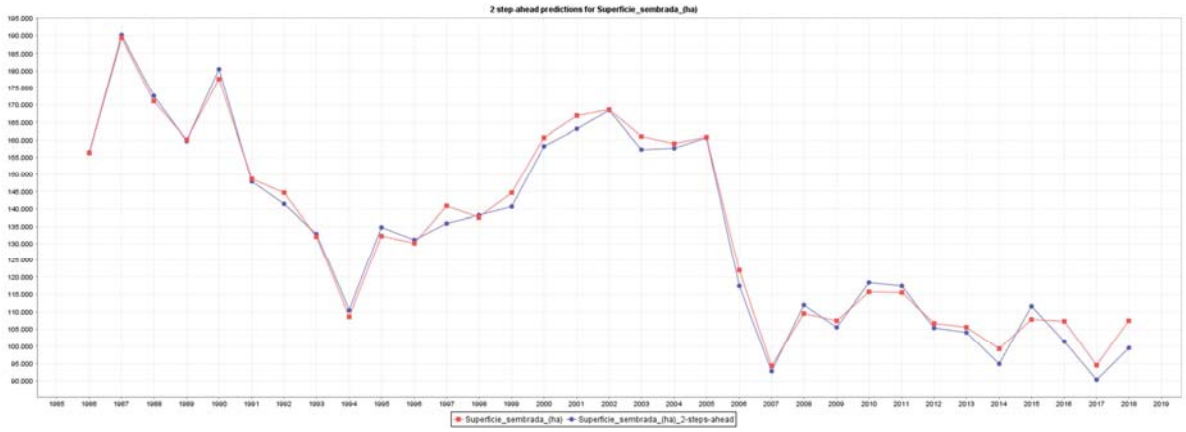

Figura B4.5 Red neuronal para siembra de trigo en La Araucanía (en función del tiempo y agua caída).

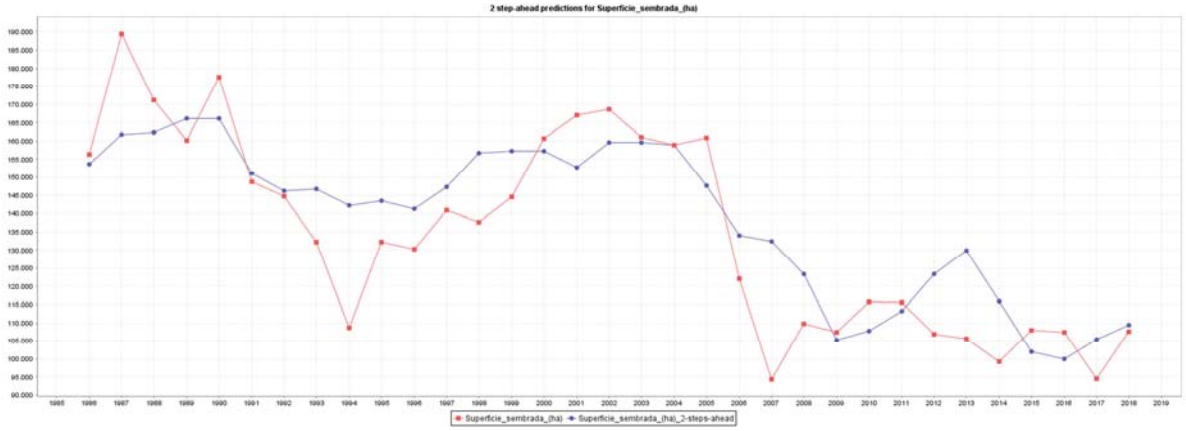

Figura B4.6 SVR para siembra de trigo en La Araucanía (en función del tiempo y agua caída).

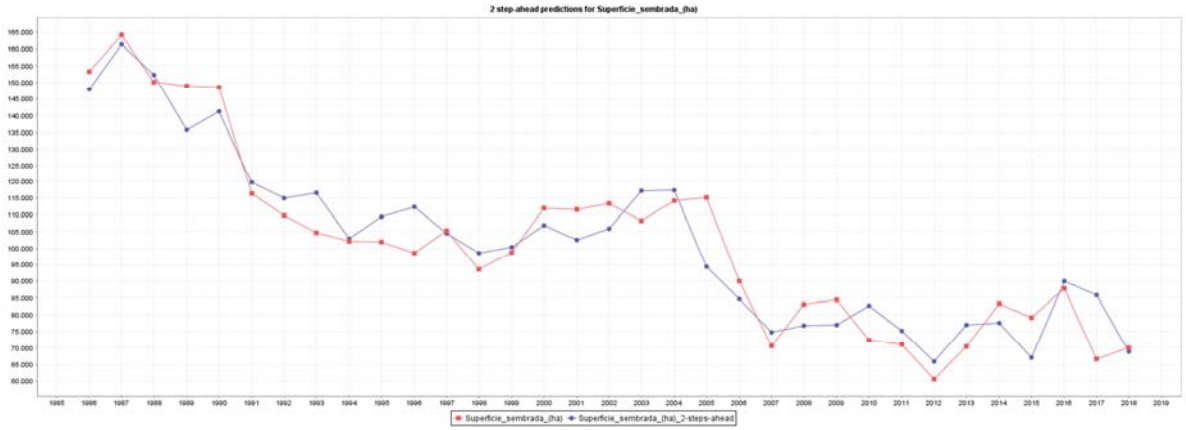

Figura B4.7 Regresión lineal para siembra de trigo en Biobío (en función del tiempo y agua caída).

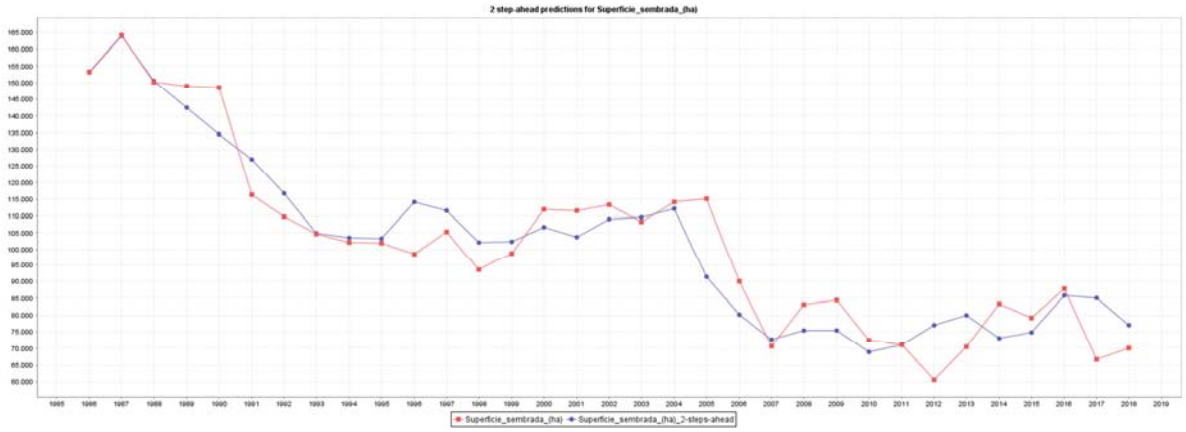

Figura B4.8 SVR para siembra de trigo en Biobío (en función del tiempo y agua caída).

