Universidad de las Ciencias Informáticas

Facultad 5

Modelo de técnicas de minería de datos aplicado a los datos de cultivos del sistema DSerp Agro.

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas.

Autor:

Jairo Alejandro Lefebre Lobaina

Tutor:

Msc. Marvyn Armando Márquez Rodríguez

Co-tutores:

Dr. Ramón Carrasco Velar

Ing. Aliander Capdezuñer González.

Ciudad de La Habana

Junio del 2012

DECLARACIÓN DE AUTORÍA:

Declaramos ser autores de este trabajo y reconocemos la Universidad de las Ciencias Informáticas los derechos patrimoniales del mismo con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____	_____
Jairo Alejandro Lefebre Lobaina.	Msc. Marvyn Armando Márquez Rodríguez
(Autor)	**(Tutor)**

AGRADECIMIENTOS:

- Mi mayor agradecimiento es para el Universo, que me ha permitido llegar a ser lo que hoy soy y a la ciencia que me aportó todos los conocimientos necesarios para llevar este trabajo a su fin.
- A los Msc. Marvyn Armando Márquez Rodríguez e Ing. Aliander Capdezuñer González por su acertada dirección y ayuda.
- Al Dr. Ramón Carrasco Velar por su valioso caudal de conocimientos y permanente asesoría puestos a disposición de la tesis.
- Al Msc. Juan Matos Borges por proporcionar la información necesaria para desarrollar este trabajo.
- Al profesor Tomas López Jiménez por tanta confianza depositada en nosotros y por convertirse en mi guía como profesional y ser humano.
- A Alberto Pacheco que tanto nos ayudó a mejorar la *falta de nivel.*
- A mis padres y abuela que durante toda la carrera me apoyaron e incentivaron.
- A Enma Farin Levys por su permanente cooperación y ayuda.
- A mi novia Aylin M. Rodríguez Jiménez por su incondicional apoyo y amor.
- A mis amigos de tantas batallas Karel Piorno, Luanner Kerton y Adrián Hernández por tantos momentos compartidos y explotar como Kafunga.
- A Lili, Rosa, José Miguel, David, Murgado, Hebert, Reiner, Dariel, Masó, Simón, Luís, Boris, Jorge Carlos, Martha, Kenia, Lisandra y Ernesto por su amistad.
- A mis tíos, tías, primos y el resto de mis familiares por haber confiado en mí.
- A los que de una forma u otra colaboraron para que este trabajo se realizara.

DEDICATORIA:

Dedico con todo mi amor este trabajo a mi madre y a todas las personas que me aman.

Resumen

El software DSerp Agro en un Sistema de Administración de Negocios Empresariales (EBMS), diseñado para gestionar los procesos empresariales relacionados con la esfera de la agricultura, se utiliza como alternativa para gestionar la información generada por las unidades de producción agropecuarias, almacena y preserva entre otros datos los relacionados con el proceso de siembra. El software almacena grandes volúmenes de información, pero no cuenta con las herramientas necesarias para extraer la información más relevante, ni realizar estudios predictivos de comportamiento, imprescindibles para la toma de decisiones por parte de la Alta Dirección.

Es propósito de este trabajo, desarrollar un modelo matemático que realice el proceso de minería de datos a la base de datos del sistema Dserp Agro, con el objetivo de extraer la información no trivial y potencialmente útil intrínseca en ella, y predecir dentro del sistema, el comportamiento de un cultivo específico, la especie Allium Cepa (cebolla).

Los experimentos realizados a partir de la muestra obtenida, permitieron predecir el comportamiento de las variables dependientes del problema con un nivel certeza superior al 90%, mediante la utilización de modelos empíricos.

.

Índice:

Introducción

El surgimiento de las Tecnologías de la Información y las Comunicaciones (TICs) revolucionó el ámbito de la sociedad en todos sus aspectos, de manera particular en el empresarial, los cambios producidos en la conformación económica de la sociedad, unidos a los avances tecnológicos han tenido una influencia significativa. Las posibilidades que brinda el empleo de las computadoras y sus aplicaciones informáticas para transformar, almacenar, gestionar, proteger, difundir y localizar los datos necesarios para cualquier actividad, ha producido cambios importantes en la manera de organizar y gestionar la actividad productiva en las Empresas, su aplicación marca diferencia en el desempeño de las organizaciones llevándolas a un status superior de desarrollo.

Las empresas manejan en su desempeño una gran cantidad de información, tanto interna como externa, que fluye y se asimila en las diferentes áreas de trabajo, lo cual constituye información valiosa que no está estructurada para aportar datos al desarrollo y mejoramiento de las entidades. El empleo de la inteligencia empresarial se transforma entonces, en un requerimiento de primer orden para el empresariado cubano.

La inteligencia empresarial constituye una herramienta gerencial, cuya función es facilitar a las administraciones el cumplimiento de los objetivos y la misión de sus organizaciones, mediante el análisis de la información relativa a su negocio y al entorno en que las mismas se desenvuelven. Desde el punto de vista operacional, es un conjunto de metodologías, procedimientos y herramientas para: la obtención, el procesamiento, el análisis y la diseminación de información, de modo que se facilite la orientación táctica, la toma de decisiones estratégicas y el desempeño de las organizaciones (Orozco, 2009).

La inteligencia empresarial permite evaluar potencialmente las capacidades de la organización en cuanto a la producción, la mercadotecnia, el estado financiero, los recursos materiales y humanos generando escenarios, reportes y pronósticos que ayudan a la toma de decisiones (Orozco, 2009).

El análisis de información para la inteligencia empresarial puede definirse como el conjunto de operaciones y procedimientos que permite extraer aquellas informaciones principales, relevantes y útiles para la toma de decisiones a partir de los datos disponibles, seleccionados de las fuentes adecuadas de información, a partir de la eficiente captura de datos y su estructuración de manera conveniente, para aplicar las técnicas métricas al estudio de las variables seleccionadas. Requiere de la utilización de herramientas informáticas que permitan automatizar gran parte del proceso, extraer

conocimiento de grandes volúmenes de información y aplicar métodos avanzados de análisis (Orozco, 2009).

Una de las herramientas utilizadas para realizar el análisis de grandes volúmenes de datos es la minería de datos (MD). Surge como una necesidad a partir del crecimiento en el volumen y variedad de la información que se encuentra almacenada en bases de datos (Han, et al., 2006), su utilización se ha incrementado exponencialmente en los últimos años.

La MD tiene el mérito que en ella concurren varias áreas del conocimiento como la Inteligencia Artificial y la Estadística, es un proceso automatizado que no se limita al análisis de sucesos pasados, si no que se encarga de extraer la información no trivial y potencialmente útil intrínseca en los datos, con el objetivo de transformarla en conocimiento, siendo capaz de realizar asociaciones y agrupamientos, predecir comportamientos y definir secuencias (J. Hernandez Orallo, 2004). Se convierte en una herramienta muy ventajosa en los procesos industriales y de investigación científica, como soporte al diseño de bases de datos, ingeniería inversa, mejora la calidad de los datos, mejora las consultas a bases de datos y para realizar pronósticos sobre el comportamiento de los datos, es muy importante como instrumento para la toma de decisiones (Orozco, 2009).

En el mundo, las aplicaciones de gestión de información son usadas cada vez con mayor frecuencia en diversas áreas dentro de las organizaciones, convirtiéndose en una herramienta fundamental dentro de la empresa, y se les exige como requisito que los mismos tengan implementado subsistemas de MD, con el objetivo de emplearlos en la inteligencia empresarial, ya que representa una poderosa herramienta para enfrentar a la competencia en el ámbito empresarial (Orozco, 2009).

En Cuba existen algunos sistemas de gestión de información, entre ellos se encuentran Versat Sarasola, Windamer, Conet y el Rodas XXI, que son empleados en múltiples sectores. En la UCI se desarrollan diversas aplicaciones de gestión de datos como el CEDRUX y el DSerp Agro.

La agricultura es uno de los sectores del país que menos progreso ha alcanzado en los últimos años, esta situación se ha visto agravada por la crisis económica y alimentaria mundial, así como el bloque económico impuesto, exigiendo retos superiores. Pero contradictoriamente, no cuenta con el desarrollo tecnológico requerido para hacer más eficiente su gestión empresarial.

Partiendo de esta realidad, estudiantes de la UCI se dieron a la tarea de desarrollar un sistema de gestión, con el propósito de salvar la brecha y poner a la agricultura a la altura que exigen las circunstancias actuales, en cuanto a la forma de almacenar y gestionar la información, es así como surge el DSerp Agro.

DSerp Agro es un sistema enfocado fundamentalmente a la automatización de procesos agrícolas, que almacena una gran cantidad de información referente al comportamiento de los cultivos, dentro de ellos se destaca el cultivo de la Allium Cepa (cebolla). Los especialistas en inteligencia empresarial que hacen uso de este software, para realizar el proceso de toma de decisiones sobre este cultivo, se basan fundamentalmente en la información histórica de la empresa almacenada por este sistema, y no disponen de los medios de análisis para grandes volúmenes de datos y la extracción de la información relevante. Sería provechoso crear un instrumento que permita realizar el proceso de MD a los datos de la cebolla almacenados por el sistema.