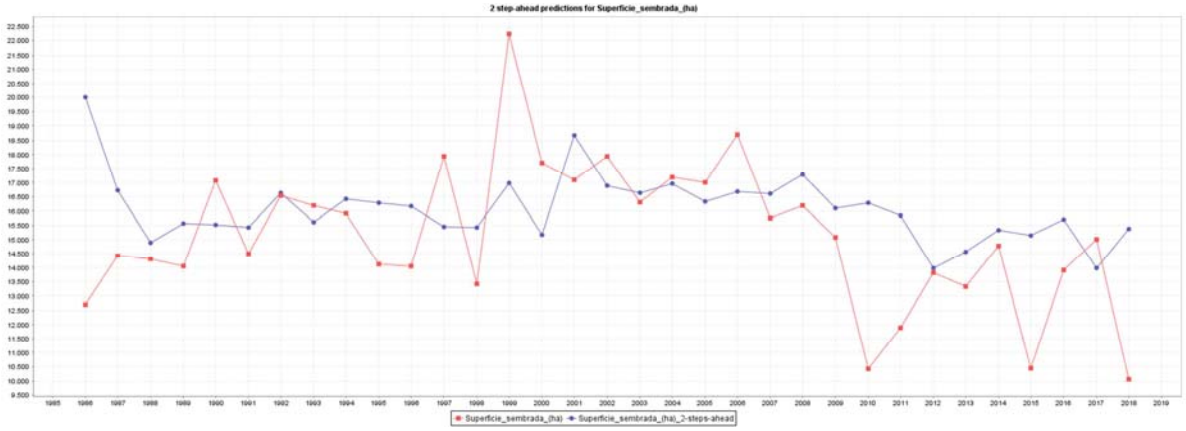

Figura B4.9 Regresión lineal para siembra de papa en Los Ríos y Los Lagos (en función del tiempo y agua caída)

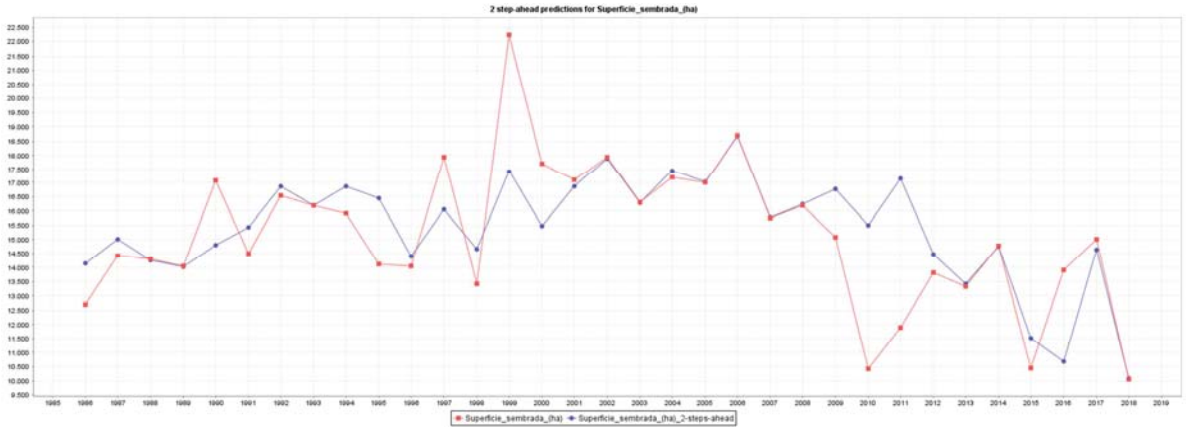

Figura B4.10 SVR para siembra de papa en Los Ríos y Los Lagos (en función del tiempo y agua caída).

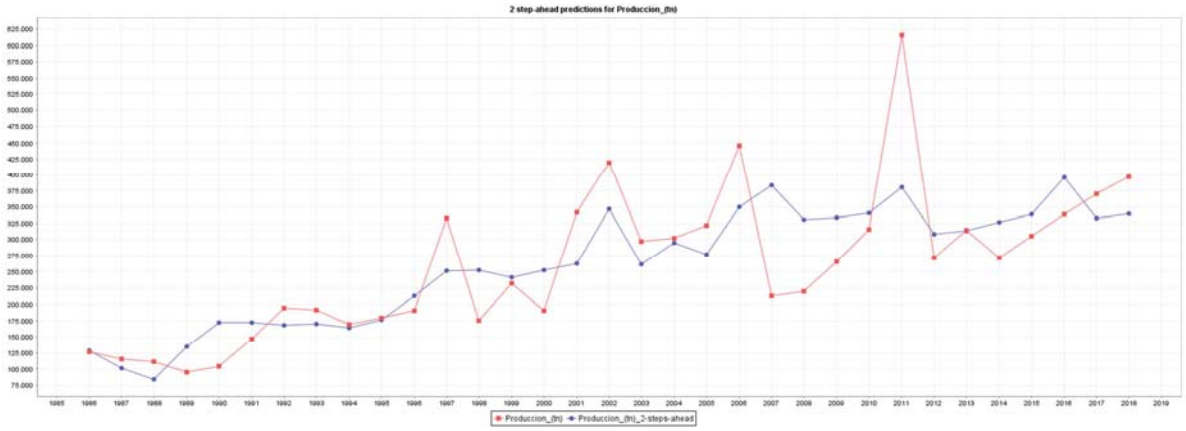

Figura B4.11 Regresión lineal para producción de papa en La Araucanía (en función del tiempo y agua caída).