De lo anteriormente planteado se derivó como **problema científico** de este trabajo, ¿Cómo predecir el comportamiento de los datos del cultivo de cebolla en el sistema DSerp Agro para viabilizar el proceso de toma de decisiones?

Este trabajo tiene propuesto como **objetivo general** realizar el proceso de minería de datos a los datos del cultivo de cebolla almacenados por el sistema DSerp Agro, como **objeto de estudio** las tecnologías de información en la inteligencia empresarial agrícola y como **campo de acción** la minería de datos aplicada a la toma de decisiones en el sistema DSerp Agro.

Para cumplir los objetivos planteados se definen las siguientes tareas a desarrollar.

- ✓ Elaborar el marco teórico a partir del estudio de las bibliografías existentes referentes al tema.
- ✓ Describir y seleccionar las variables que definen el comportamiento de la cebolla como miembro representativo de la familia de las Liliaceae.
- ✓ Seleccionar los modelos matemáticos predictivos que arrojen los resultados más precisos.
- ✓ Ajustar los modelos seleccionados al problema planteado.
- ✓ Comparar los resultados obtenidos de los modelos.
- ✓ Probar el modelo mediante un prototipo de aplicación.

Para todo el proceso de investigación y desarrollo de este trabajo se tomó en cuenta la utilización de los siguientes métodos científicos.

Métodos Teóricos:

Analítico-Sintético: Se usa este método para definir las características distintivas de los sistemas de MD y los algoritmos que los componen, usando como apoyo la bibliografía y los estudios realizados por los especialistas en el tema para obtener la mayor información posible respecto al problema.

Histórico-Lógico: Se hace uso de este método para revelar y analizar la evolución de los sistemas de MD y su aplicación en las diversas esferas de la economía haciendo énfasis en la inteligencia empresarial.

Inductivo-Deductivo: Se arriba a un conocimiento o vía de solución al problema planteado, teniendo como punto de partida la información que se ha encontrado.

Métodos Empíricos:

Entrevista: Se utilizó para poder obtener una información más específica sobre la manera de adaptar los algoritmos de inteligencia artificial utilizados al problema planteado, con el apoyo de profesores que han trabajado el tema.

Experimento: Este método se empleó para aplicar distintas técnicas y establecer comparaciones hasta lograr los resultados esperados.

1. Capítulo 1: Fundamentación teórica

1.1. Historia y definición

Las áreas del conocimiento, como la inteligencia artificial y la estadística, que conforman el núcleo de la minería de datos, han estado en desarrollo durante décadas, pero el surgimiento de las técnicas de minería de datos como tal, es el resultado de un largo proceso evolutivo, el cual inició cuando se comenzó a almacenar la información de organizaciones en computadoras, y continuó con la evolución de las tecnologías de acceso a la información de forma remota y, más recientemente, en tiempo real (Han, et al., 2006).

Aunque el diseño de algunos de los algoritmos utilizados hoy en día fue desarrollado en la segunda mitad del siglo pasado, la tecnología de ese momento no permitía su implementación de forma tal que pudieran ser aplicados con resultados satisfactorios. Hoy en día esto es posible, porque se cuenta con tres tecnologías lo suficientemente maduras que soportan su aplicación como son la recopilación de datos de forma masiva, los algoritmos de minería de datos y computadoras poderosas con multiprocesamiento.

Ya en la década del 60 se comenzaban a emplear términos como arqueología de datos o pesca de datos, por algunos estadísticos, pues es una de las principales áreas del conocimiento que intervienen en el proceso de minería, pero no fue hasta principio de los 80 donde especialistas como: Robert Blum, Gio Wiederhold, Gregory Piatetsky-Shapiro y Rakesh Agrawal, empezaron a consolidar el término minería de datos.

A partir de ese momento las empresas se comenzaron a interesar en el uso y aplicación de esta tecnología por ser muy provechosa en el ámbito comercial, convirtiéndose en un punto de encuentro entre el ámbito académico y de los negocios.

De esta relación parten los conceptos utilizados actualmente, uno más tradicional que plantea: "…*que es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos…*" (U. Fayyad, 1996), y otro desde el punto de vista empresarial que lo define como: "…*la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión…*" (MOLINA, et al., 2001). Desde la apreciación de los usuarios esta evolución del manejo de la información ha permitido brindarle respuestas a nuevas y cada vez más complejas preguntas sobre los negocios.

1.2. El proceso de minería de datos

Durante el desarrollo de un proyecto de minería de datos se usan diversas aplicaciones, relacionadas directamente con cada una de las etapas del proceso por el que transitan los datos, principalmente se usan herramientas de estadísticas, inteligencia artificial y visualización de datos (Han, et al., 2006).

El proceso tiene como objetivo fundamental extraer los datos de las bases de datos y analizarlos, para a partir de estos obtener información valiosa, pero los datos se encuentran almacenados de diferentes sistemas de almacenamiento que pueden ser, bases de datos relacionales, almacenes de datos o incluso simples archivos.

Generalmente son grandes volúmenes de datos históricos con años de antigüedad, de lo que se deriva algunas veces la necesidad de utilizar procesamiento paralelo. Como resultado final del proceso se obtiene cinco tipos de información que pueden ser: secuencias, clasificaciones, pronósticos, agrupamientos y asociaciones (J. Hernandez Orallo, 2004). Por lo general el minero o usuario final carece de los conocimientos y habilidad de programación, y solo se limita al uso de estas poderosas herramientas para obtener respuestas.

Es necesario señalar, que el proceso de minería de datos invierte la dinámica del método científico, ya que a diferencia del mismo, se coleccionan los datos y se espera a que emerjan hipótesis de ellos, se busca que los datos describan por qué son como son y después se valida la hipótesis en los datos, de ahí que en la minería de datos se presente un enfoque exploratorio y no confirmador (Han, et al., 2006).

1.3. Relación con otras tecnologías

La procedencia de los datos es variada, las formas de almacenamiento poseen diferentes estructuras y se usan diferentes tecnologías de acceso a las mismas. Dependiendo de las características del origen de los datos varía la rigurosidad y complejidad del proceso de minería de datos, si los datos se encuentran almacenados en un almacén de datos o un *Data Mart*[1] el proceso es menos complejo pues los datos han sido previamente limpiados y su configuración resuelve el problema de integridad de los datos, de no ser así se debe realizar este proceso para garantizar un correcto procesamiento de los datos.

[1] Data Mart: Es un pequeño almacén de datos aplicado únicamente a un área funcional de la empresa.

Del origen también dependen las posibilidades de acceso a los datos, y potencialidades que brindan el uso de otras tecnologías, pues si la información parte de una base de datos multidimensional se puede usar *OLAP*[2] que brinda un espectro de herramientas para la toma de decisiones, y a diferencia de la minería de datos genera una serie de modelos hipotéticos que trata de verificar a través de consultas (Han, et al., 2006).

Cuando la cantidad de variables a analizar crece se necesita gran cantidad de tiempo para encontrar una hipótesis que pueda ser comprobada por el sistema. Aunque en las fases iniciales del proceso de minería de datos *OLAP* es una herramienta complementaria porque puede ayudar a explorar los datos enfocando atención en variables importantes, e identificar excepciones o hallazgos (Han, et al., 2006).

1.4. Niveles de conocimiento

Los métodos de almacenamiento de los datos definen las características iniciales del proceso de minería de datos, pero existen diversos niveles de conocimiento a extraer.

Los sistemas de almacenamiento de datos se caracterizan por tener tres niveles fundamentales de conocimiento, cada uno es más complejo y profundo que el anterior, y en cuanto a las características de la información existente en ellos es necesario el uso de técnicas cada vez más complejas (Han, et al., 2006).

El primer nivel es el del conocimiento evidente y explícito al cual se accede mediante simples consultas *SQL*[3] motivada por una hipótesis muy concreta, siendo la información necesaria para administrar la actividad del negocio de forma frecuente.

Al segundo nivel le corresponde el conocimiento multidimensional, con el cual se usa *OLAP* para organizar los datos y ajustarlos al modo de análisis y administración de los usuarios, facilitando el proceso de creación de informes. Al crear un cubo *OLAP* a partir de una consulta se convierte al conjunto de registros planos en un cubo, una jerarquía estructurada que permite que los informes tengan el nivel de detalle deseado, y se predefinen los valores de resumen de los informes, para acelerar el proceso de cálculo de los mismos.

El tercer nivel posee el conocimiento oculto, el más útil y a la vez el más difícil de extraer, pues es necesaria la aplicación de técnicas de diversas áreas del conocimiento, para conformar el conjunto de

[2] OLAP: Procesamiento Analítico en Línea
[3] SQL: Lenguaje de consulta estructurado.

algoritmos necesarios para el proceso de extracción, conocido como minería de datos (Han, et al., 2006).

1.5. Fases del proceso de minería de datos

El proceso de minería de datos consta de cinco fases fundamentales, las cuales definen las características del proceso. Del correcto desarrollo de cada una de estas fases depende la efectividad y la eficiencia de todo el proceso.