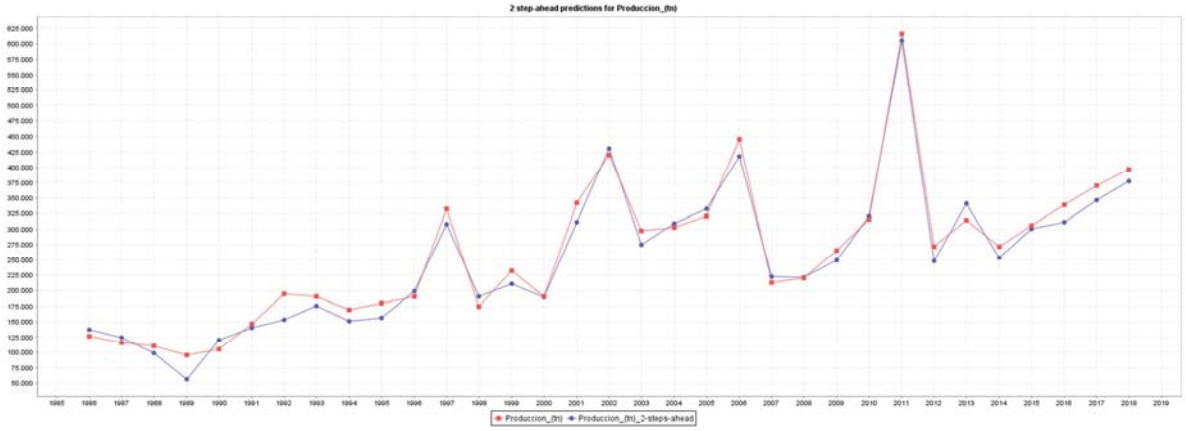

Figura B4.12 Red neuronal para producción de papa en La Araucanía (en función del tiempo y agua caída).

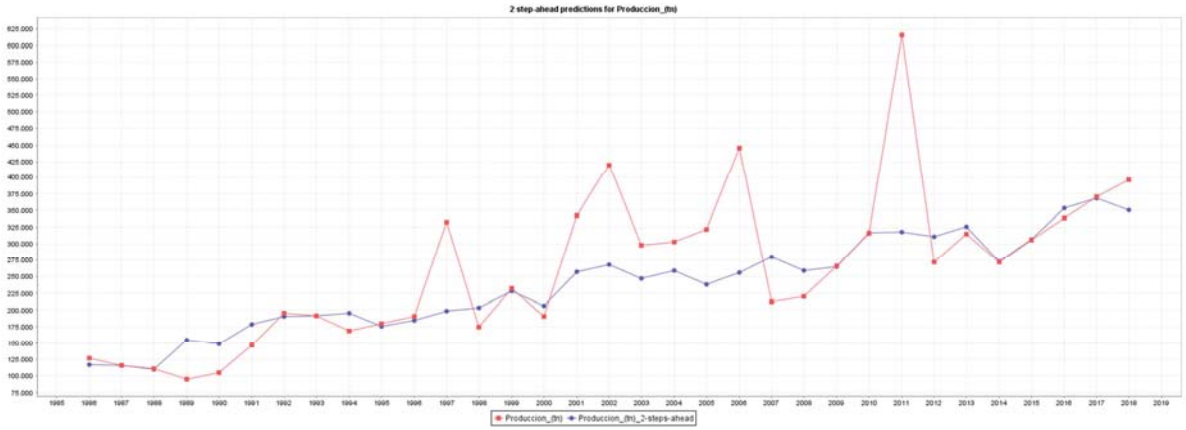

Figura B4.13 SVR para producción de papa en La Araucanía (en función del tiempo y agua caída).

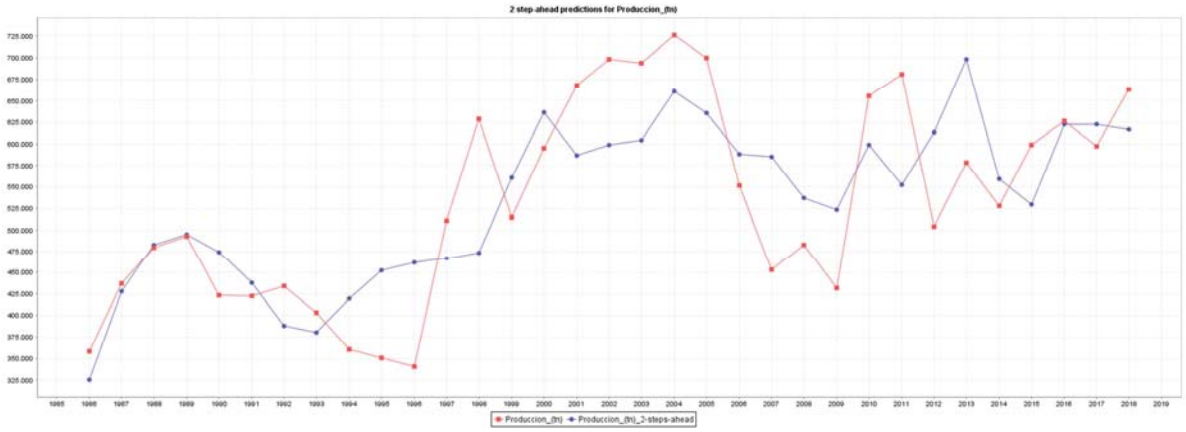

Figura B4.14 Regresión lineal para producción de trigo en La Araucanía (en función del tiempo y agua caída).

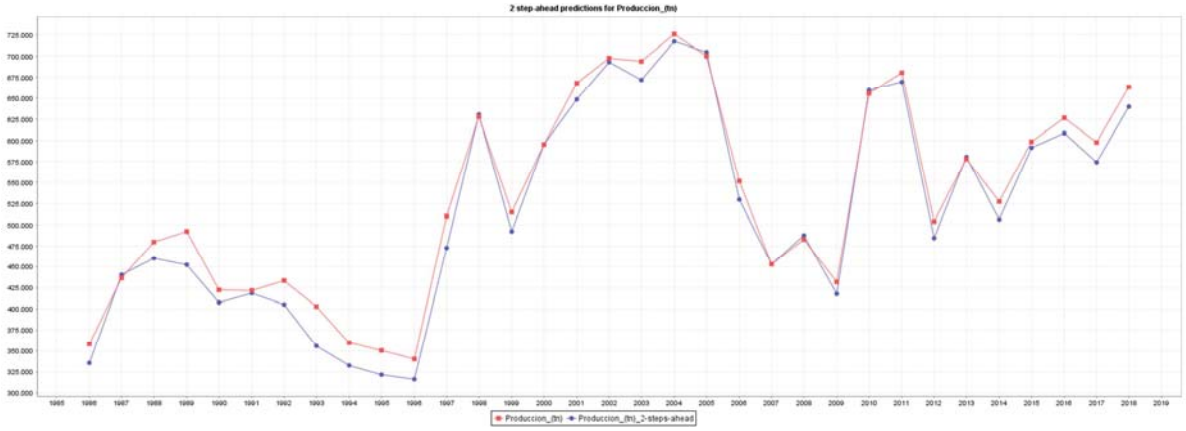

Figura B4.15 Red neuronal para producción de trigo en La Araucanía (en función del tiempo y agua caída).