Comprensión del negocio y del problema:

Esta fase del proceso de minería de datos es una de las más delicadas por ser la que da inicio, de ella depende la efectividad del proceso, pues los datos contenidos en la fuente tienen un formato que frecuentemente no es el adecuado, sobre el cual es muy difícil la aplicación de algún algoritmo de minería; por tanto los datos iniciales requieren ser limpiados y transformados para eliminar el ruido y los errores humanos (J. Hernandez Orallo, 2004). Es importante señalar que aunque la fuente sea un almacén de datos o una base de datos multidimensional generalmente es necesario realizar algunas de estas transformaciones.

Todo este proceso se debe realizar teniendo en cuenta la integridad de los datos y del sistema de manera tal que no afecte su funcionamiento ni la lógica del negocio, además se obtienen muestras de los datos en busca de mayor eficiencia y velocidad de los algoritmos o se reducen el número de valores posibles para los atributos de análisis.

Filtrado de datos:

Como es muy difícil la aplicación de los algoritmos de minería de datos sobre los datos en bruto, es necesario realizar un proceso de filtrado. Este proceso debe tener en cuenta los parámetros que pueden procesar los algoritmos, como la presencia de valores faltantes o el trabajo con cadenas.

Dependiendo de la característica de la fuente, y los algoritmos a utilizar posteriormente, se eliminan valores no válidos o innecesarios, además para acelerar el proceso se reducen el número de valores posibles mediante métodos de agrupamiento y redondeo, ajustándose así a las necesidades del algoritmo (J. Hernandez Orallo, 2004).

Selección de variables:

Después del pre-procesamiento y reducción al cual fueron sometidos los datos, por lo general queda una gran cantidad de datos para analizar. El proceso de selección de variables tiene como principal objetivo reducir la dimensionalidad del espacio de variables, seleccionando sobre todo las variables relevantes, obviando las redundantes que son las que poseen información contenida en otra variable. También existen variables consideradas como irrelevantes pues es muy poca la cantidad de información que aportan.

Con las variables seleccionadas se construye un conjunto más pequeño de nuevas variables aplicando una transformación lineal o no lineal al conjunto original. Para la selección de este subconjunto se usan básicamente dos métodos, aquellos que priorizan la selección de los mejores atributos para el problema y los que mediante algoritmos de distancia, heurísticos o test de sensibilidad buscan las variables independientes (J. Hernandez Orallo, 2004).

Este proceso se caracteriza por tener cinco pasos fundamentales:

- Proceso de generación: En este paso se trata de encontrar el subconjunto óptimo para el problema el cual puede ser heurístico, aleatorio o completo.
- Función de evaluación: Se persigue encontrar las métricas del conjunto seleccionado, se calculan las medidas de distancia, de información, de consistencia de dependencia y la tasa de error de clasificación.
- Criterio de parada: Es el criterio bajo el cual se asume que el subconjunto resultante posee la riqueza de información y las características necesarias para ser utilizado, este criterio puede ser un umbral determinado, una cantidad de iteraciones predefinidas o un tamaño predefinido del mejor subconjunto de variables.
- Proceso de validación: Es el único paso opcional, pero muy recomendado, aquí se verifica la validez del subconjunto seleccionado, y de no cumplir los requisitos necesarios se repite todo el proceso nuevamente realizando modificaciones en los pasos anteriores, hasta obtener el resultado deseado.

Extracción del conocimiento:

Existen diversos métodos de extracción la información relevante de los datos que pueden ser usados para entender los datos, reorganizarlos en las bases de datos, determinar sus características de manera simple y concisa, y descubrir relaciones entre los datos espaciales y no espaciales (Han, et al., 2006).

Existen cinco grupos de métodos de minería de datos:

- Métodos basados en generalización: Los cuales requieren la implementación de jerarquías de conceptos, en el caso de las bases de datos espaciales estas jerarquías pueden ser temáticas o espaciales.
- Métodos de reconocimiento de patrones: Estos pueden ser usados para realizar reconocimientos y categorizaciones automáticas de fotografías, imágenes y textos, entre otros, por lo general usan algoritmos de inteligencia artificial.
- Métodos que usan agrupamiento: Consisten en crear agrupaciones o asociaciones de datos, cuando en estos existan nociones de similitud (por ejemplo, distancia Euclidiana). Agrupamiento o clustering es el proceso de agrupar datos en grupos o clusters de tal forma que los objetos de un cluster tengan una alta similitud entre ellos, y baja con objetos de otros clusters.
- Métodos explorando asociaciones espaciales: Permiten descubrir reglas de asociaciones espaciales, es decir, reglas que asocien uno o más objetos espaciales con otro u otros objetos espaciales Su aplicación está en bases de datos grandes, donde puede existir una gran cantidad de asociaciones entre los objetos, pero la mayoría de ellos serán aplicables solamente a un pequeño número de objetos, teniendo en cuenta que la confianza de la regla puede ser baja.
- Métodos que utilizan aproximación y agregación: Descubren conocimiento en base a las características representativas del conjunto de datos. La proximidad agregada es la medida de proximidad del sistema de puntos en el grupo en base a una característica en comparación con el límite del grupo y el límite de una característica. Las consultas de proximidad solicitan objetos que se hallen cerca de una posición específica

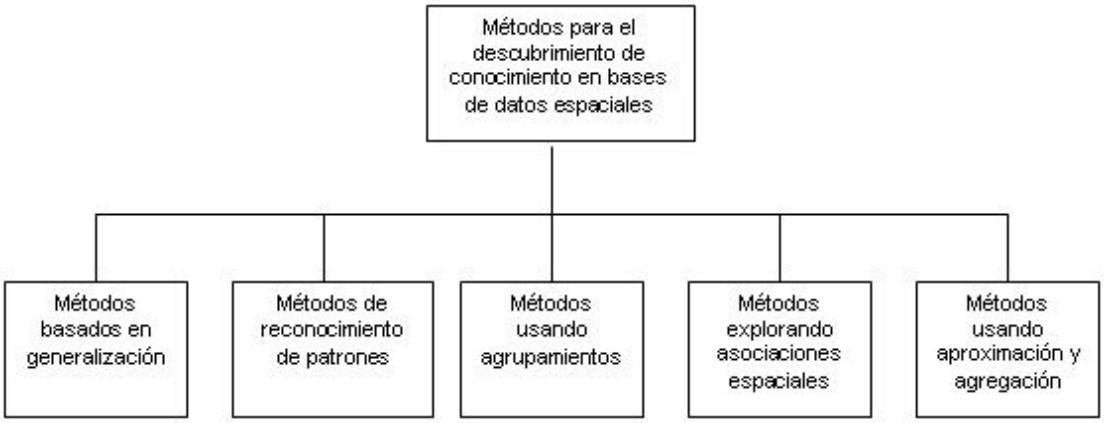

Figura 1: Clasificación de los métodos de minería de datos.

A partir de la aplicación de una de estas técnicas se obtiene un modelo de conocimiento, que representa patrones de comportamiento o relaciones de asociación existentes en los valores de las variables; aunque en algunas ocasiones se utilicen varias técnicas a la vez con el objetivo de generar distintos modelos.

Interpretación y evaluación:

Después de haber obtenido el modelo es necesario validarlo, comprobando la autenticidad de las conclusiones arrojadas, si estas son suficientemente satisfactorias, determina la efectividad del modelo seleccionado.

En caso de ser empleadas diversas técnicas para generar disimiles modelos, es necesario realizar una comparación entre cada uno, para determinar cuál de los modelos generados se ajusta más al problema planteado, en caso que ninguno de los modelos alcance los resultados esperados, entonces es necesario alterar algunos de los pasos anteriores, para generar nuevos modelos de mejores resultados.

1.6. Algoritmos utilizados en los distintos modelos

Independientemente de las características específicas de un proyecto de minería de datos y los métodos de extracción del conocimiento que se utilicen, generalmente se buscan tres tipos de resultados diferentes que son: clasificaciones, predicciones y asociaciones (Han, et al., 2006). Según el resultado perseguido se aplica alguno de los diversos algoritmos correspondiente a cada grupo. En la Figura 2 se muestra la taxonomía de los algoritmos correspondientes a las técnicas utilizadas.

Figura 2: Taxonomía de las técnicas y algoritmos de minería de datos.

1.6.1. Clasificación

La clasificación de los datos es un proceso de dos etapas. En la primera etapa se construye un clasificador para describir un conjunto de datos predeterminado, esta etapa es considerada como fase de aprendizaje o entrenamiento, donde el algoritmo de clasificación construye el clasificador, analizando en conjunto de datos de entrenamiento. Después se aplican las reglas del clasificador al conjunto de datos objetivo para obtener el conjunto de nuevos datos.

Árboles de decisión:

Al representar este algoritmo en un diagrama se comporta como el flujo de un árbol, en el cual a cada nodo que no es hoja le corresponde una prueba o pregunta a un atributo determinado, las ramas representas las salidas de la prueba y cada hoja contiene una etiqueta. Para realizar la clasificación usando árboles de decisión se toma una *tupla*[4] cualquiera con su etiqueta asociada desconocida, los valores de los atributos se prueban usando las preguntas del árbol de decisión y se obtiene como

[4] Tupla: Notación básica para formular la definición de la relación en términos de las relaciones de la base de datos.

resultado la etiqueta asociada a la *tupla*. Se caracterizan por ser recorridos rápidamente después de ser construidos.

Clasificadores bayesianos:

Están basados en el teorema de Bayes y son clasificadores estáticos para realizar predicciones. Pueden predecir las probabilidades de pertenencia a una clase, así como la probabilidad que tiene una *tupla* dada de pertenecer a una clase específica. Basándose en el mismo teorema existen diversas modificaciones del algoritmo que dan paso a diversos algoritmos como Naïve Bayesian y Bayesian belief networks, cada uno tiene sus características propias, pero comparten el nivel de eficiencia y velocidad.

1.6.2. Predicción

Predecir el comportamiento de una o varias variables es otro de los objetivos del proceso de minería de datos; las predicciones se realizan modelando la relación existente entre una o varias variable independientes y una variable dependiente, y al definirse esta relación se puede predecir el comportamiento de la variable dependiente con cierto nivel de certeza.

Regresión lineal:

Es un método matemático que modela la relación entre una variable dependiente Y las variables independientes X_i teniendo una muestra de n individuos para los que se dispone los valores de las variables y visualizándolos en un gráfico de dispersión tiene como objetivo encontrar una recta que se ajuste a la nube de puntos del diagrama y que pueda ser utilizada para predecir los valores de Y a partir de los de X.

Teniendo la ecuación general de la recta de regresión la forma: $Y = a + bX$; el problema sería encontrar la recta que se ajuste mejor a los datos, para solucionar este problema se ha recurrido al método de mínimos cuadrados, el cual es capaz de encontrar la línea más minimizando el error entre el dato actual y el estimado de la línea. En la Figura 3 se puede apreciar u ejemplo gráfico de una regresión lineal en un eje cartesiano.

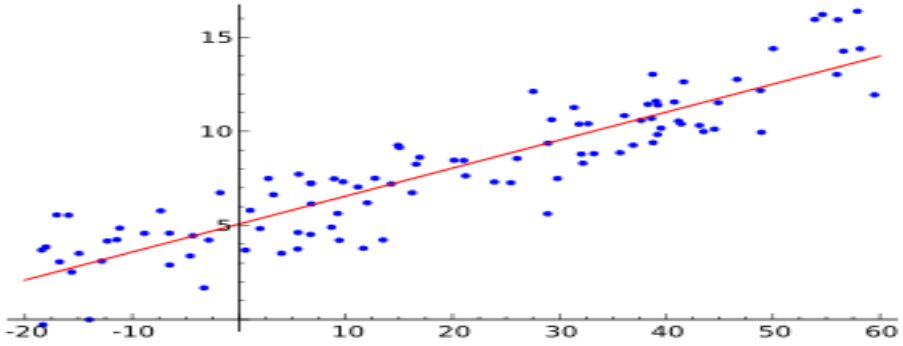

Figura 3: Ejemplo de regresión lineal.

Regresión no lineal:

La regresión no lineal es un problema de inferencia para un modelo que contiene una función no lineal F. Existen algunos casos que la función F pude ser sometida a un proceso de linealización para transformarla en una función lineal. Este método pretende obtener los valores de los parámetros asociados con la mejor curva de ajuste usando generalmente el mismo método de la regresión lineal. En la Figura 4 se exponen varios ejemplos de funciones de regresión no lineal.

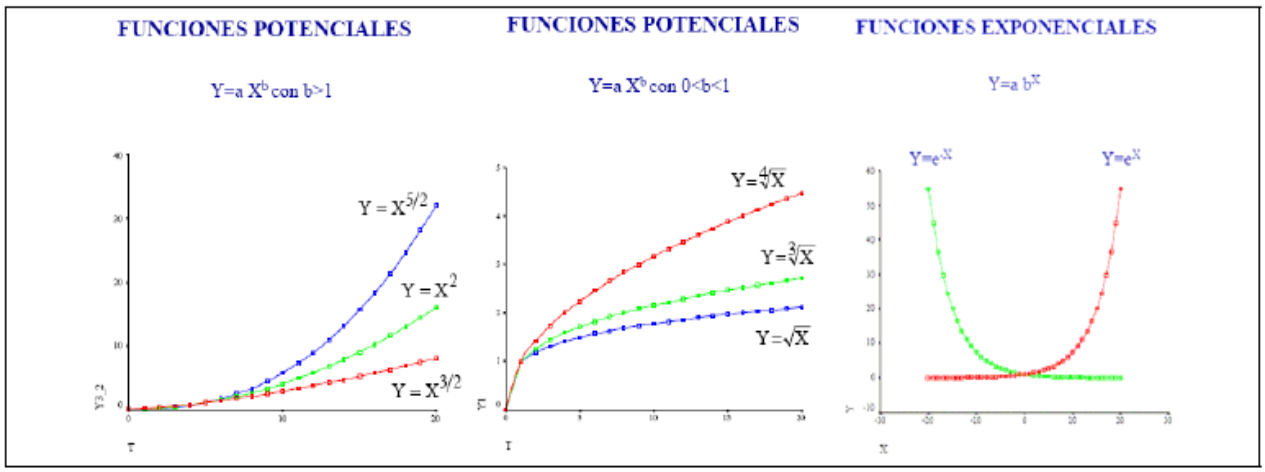

Figura 4: Ejemplos de funciones de regresión no lineal.

1.6.3. Agrupamiento

Es el proceso de agrupar un conjunto de objetos en clases de objetos similares, los elementos pertenecientes a estos grupos tienen la característica que son similares entre ellos mismo y diferentes de los elementos pertenecientes a otro grupo y son tratados como un grupo. Esta definición de grupos es muy efectiva para distinguir objetos de la misma clase pero también es muy utilizada para detectar

comportamientos irregulares que se manifiestan como objetos que no pertenecen a ningún conjunto en específico u *outliers*.

Los algoritmos de agrupamiento se dividen en tres categorías fundamentales, pero esto ha cambiado un poco, pues con el tiempo muchos de estos algoritmos han evolucionado, incluso han surgido nuevos métodos como resultado de la mezcla de los algoritmos de categorías diferentes, con el objetivo de obtener mejores resultados.

En la Figura 5 se muestran los algoritmos de agrupamiento por categorías y además se muestran sus relaciones, es decir los algoritmos que surgieron a partir de otros.

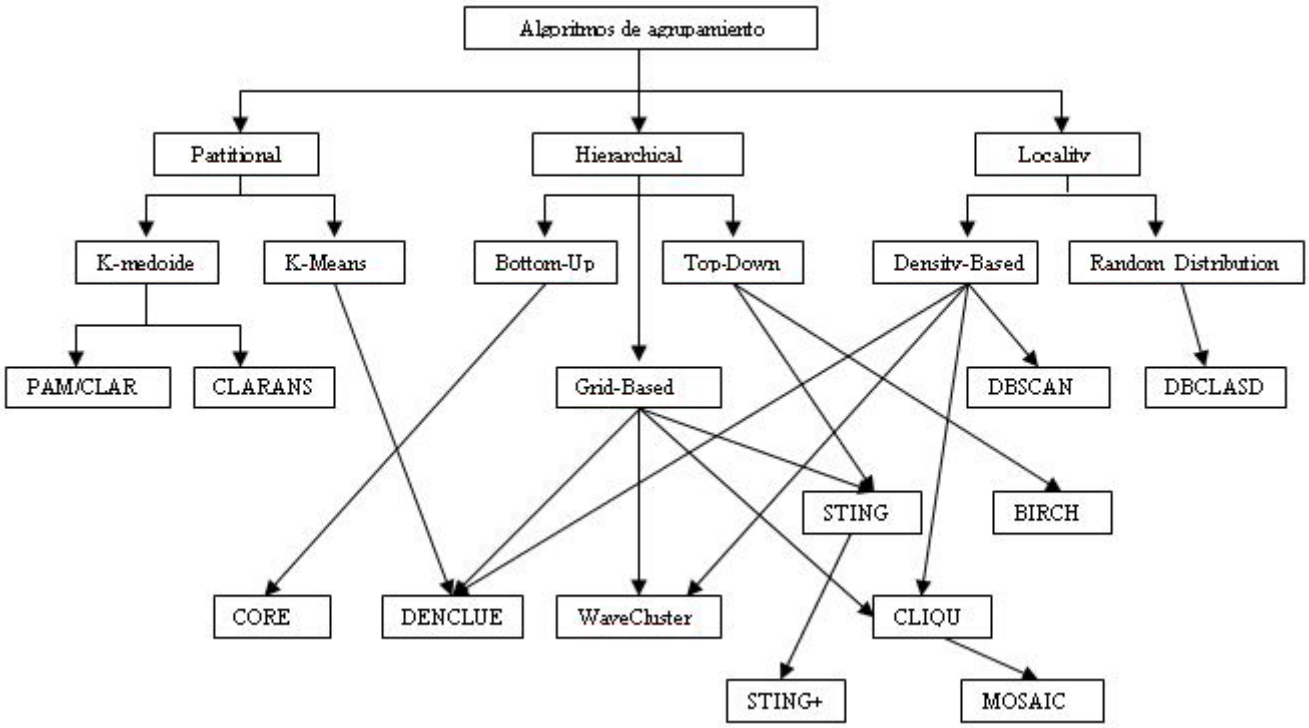

Figura 5: Clasificación y relación de los algoritmos de agrupamiento.