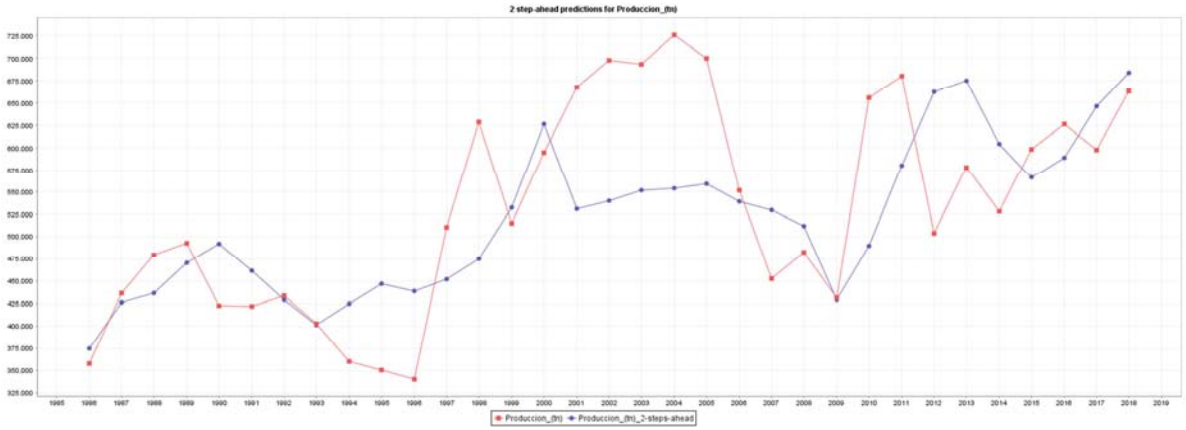

Figura B4.16 SVR para producción de trigo en La Araucanía (en función del tiempo y agua caída).

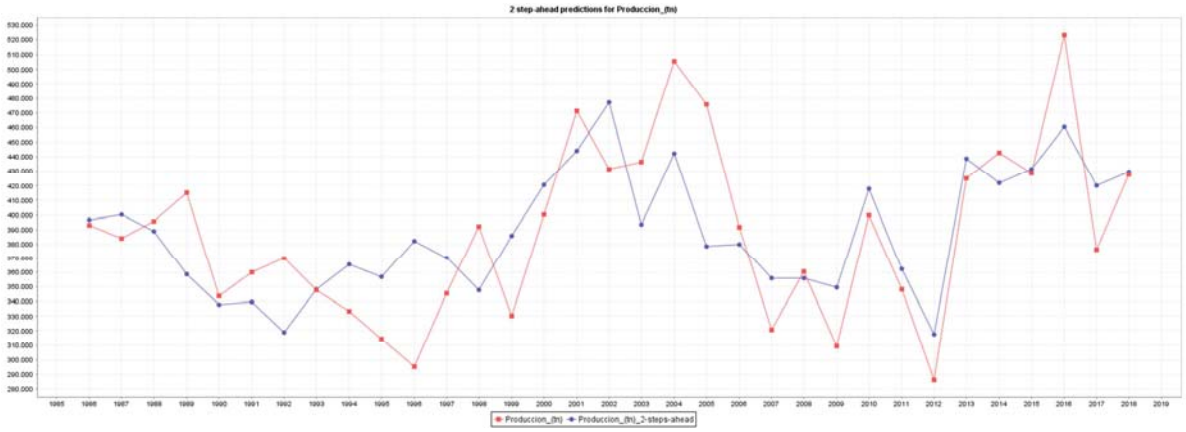

Figura B4.17 Regresión lineal para producción de trigo en Biobío (en función del tiempo y agua caída).

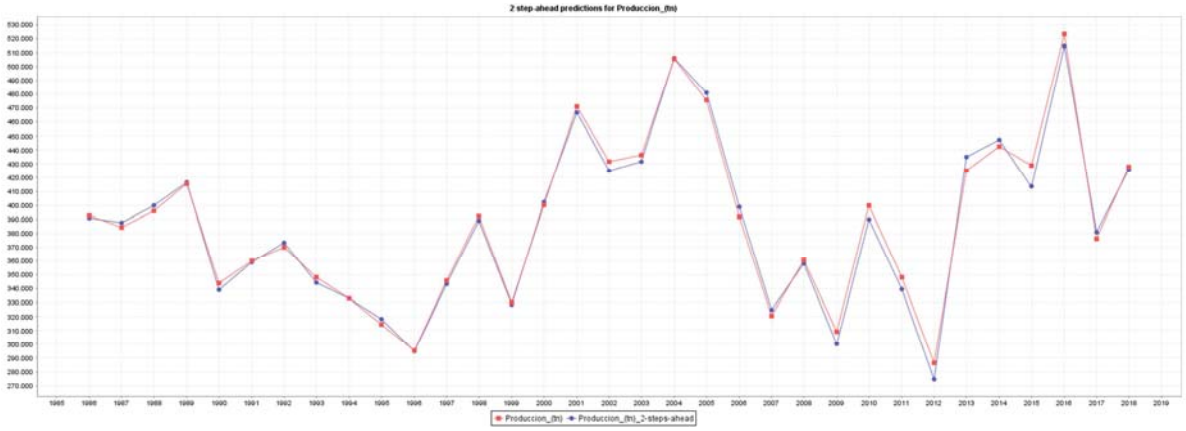

Figura B4.18 Red neuronal para producción de trigo en Biobío (en función del tiempo y agua caída).

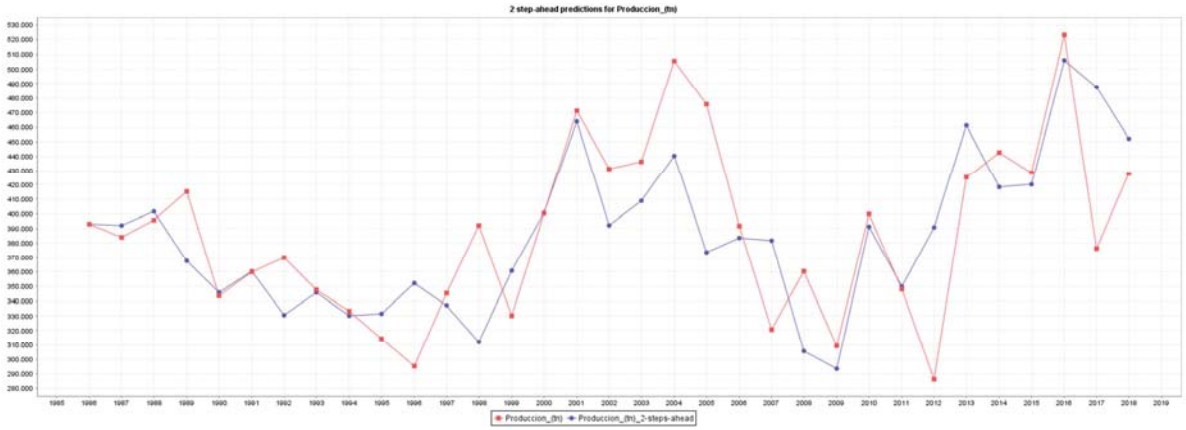

Figura B4.19 SVR para producción de trigo en Biobío (en función del tiempo y agua caída).

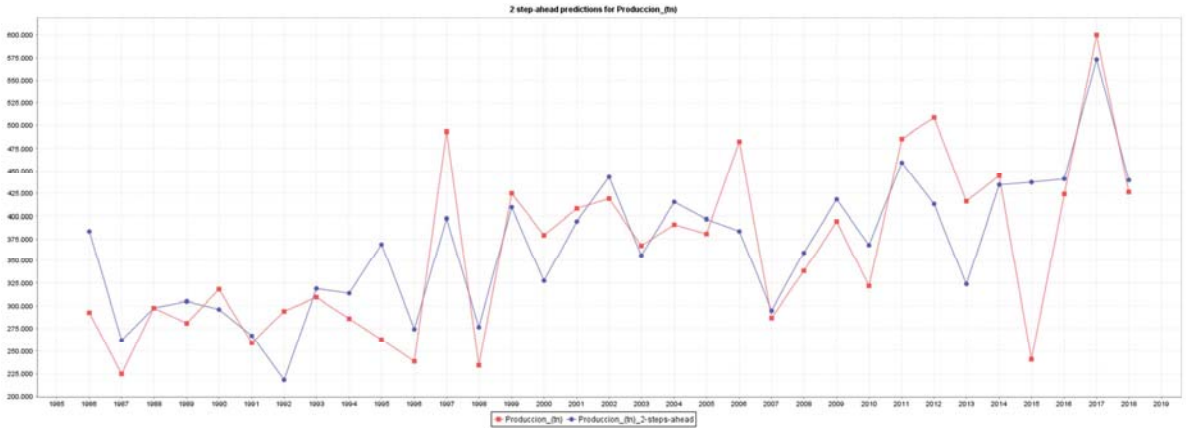

Figura B4.20 Regresión lineal para producción de papa en Los Ríos y Los Lagos (en función del tiempo y agua caída).

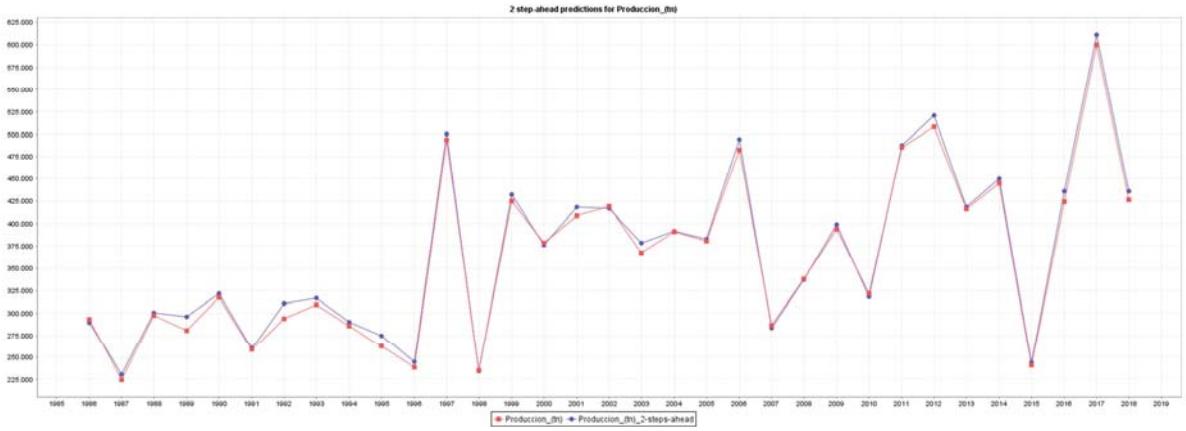

Figura B4.21 Red neuronal para producción de papa en Los Ríos y Los Lagos (en función del tiempo y agua caída).

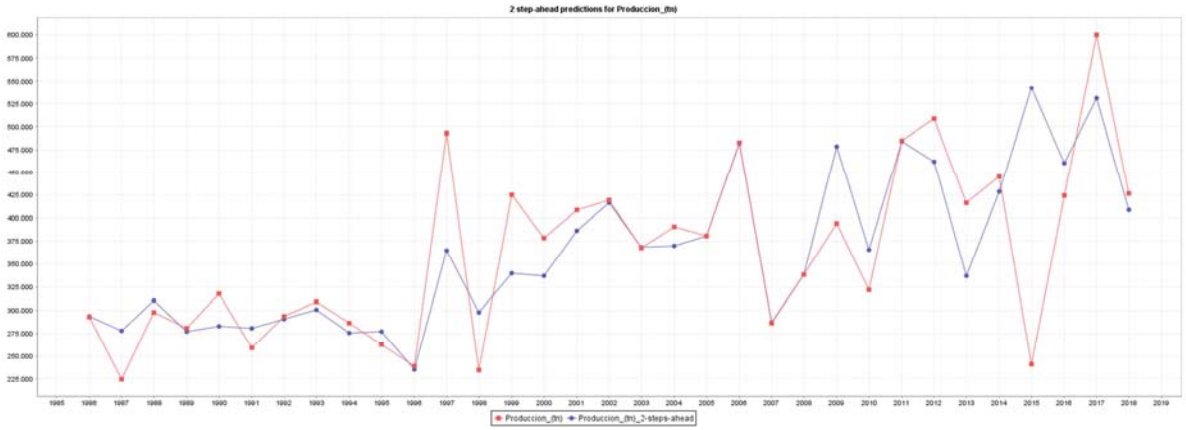

Figura B4.22 SVR para producción de papa en Los Ríos y Los Lagos (en función del tiempo y agua caída).