1.7. Aplicación de la minería de datos

Por el gran potencial que poseen las técnicas de extracción de la información son aplicables a diversas áreas de la sociedad, de hecho es potencialmente aplicable a cualquier entidad que recopile datos con cierta frecuencia (J. Hernandez Orallo, 2004).

En el campo empresarial es utilizado para detectar fraudes de tarjetas de crédito, un problema que anualmente producía gran cantidad de pérdidas a las entidades bancarias, a raíz de esto se desarrolló

el software Falcon Fraud Manager el cual permite detectar y mitigar fraudes en tarjetas bancarias, comerciales, de combustible y de débito.

Otro aspecto que preocupa a las empresas es la relación cliente proveedor, porque en muchos casos las empresas tienen perdidas de clientes sin el debido conocimiento de las causas del fenómeno, causas que en muchos casos puede ser mitigables, este problema es solucionado utilizando técnicas de minería de datos aplicadas en función de predecir el comportamiento los clientes. Pero no solo se utiliza para reducir la pérdida de clientes también permite a las empresas brindar un servicio más personalizado y selectivo, adelantándose a sus necesidades en el momento requerido.

No se limita la aplicación de la minería de datos al ámbito de los negocios y las empresas, también es aplicado a otras esferas.

En el deporte está el caso del software Advanced Scout que se emplea para apoyar a los entrenadores de basquetbol utiliza técnicas de minería de datos para detectar patrones estadísticos y eventos raros. Las universidades lo aprovechan para conocer el perfil que caracteriza a los estudiantes durante su estancia en la institución y ajustar el programa de clases.

Fue aplicado en investigaciones espaciales, como en el caso del software SPYCAT, que fue capaz de clasificar los objetos celestes captados en fotos tomadas desde el observatorio POSS-II, manejando técnicas de agrupamiento y árboles de decisión.

Minería de web

De las aplicaciones de la minería de datos que se usan hoy en día una de las más importantes es la minería sobre sitios web. Actualmente es una de las principales líneas de investigación por el gran potencial que ofrece a causa de la gran cantidad de información que transita vía web. Como es muy común registrar las acciones de los usuarios mientras navegan, es de mucha utilidad la información generada en el proceso, la minería de web se encarga de analizar estos datos para obtener información de los mismos.

De acuerdo con la naturaleza de los datos se clasifica en tres dominios de extracción del conocimiento:

Minería de contenido web: Utiliza como fuente el contenido de los documentos o sus descripciones para extraer la información basándose en conceptos de indexación o localizando patrones en el texto.

Minería de la estructura de la web: Es el proceso de inferir conocimiento utilizando la estructura y organización de la WWW.

Minería de uso de la web: Se extraen los modelos de alto interés utilizando como fuente fundamental de datos los log de accesos al sitio.

Minería de textos

Independientemente del desarrollo tecnológico alcanzado actualmente por la sociedad la mayoría de la información de las compañías esta almacenada en documentos. Utilidades como el procesamiento del lenguaje natural, el aprendizaje automático, la categorización del texto y la extracción y recuperación de la información, son de las grandes ventajas proporcionadas por esta técnica, que se basa en examinar una colección de documentos y descubrir la información no contenida en ninguno de forma individual. Esta podría ser la herramienta definitiva para los investigadores pues una persona de forma individual es capaz de procesar una fracción de la información publicada y contenida en los textos referentes a una materia en específico.

1.8. Debilidades de los sistemas de minería de datos

Aunque la aplicación de los sistemas de minería de datos reportan muchas ventajas, poseen debilidades las cuales es necesario conocer pues de no tenerlas en cuenta pueden poner en peligro el desarrollo efectivo de proceso de extracción de la información.

Es necesario señalar, como primer aspecto, la facilidad con la que se puede malinterpretar la información, esta característica parte de una de las áreas del conocimiento que intervienen en el proceso de minería de datos, la estadística. La posibilidad de ser engañado por las estadísticas es conocida, pues muchos de los indicadores y resúmenes son generalizaciones de grandes volúmenes de datos donde se hace una inferencia, la cual posee un riesgo o margen de error. Por eso en todo momento se debe tener en cuenta que usualmente estos resultados representan poblaciones y no a individuos.

El riesgo de equivocarse utilizando técnicas de minería de datos no es bajo, el hecho de que se encuentren correlaciones estadísticamente significativas no quiere decir que se encontró la relación causa efecto. Como la minería de datos es un proceso explorativo arroja hipótesis sobre los datos, las cuales se infieren como verídicas, pues se formulan creencias partiendo de un número de experiencia finita, resulta de vital importancia revisar la hipótesis usando otras experiencias para validar su veracidad.

El utilizar computadoras para modelar el tiempo es una tarea compleja a la hora de hacer inferencias y encontrar patrones, pues se convierten en una carga difícil de procesar por los algoritmos de

aprendizaje automático. El hacer inferencias y análisis de datos en un período de tiempo dado es una de las tareas más comunes solicitadas a un sistema de minería de datos, pero existe la probabilidad que en el periodo seleccionado no se hayan registrado el mismo número de variables, estas no tengan la misma interpretación o que no tengan la misma precisión.

Cuando se trabaja con intervalos de tiempo también hay que tomar en consideración la discontinuidad de las variables con respecto al tiempo, si se realizan recopilaciones semanales resulta imposible realizar una predicción diaria en función de la misma variable. Para realizar comparaciones entre los resultados del proceso de minería de datos en tiempos diferentes es necesario tener en cuenta que después de ser procesados los datos no se puede recopilar los conjuntos y variables que le dieron origen.

Otro aspecto delicado a considerar es la privacidad de la información y el manejo de los resultados de los análisis. Cuando los datos son referentes a personas puede arrojar información reservada, por lo tanto la minería de datos se convierte en un proceso que puede atentar seriamente contra la privacidad de los datos de clientes o usuarios, aunque actualmente las definiciones de privacidad y propiedad de los datos necesitan ser reevaluadas, es necesario tener este aspecto presente a la hora de desarrollar un proyecto.

1.9. Revisión de modelos anteriores

En general existen tres tipos de modelos matemáticos: modelos empíricos, modelos mecanicistas y modelos teleonómicos.

Modelos empíricos son descripciones directas de datos y proporcionan relaciones observables entre las variables de un sistema o fenómeno, sin proporcionar ninguna explicación de los organismos subyacentes. Estos modelos son un poderoso medio para describir y resumir datos, ejemplos de ellos son los modelos de regresión simple y regresión múltiple, redes neuronales, modelos difusos o modelos neuro-difusos.

Modelos teleonómicos son aplicables a comportamientos dirigidos por metas y se formula explícitamente en términos de objetivos.

Modelos explicativos o mecanicistas generalmente son modelos determinísticos (Thornley, et al., 2000).

"*…Durante los últimos 30 años el enfoque mecanicista se ha empleado exitosamente en la modelación de procesos de crecimiento de cultivos…"* (Thornley, et al., 2000) (Goudriaan, et al., 1994). Dentro de

los resultados más conocidos se encuentra el modelo SUCROS (a Simple and Universal CROp growth Simulator), por su eficiencia es considerado como punto de partida para la ceración de nuevos modelos de crecimiento y desarrollo de cultivos más complejos. El modelo SUCROS simula el crecimiento potencial de un cultivo al aire libre partiendo de la suposición de que el terreno se encuentra libre de plagas, enfermedades y la planta tiene una amplia disponibilidad de nutrientes y agua.

En este modelo el crecimiento y desarrollo solo depende de variables ambientales radiación, temperatura, concentración de CO^2 y las características genéticas de la planta (Goudriaan, et al., 1994).

Para el cultivo de la cebolla en específico se desarrolló el modelo matemático de crecimiento potencial ALCEPAS (VISSER, 1994).

Es necesario resaltar que estos algoritmos cuentan con una mayor cantidad de variables y están enfocados en el proceso de crecimiento de los cultivos, son más generales que el modelo que se pretende desarrollar.

1.10. Software de minería de datos

Para definir el modelo matemático que será aplicado sobre los datos existen diversos software específicos, que se caracterizan por centrarse en un modelo único. Para realizar un estudio de los datos se requieren herramientas más completas y generales, estas se puede dividir en 2 grandes grupos: comerciales y académicos (Piatetsky-Shapiro).

Entre las herramientas comerciales destaca SAS Enterprise Miner que es uno de los sitios comerciales que proporciona el SAS Institute para tareas de minera de datos. Posee una arquitectura distribuida con una potente interfaz de usuario, entre las tareas que realiza la herramienta están: pre-procesamiento de los datos, modelos matemáticos, evaluación para comprobar la eficiencia y eficacia de la herramienta, y la visualización de resultados.

Otra herramienta comercial destacada es Oracle Data Xining Suite (ODMS), diseñado sobre una arquitectura cliente servidor y ofrece gran versatilidad para acceder a grandes volúmenes de información. Se caracteriza principalmente por su acceso a datos en diversos formatos, entre los que se encuentran bases de datos relacionales y archivos de texto plano. Realiza pre-procesamiento de los datos, aplica modelos de aprendizaje y posee herramientas de visualización que muestra resultados estadísticos, y es capaz de importar datos en Excel o Word.