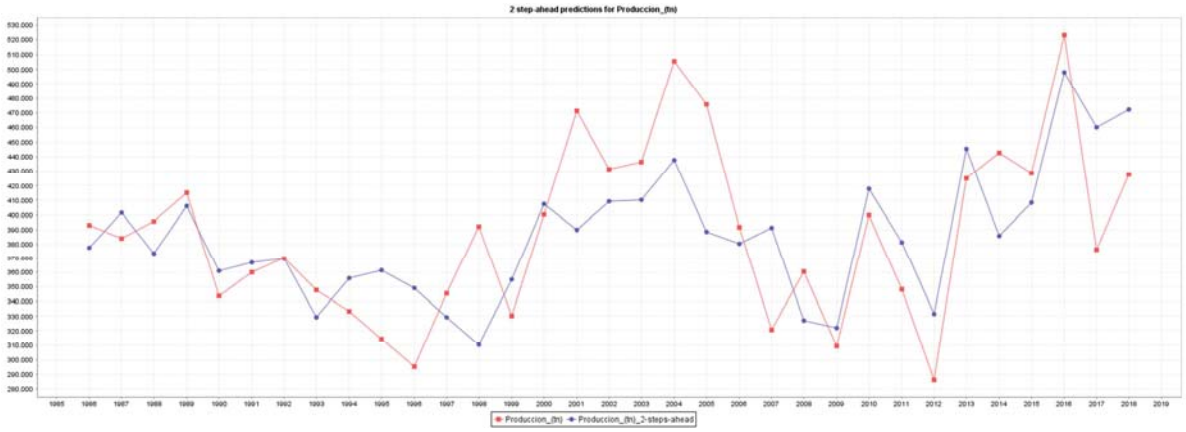

Figura B4.23 Regresión lineal para producción de trigo en Biobío (en función del tiempo, agua caída y superficie sembrada).

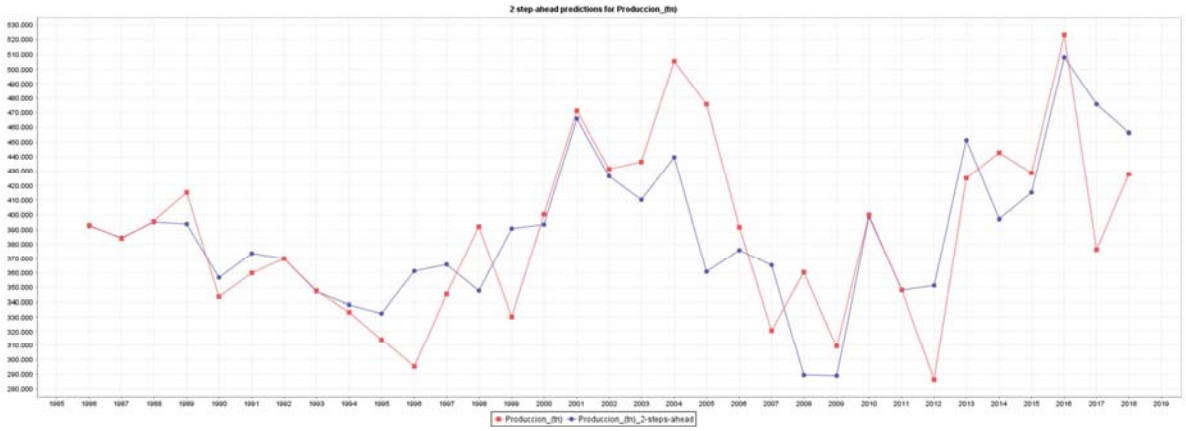

Figura B4.24 SVR para producción de trigo en Biobío (en función del tiempo, agua caída y superficie sembrada).

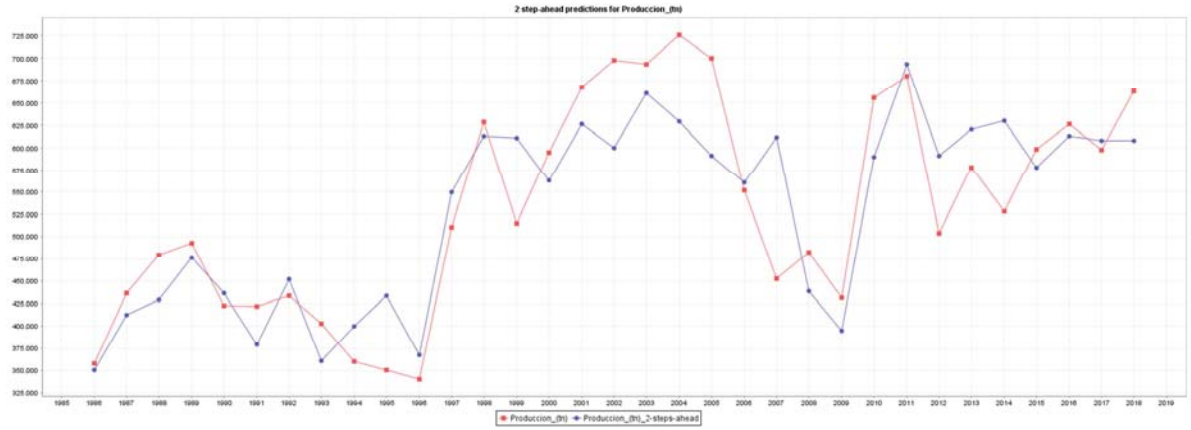

Figura B4.25 Regresión lineal para producción de trigo en La Araucanía (en función del tiempo, agua caída y superficie sembrada).