Entre las herramientas académicas se destacan XLminer y WEKA. XLminer es un complemento para Excel con funcionamiento mediante macros[5] que permite varios tipos de análisis, además es capaz de rellenar los datos faltantes, realizar predicciones, aplicar varios modelos matemáticos y tiene facilidad para la entrega de reportes. Cuenta con opciones de configuración de trabajo y una interfaz amigable para cada método, y los formatos de presentación de resultados como gráficos y tablas son bastante ordenados. También se caracteriza por no poseer indicadores de errores claro, al parametrizar alguna operación con datos inadecuados la operación se rompe y cae la aplicación en la versión de prueba. La versión de prueba solo es utilizable por un periodo de 30 días y se debe pagar para tener acceso a la versión que no limita el tamaño de la base de datos.

Waikato Enviroment for Knowledge Analysis (WEKA por sus siglas en inglés) es una herramienta visual libre, utilizada para el aprendizaje automático y minería de datos, desarrollada por los investigadores de la Universidad de Waikato en Nueva Zelanda el año 1993. Permite el acceso a datos desde archivos en formato arff[6] y cuenta con una serie de algoritmos para realizar el pre-procesamiento de los datos, además de diferentes modelos de aprendizaje, y posee una interfaz gráfica que unifica las herramientas para que estén a una mejor disposición.

Es una herramienta que soporta varias de las tareas comúnmente realizadas en el proceso de minería de datos y también por medio de SQL permite el acceso a instancias de bases de datos. Dentro de sus ventajas resalta que está disponible bajo la licencia GNU-GPL, contiene una gran variedad de técnicas para el modelado y procesamiento de datos, es multiplataforma y posee una interfaz de usuario sencilla. Tiene como principales desventajas que no incluye algoritmos de modelado de secuencias, además al utilizar métodos de combinación de modelos los resultados tienden a complicarse perdiendo comprensibilidad y solo puede tener una variable dependiente como objetivo de los algoritmos.

A parte de las herramientas mencionadas existen un gran número de herramientas para la utilización de las técnicas de minería de datos, en la Tabla 1 se muestra una comparación de varias herramientas entre las que se encuentran algunas de las mencionadas.

[5] macros: Instrucciones de seguimiento cronológico para realizar tareas.
[6] arff: Es un archivo de texto plano organizado en filas y columnas.

Tabla 1: Comparación entre aplicaciones de minería de datos.

Característica	Clementine	SAS Enterprice Miner	Tariykdd	WEKA
Licencia libre	No	No	Si	Si
Requiere conocimientos avanzados.	No	No	No	No
Acceso a SQL	Si	Si	Si	Si
Multiplataforma	No	-	Si	Si
Requiere de bases de datos especializadas	No	Si	No	No
Métodos de máquinas de soporte vectorial	Si	-	No	Si
Métodos bayesianos	Si	Si	No	Si
Puede combinar modelos	Si	Si	No	Si
Modelos de clasificación	Si	Si	Si	Si
Implementa árboles de decisión	Si	Si	Si	Si
Modelos de regresión	Si	Si	No	Si
Clustering y agrupamiento	Si	Si	No	Si
Interfaz amigable	Si	Si	Si	Si
Permite visualización	Si	Si	Si	Si

2. Capítulo 2: Descripción de variables y algoritmos

Existe una gran diversidad de técnicas y algoritmos, que se pueden utilizar, para construir variedad de modelos matemáticos enfocados en la solución de un problema determinado. En este capítulo se describen las características y la estructura en la cual están almacenadas las variables y se mencionan los algoritmos matemáticos para procesar los datos almacenados por el sistema DSerp Agro.

2.1. Data Mart Siembra

El software DSerp Agro es un Sistema de Administración de Negocios Empresariales (EBMS por sus siglas en inglés) el cual tiene divididas las actividades empresariales por procesos, con una base de datos centralizada. Para garantizar la persistencia de los datos del proceso de siembra el sistema cuenta con un *Data Mart*, el cual almacena e integra un conjunto de variables relacionadas con estas áreas de procesos.

El *Data Mart* posee una arquitectura *snowflake* (Han, et al., 2006) y en una tabla *fact*[7] contiene los valores relacionados con el cultivo, la calidad y el índice de producción del producto por cosecha, siendo estos aspectos la medida fundamental del negocio. En las tablas *lock_up*[8] se almacenan los datos adicionales que se registran en el sistema, es decir, el resto de las variables que influyen sobre los valores de la tabla *fact*.

En las tablas del *Data Mart Siembra*, cada una de las dimensiones representadas en las tablas *look_up*, que describen el fenómeno, se toman como las variables que influyen sobre el mismo y las dimensiones de la tabla *fact* son las variables que representan el fenómeno.

2.2. Origen de los datos

La efectividad del modelo matemático depende en gran medida de la calidad de los datos usados para el entrenamiento de los algoritmos. La calidad de la muestra es determinante, ya que requiere de variedad suficiente en los datos. De no ser así, los algoritmos se limitan únicamente a describir la muestra, y por lo tanto, no cumplen con el propósito fundamental que consiste en describir el fenómeno.

[7] tabla fact: contienen los valores de las medidas de negocios
[8] tabla lock_up: contienen el detalle de los valores que se encuentran asociados a la tabla fact.

Como el software DSerp Agro es de reciente creación, posee muy poco tiempo de aplicación y por lo que no cuenta con suficiente información en su base de datos.

La muestra de cien casos de cultivo de cebolla forma parte de los datos personales del profesor Msc. Juan Matos Borges, investigador de la Universidad de Guantánamo los cuales se ajustaron al formato arff requerido. El estudio del comportamiento de los datos fue realizado a partir de esa muestra, para el entrenamiento de los algoritmos y la determinación de su efectividad.

2.3. Descripción de las variables seleccionadas

Las variables que representan el problema son definidas por las dimensiones del *Data Mart Siembra*, pero no se utilizaron todas las variables. Como uno de los objetivos de este trabajo se enfocó en el estudio de los datos referentes a la cebolla como cultivo representativo del genero Liliaceae, no se consideraron aquellas que hacían referencia a otro tipo de cultivo, reduciendo de esta forma la cantidad de dimensiones a dos (Calidad y Rendimiento) y el dominio de las variables.

2.3.1. Variables independientes

Después de eliminar las variables innecesarias, se asumieron como variables independientes las siguientes:

PH del suelo: Describe el nivel de acidez del suelo, factor importante en el crecimiento del cultivo.

Humedad del suelo: Representa el por ciento de humedad relativa del suelo, que viene dado por el nivel de retención del agua en las capas superiores.

Salinidad del suelo: Representa el nivel de salinidad del suelo, que se obtiene mediante una prueba de acidez, y se tiene en cuenta para la aplicación de determinados tipos de fertilizantes que afectan este valor.

Profundidad del suelo: Describe el nivel de profundidad del suelo cultivable, que influye en la correcta absorción de los nutrientes.

Población: Cantidad de plantas sembradas en un área determinada, base fundamental para el estudio del rendimiento.

Estado de las semillas: Valor cualitativo que describe la calidad de las semillas utilizadas para la siembra. Este valor es determinado mediante una prueba de germinación.

Forma de siembra: Describe de manera cualitativa las diferentes formas en que puede ser sembrado el cultivo, que en el caso específico de la cebolla pueden ser: siembra directa, mediante trasplante o producción de bulbillos.

Forma de recolección: Describe de manera cualitativa las vías utilizadas para recolectar el cultivo, que para el caso específico de la cebolla puede ser manual o mecanizada.

Temperatura en el día: Promedio de temperatura durante el día durante el periodo de desarrollo del cultivo.

Temperatura en la noche: Promedio de temperatura durante la noche durante el periodo de desarrollo del cultivo.

Personal: Cantidad de personas atendiendo el cultivo, factor que en caso de no alcanzar los valores adecuados puede influir negativamente en la productividad y la calidad.

Riego: Cantidad de agua aplicada al cultivo. En el caso del cultivo de la cebolla, para realizar el riego se necesita tener en cuenta las precipitaciones ocurridas durante el desarrollo de la planta.

Variables de abonos y fertilizantes: Es un conjunto de variables que representan los elementos químicos, que componen los abonos y fertilizantes administrados a las plantas. La proporción de estos elementos químicos en el abono o fertilizantes varía de acuerdo con la sustancia de que está compuesta. Estos elementos son nitrógeno, fósforo, potasio, calcio, magnesio, sodio, azufre, boro, cobalto, cobre, hierro, manganeso, molibdeno y zinc. Cada elemento es una variable que representa la cantidad (en kilogramos) del elemento distribuido por unidad de área (hectáreas).

Período de siembra: Contiene la información de la fecha en que se sembró la planta, de la cual se deduce la estación del año.

Período de recolección: Contiene la información de la fecha en que se recolectó el producto, que en el caso específico de la cebolla se arranca del terreno y se deja reposar en la tierra de 3 a 5 días. Estos días son tenidos en cuenta, por tanto, la fecha que se almacena es cuando ya está lista para su comercialización.

2.3.2. Variables dependientes

Como variables dependientes del modelo se tomaron Calidad y Rendimiento, que son las que representan el resultado final en el proceso completo. La dependencia de ambas está descrita por los documentos técnicos referentes al cultivo de la cebolla (Pérez Domínguez, 1988). Según ese

documento, la Calidad está determinada directamente por la correcta aplicación y distribución de la cantidad de elementos químicos correspondiente de los abonos y fertilizantes usados, mientras que el Rendimiento está dado por la población inicial, el PH, la salinidad y la humedad del suelo, el estado de las semillas y la correcta aplicación de agua en el cultivo.

No obstante, es importante tener en cuenta que las dependencias descritas anteriormente no deben ser rigurosamente exactas. Para demostrar esto se consideraron todas las variables independientes descritas en el epígrafe 3.3.1, para así poder explorar nuevas relaciones a partir de los datos, relaciones no mencionadas en la bibliografía, pero que pueden resultar de interés y utilidad para los especialistas.

2.4. Pre-procesamiento de las variables

El *Data Mart Siembra* almacena todos los valores de las variables, copiándolos de la base de datos del sistema, donde inicialmente se encuentran, en un formato inadecuado para ser procesadas. Cada una de las variables ocupa una columna en el *Data Mart* y representan una dimensión, y pueden tomar diferentes valores según sus características.

Por las especificaciones del sistema se obliga a los usuarios al llenado solamente de los campos realmente necesarios para evitar la entrada de valores erróneos, pero la información ahí almacenada es aún difícil de procesar.

Al enviar la información hacia el *Data Mart Siembra* se realiza un procesamiento, para reducir su volumen, fusionar los datos comunes, eliminar u omitir los datos redundantes y discretizar todos los valores. De esta forma, los valores se almacenan de manera uniforme y son más fáciles de procesar porque muchos algoritmos de procesamiento de datos no soportan variables nominales o continuas.

Independientemente de la forma en que los datos estén organizados y del proceso de limpieza, integración y discretización de los datos, enviar los datos al *Data Mart Siembra* no es suficiente para trabajar directamente con estos valores, pues no poseen la uniformidad necesaria. Para poder procesar estos datos utilizando los algoritmos de procesamiento de la información se necesita normalizarlos y aplicar técnicas de reducción de atributos.

Después de modificar el estado y contenido de las variables sin que pierdan información, se hace factible la aplicación de los algoritmos de minería de datos para extraer la información. El pre-procesamiento se necesita para aligerar la carga, acelerar la ejecución e incrementar el nivel de precisión de los algoritmos aplicados sobre los datos (Han, et al., 2006).

2.5. Características de las variables independientes.

En dependencia de las características de las variables independientes, es que se evalúan los modelos matemáticos para posteriormente definir los más apropiados. Como el problema planteado consta de dos variables independientes se analizan inicialmente por separado. De esta forma se define cuál es el modelo más adecuado para cada una.

Los modelos se definen de acuerdo con el objetivo del problema, que en este caso consiste en predecir el comportamiento de las variables (Figura 2), Dado el objetivo, se conocen los posibles algoritmos que pueden llegar a conformar el modelo final.

Descripción de la variable Rendimiento:

La variable Rendimiento es el valor del resultado final de la cosecha, que representa la cantidad de plantas que se pueden comercializar. Contiene valores discretos los cuales se encuentran en el conjunto de los números naturales y su valor está influido altamente por el comportamiento de la población inicial y la cantidad de personal utilizado para atender el cultivo. Dado que los valores de las variables Personal y Población son proporcionales se utiliza solo Personal.

Descripción de la variable Calidad:

La variable Calidad contiene los valores correspondientes a la evaluación cualitativa dada a la calidad del producto, estos valores o clases pueden ser: *buena*, *mala* o *regular*. Como es una variable nominal la predicción de sus valores se resuelve como un problema de clasificación. Dependiendo de los valores tomados por el conjunto de variables independientes en la *tup*la, se clasifica en una de las clases enunciadas.

2.6. Experimentos realizados

Para definir el mejor modelo para cada una de las variables dependientes, se evaluaron diversos algoritmos. En un estudio preliminar se evaluaron algoritmos basados en reglas y los árboles de decisión, reportados en trabajos similares. Posteriormente se incluyeron en el experimento el Perceptrón Multicapas y las máquinas de soporte vectorial.

Criterio de comparación de algoritmos:

En el caso de la variable Calidad el criterio fundamental fue el por ciento de acierto de los algoritmos. También se tomó en cuenta en cuales de las clases se cometió el mayor error, pues era de mayor interés que se clasificara correctamente las clases *buena* y *regular*, porque el cultivo es tratado durante

todo el proceso agrícola bajo normas técnicas establecidas que limitan la obtención de productos de mala calidad.

En el caso de la variable Rendimiento el criterio de comparación de mayor peso fue el coeficiente de correlación, además se tomaron en cuenta otros parámetros como son los errores absoluto medio, cuadrático medio, relativo absoluto y cuadrático relativo. En la Tabla 4 se muestran los resultados obtenidos.

Calculo del error máximo:

Para ambas variables dependientes el estudio de los algoritmos se realizó tomando como punto de partida el ZeroR, el cual predice la media para los atributos numéricos y la moda para los atributos nominales (Bouckaert, 2009). La aplicación de este algoritmo fue considerado como punto de partida para definir el error máximo, a superar con otros, en cada una de las variables independientes, y se asumió como criterio base para establecer las comparaciones entre los diferentes algoritmos probados.

2.7. Algoritmos probados

En el epígrafe 3.6.2 se describe como se efectuaron las pruebas iniciales con el algoritmo ZeroR para ambas variables. En este epígrafe se describen brevemente los diversos algoritmos probados y los valores de sus respectivos parámetros.

ConjunctiveRule:

ConjunctiveRule es un algoritmo capaz de predecir el comportamiento de variables de tipo numérico y nominal. Consta de antecedentes y consecuencias, que en el caso de las variables nominales es la distribución de las clases disponibles y para las variables numéricas es la media aritmética. Si la instancia probada no cumple con la regla se predice su valor usando la clase o valor por defecto. (Bouckaert, 2009).

El algoritmo se aplicó sin considerar las expresiones exclusivas para atributos nominales, con tres ciclos, una semilla de valor uno, un peso mínimo de instancias de una regla de 2.0 y poda de error reducido.

J48graft:

El algoritmo *J48graft* construye un árbol de decisión C4.5, es una extensión del algoritmo ID3 (Mitchell, 1997). Este algoritmo genera un árbol de decisión a partir de los datos mediante particiones realizadas

recursivamente basándose en la estrategia de profundidad primero, y considera todas las posibles pruebas que pueden dividir el conjunto de datos (Quinlan, 1993).

El algoritmo se aplicó sin utilizar separaciones binarias, un factor de confianza utilizado para la poda de 0.25, un número mínimo de instancias por hora de dos, sin reetiquetar durante la inserción y sin usar el modelo de Laplace para el suavizado en las hojas.

DecisionStump:

DecisionStump es un algoritmo capaz de procesar variables numéricas y nominales además de tratar con datos faltantes como valores separados. Construye un árbol que utiliza como base el error cuadrático medio para realizar el proceso de regresión y para realizar la clasificación utiliza como base la entropía (Bouckaert, 2009).

Red neuronal Perceptrón Multicapas:

Una red neuronal artificial consiste en un conjunto de elementos de procesamiento simples conectados entre sí y entre los que se envía información mediante estas conexiones (Rojas, 1996). Existen diversos tipos de redes neuronales, pero se seleccionó el Perceptrón Multicapas que es una red neuronal artificial formada por múltiples capas que pueden ser: capas de entrada, ocultas y de salida. La diversidad en sus capas es lo que le permite resolver problemas que no son linealmente separables y sus funciones de activación deben ser derivables (Rojas., 1996).

Máquinas de soporte vectorial:

La teoría de las Máquinas de Soporte Vectorial o SVM (Support Vector Machines) por sus siglas en inglés es una técnica de clasificación (Smola., 1999) basada en la idea de minimización de riesgo estructural (Vapnik., 1995). El principio básico de las SVM es hacer un redimensionamiento del espacio de parámetros para permitir la separabilidad lineal de los datos.

La descripción dada por los datos de los vectores de soporte, es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje, con muy poco o ningún conocimiento de los datos fuera de esta frontera. Para mapear los datos se utiliza un función kernel (Smola, 1999) que busca la máxima separación entre las clases formando un agrupamiento.

3. Capítulo 3: Discusión y resultados

Después de analizar las variables y su procedencia, se determinaron los algoritmos que se ajustaban mejor al problema para conformar el modelo matemático a utilizar. Se tomaron en consideración trabajos realizados con anterioridad sobre modelos matemáticos para describir fenómenos agrícolas, prestando mayor atención a los relacionados con el desarrollo de los cultivos, y en especial con el cultivo de la cebolla.

3.1. Análisis de correlaciones

La existencia de un número excesivo de atributos incrementa complejidad del modelo y traer como consecuencia un *overfitting*[9]. Para definir el nivel de influencia de las variables independientes sobre las variables dependientes se realizó un análisis de correlaciones, que consiste en la aplicación de técnicas estadísticas para medir la intensidad de la asociación entre dos variables.