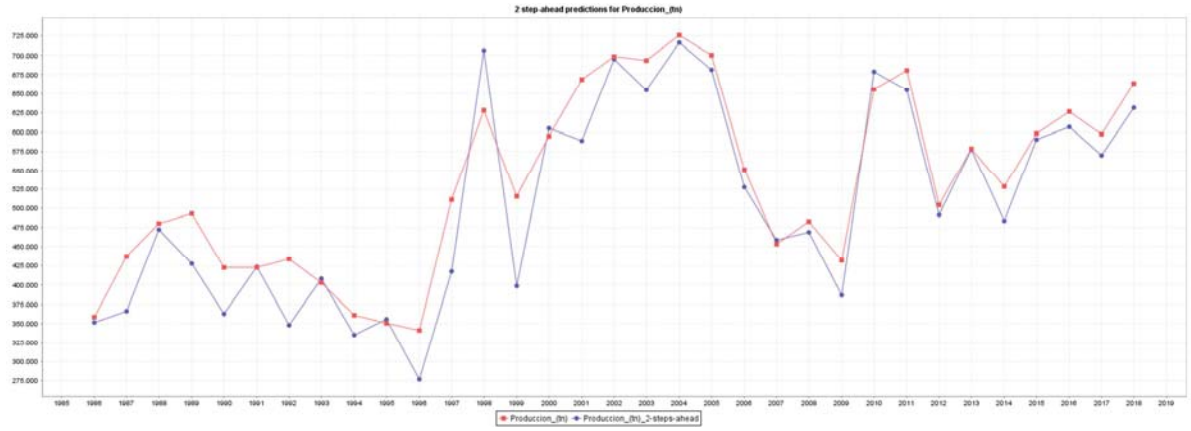

Figura B4.26 Red neuronal para producción de trigo en La Araucanía (en función del tiempo, agua caída y superficie sembrada).

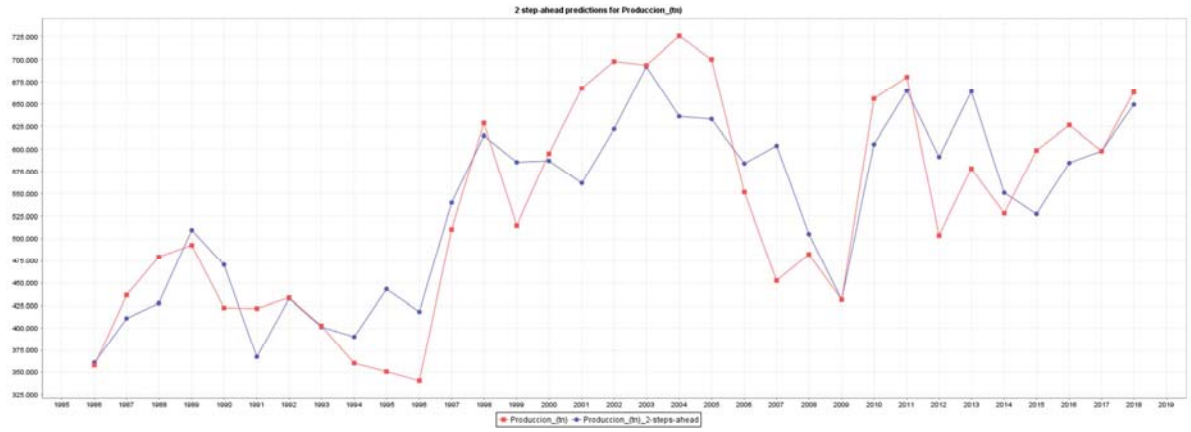

Figura B4.27 SVR para producción de trigo en La Araucanía (en función del tiempo, agua caída y superficie sembrada).

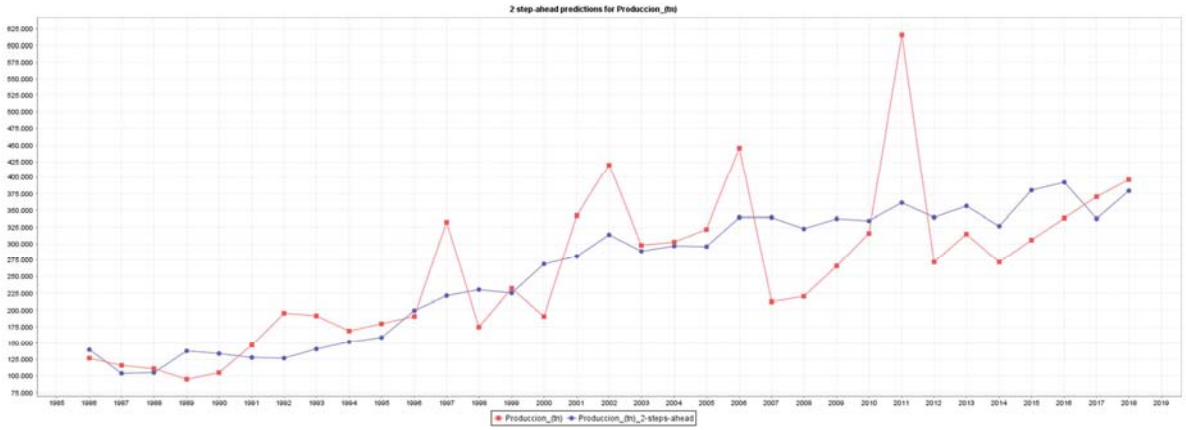

Figura B4.28 Regresión lineal para producción de papa en La Araucanía (en función del tiempo, agua caída y superficie sembrada).

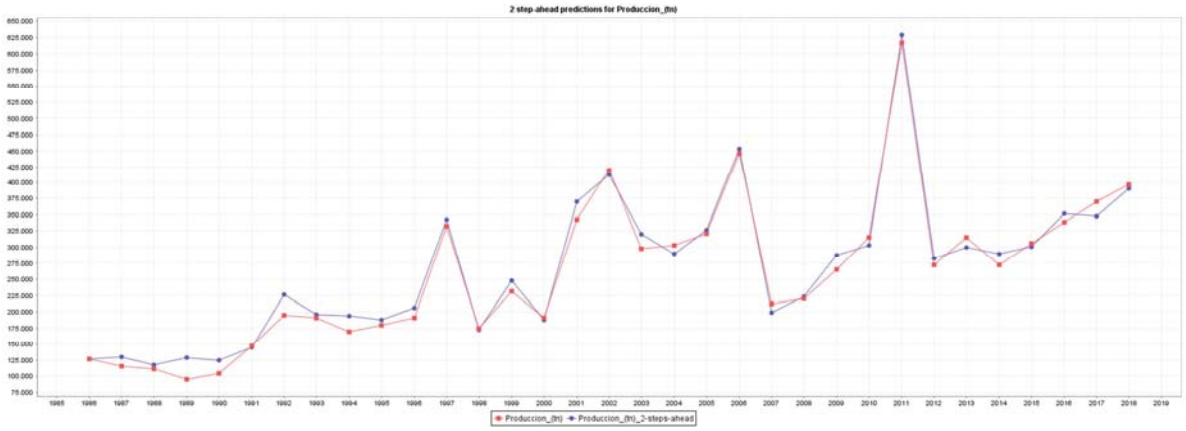

Figura B4.29 Red neuronal para producción de papa en La Araucanía (en función del tiempo, agua caída y superficie sembrada).

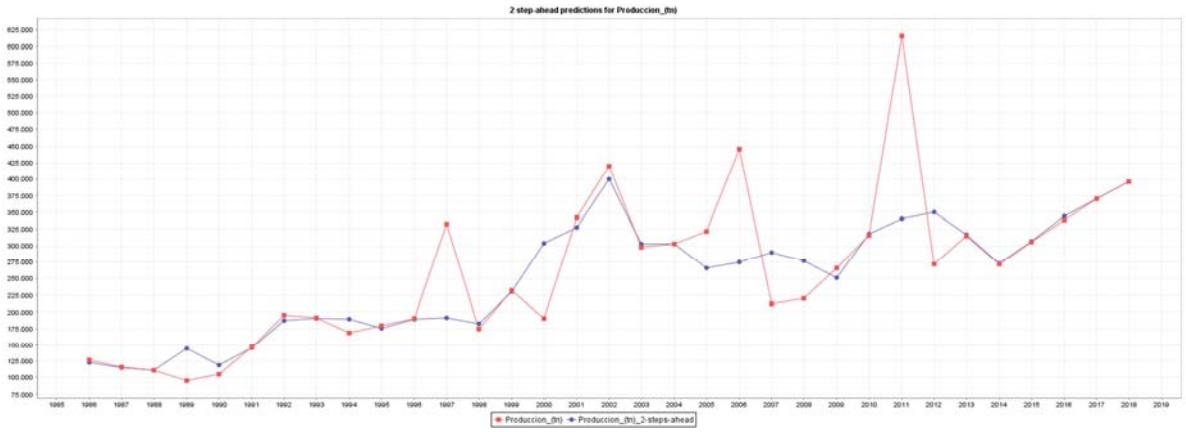

Figura B4.30 SVR para producción de papa en La Araucanía (en función del tiempo, agua caída y superficie sembrada).

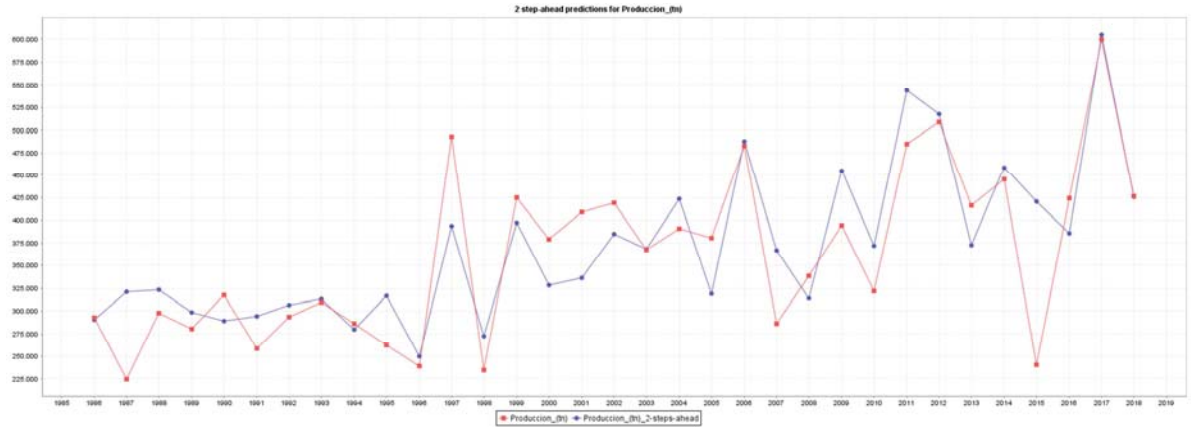

Figura B4.31 Regresión lineal para producción de papa en Los Ríos y Los Lagos (en función del tiempo, agua caída y superficie sembrada).

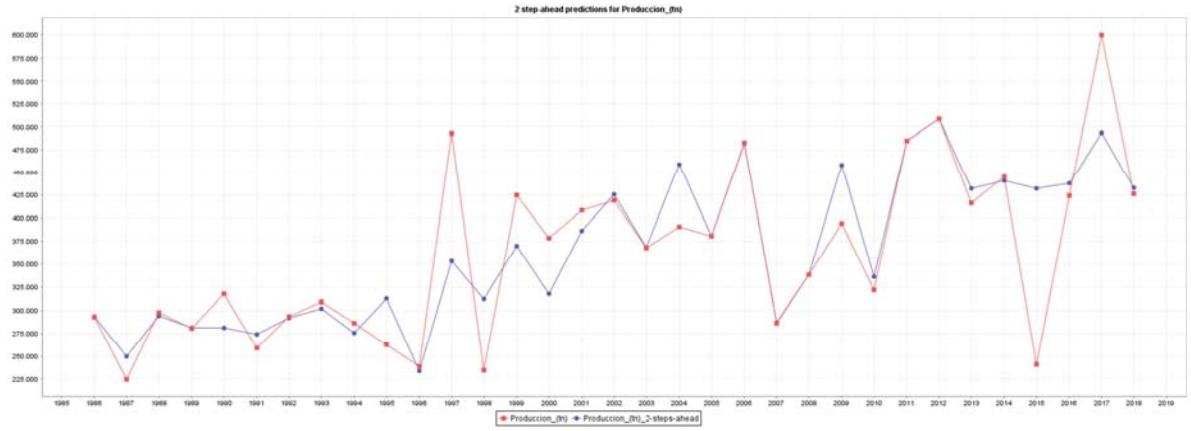

Figura B4.32 SVR para producción de papa en Los Ríos y Los Lagos (en función del tiempo, agua caída y superficie sembrada).