Para efectuar este proceso en el Weka, se seleccionó un método de búsqueda y uno de evaluación, tomando como algoritmo de búsqueda BestFirst, que busca en el espacio de subconjuntos de atributos mediante un algoritmo *greedy*[10] (Brassard, et al., 1997) *hill climbing*[11] (Russell, et al., 2003) aumentado con facilidad de *backpropagation*[12] (Rojas., 1996).

Como método de evaluación se utilizó el CFsSubsetEval que evalúa el valor de un subconjunto de atributos teniendo en cuenta la capacidad individual de predicción de cada característica, junto con el grado de redundancia entre ellos (Bouckaert, 2009).

La tabla 2 muestra los resultados de la aplicación de ambos algoritmos con validación cruzada de diez ciclos, mostrando el nivel de influencia de cada variable independiente sobre las dependientes.

Tabla 2: Análisis de correlaciones.

Variables	Calidad	Rendimiento
PH del suelo	0%	0%
Humedad del suelo	0%	0%

[9] overfitting: Sobre aprendizaje del modelo matemático.
[10] greedy: Algoritmo ávido que sigue una heurística basada en la selección de óptimos locales.
[11] hill climbing: Método de búsqueda iterativo basado en una heurística para mejorar la eficiencia.
[12] backpropagation: Algoritmo de ajuste del peso de los nodos para reducir el error.

Salinidad del suelo	0%	0%
Profundidad del suelo	100%	0%
Población	-	-
Estado de las semillas	10%	80%
Período de siembra	0%	0%
Período de recolección	0%	30%
Forma de siembra	20%	20%
Forma de recolección	0%	0%
Temperatura promedio durante el día	0%	20%
Temperatura promedio durante la noche	0%	70%
Personal	0%	100%
Riego	90%	0%
Nitrógeno	50%	0%
Fosforo	0%	10%
Potasio	0%	0%
Calcio	10%	50%
Magnesio	0%	0%
Sodio	0%	0%
Azufre	0%	0%
Boro	10%	0%
Cobalto	0%	40%

Cobre	10%	0%
Hierro	0%	0%
Manganeso	10%	0%
Molibdeno	20%	0%
Zinc	50%	0%

El análisis de correlaciones permitió descartar del modelo todas las variables que no tenían ningún nivel de influencia. Como resultante se obtuvieron dos conjuntos que, aunque poseen elementos comunes, son divergentes y caracterizan a diferentes variables.

3.2. Transformaciones realizadas

A las variables seleccionadas se le aplicaron diversas transformaciones teniendo en cuenta sus características. Las variables nominales como el estado de las semillas y la forma de siembra se transformaron a binario, con una cantidad de bits igual a los posibles estados, según el estado que tenga será el bit activo, para el resto de los estados sus respectivos bits tendrán valor cero.

Las variables de valor numérico se normalizaron en un intervalo entre cero y uno. Las variables de tipo fecha se transformaron al valor numérico, correspondiente a la cantidad de días sin tener en cuenta el valor del año, de esta forma se logra un énfasis en los datos correspondientes a la época del año en que son sembrados y recolectados**.**

3.3. Matriz de confusión

La matriz de confusión es uno los datos de salida de los algoritmos al ser procesados las variables de tipo nominal en el Weka. Es una matiz cuadrada que describe el proceso de clasificación en función de los valores que pueden tomar la variables como se muestra en la Figura 6.

```
a  b  c   <-- classified as
8  0  3 |  a = buena
0  3  0 |  b = mala
0  0  6 |  c = regular
```

Figura 6: Matriz de confusión.

La suma de todos los valores de la matriz es la cantidad de casos de la muestra, la Figura 6 expone una muestra de 20 casos diferentes. Los datos de la matriz son clasificados por columnas en: buena, mala o regular.

Los casos clasificados correctamente son los que se encuentran en la diagonal principal, sin embargo se puede apreciar una incorrecta clasificación de 3 casos, aparecen clasificados como C (regular) cuando en realidad pertenecen a la clase A (buena), estos casos son un error en el que incurrió el algoritmo. En base a estos errores es que se determina el nivel de acierto del algoritmo.

3.4. Algoritmos propuestos para los modelos matemáticos

Al aplicar el ZeroR a la muestra teniendo como objetivo la variable Rendimiento se obtuvo como resultado el valor de la media de 252.17, con un coeficiente de correlación de -0.36. Al aplicar ZeroR teniendo la variable Calidad como objetivo el resultado es que clasifica correctamente al 40% de la muestra lo que nos da un punto de partida a superar.

Teniendo en cuenta las características del Weka, los algoritmos solo pueden tener una variable dependiente como objetivo, sin importar la cantidad de atributos que describan a esta variable, por tanto los algoritmos son probados por cada variable dependiente. También se tiene en cuenta las características de los algoritmos, como las variables dependientes son de diferente tipo, no siempre es posible la aplicación de un mismo algoritmo, porque alguno no soportan datos nominales o numéricos como objetivo.

Por cada conjunto de algoritmos se realizó la prueba a un elemento como representante del conjunto, este proceso se efectuó para cada una de las variables dependientes. Cuando el algoritmo no podía ser probado con la otra variable, entonces se tomaba otro del dominio en su lugar. Todos los algoritmos utilizados fueron probados usando una validación cruzada de diez ciclos.

Fue de gran interés la clasificación correcta de las cosechas en los valores regular y bien, correspondientes a la variable nominal Calidad; analizando el comportamiento de los datos desde el punto de vista fenomenológico, las probabilidades de mala de calidad en las cosechas es muy baja.

Existen normas establecidas para el trabajo con los cultivos de cebolla que son exigidas, aspecto considerado en el momento de seleccionar el algoritmo más adecuado.

3.4.1. Basados en reglas

Los algoritmos basados en reglas son deterministas y constituyen la técnica más sencilla utilizada en la extracción del conocimiento. Se aplican cuando la complejidad del problema es alta y por lo general es necesario emplear aproximaciones. Tienen como base un conjunto de reglas que son definidas al construirse el clasificador con el conjunto de entrenamiento, se realiza utilizando diversos métodos, cada algoritmo usa métodos de construcción diferentes. Al final se obtiene un resultado similar para dar solución al problema, es construido un conjunto de reglas SI A ENTONCES B, donde A es una condición determinada y B un resultado asociado a dicha condición.

De este conjunto de algoritmos fue probado el ConjunctiveRule para ambas variables dependientes, y es explicado más detalladamente en el epígrafe 2.7

```
 a  b  c   <-- classified as
40  0  0 |  a = buena
 0 30  0 |  b = mala
28  2  0 |  c = regular
```

Figura 7: Matriz de confusión algoritmo ConjunctiveRule.

Como resultado de la aplicación del algoritmo se alcanzó la clasificación correcta del 70% de la muestra. El algoritmo no pudo clasificar correctamente los elementos regulares en la muestra (Ver Figura 7), esto resulta un aspecto negativo a valorar, por ser una de las clases de mayor interés en el proceso de clasificación para esta variable. Al aplicar el mismo algoritmo para la variable Rendimiento el coeficiente de correlación obtenido fue de 0.79, el error absoluto medio fue 62.6%, el error cuadrático relativo fue 74.2%, el error relativo absoluto fue 59.1% y el error cuadrático relativo fue 60%.

3.4.2. Arboles de decisión

Los arboles de decisión son algoritmos de rápido aprendizaje y después de ser construidos son recorridos rápidamente. Para la variable Calidad se utilizó el algoritmo J48graft.

```
 a  b  c   <-- classified as
40  0  0 |  a = buena
 0 30  0 |  b = mala
 3  0 27 |  c = regular
```

Figura 8: Matriz de confusión algoritmo J48graft.

Al ser probado utilizando la variable Calidad como objetivo se realizó una correcta clasificación del 96% de la muestra, la matriz de confusión se representa en la Figura 8. Como se puede apreciar el algoritmo presenta dificultad para clasificar los elementos pertenecientes a la clase de valor regular.

Para la variable Rendimiento se utilizó el algoritmo DecisionStump debido a que el J48graft solo puede ser utilizado en clasificación. De su aplicación se obtuvo un coeficiente de correlación de 0.78, el error absoluto medio fue 61.8%, el error cuadrático relativo fue 75.3%, el error relativo absoluto fue 58.4% y el error cuadrático relativo fue 60.8%.

3.4.3. Red neuronal Perceptrón Multicapas

Este algoritmo es capaz de procesar variables de tipo nominal y numérica, fue probado para ambas variables dependientes del problema.

La arquitectura inicial de la red para la variable Calidad fue:

- Cantidad de capas ocultas: 11
- Cantidad de neuronas por capa: 11
- Velocidad de aprendizaje: 0,3
- Impulso: 0.2
- Función de activación: Sigmoide
- Función de salida: y=x

La arquitectura inicial de la red para la variable Rendimiento fue:

- Cantidad de capas ocultas: 13
- Cantidad de neuronas por capa: 13
- Velocidad de aprendizaje: 0,3
- Impulso: 0.2

- Función de activación: Sigmoide
- Función de salida: y=x

```
 a  b  c   <-- classified as
37  0  3 |  a = buena
 0 26  4 |  b = mala
 4  2 24 |  c = regular
```

Figura 9: Matriz de confusión algoritmo Perceptrón Multicapas.

En la aplicación inicial del algoritmo para la variable Calidad obtuvo un nivel de acierto del 87.1%, la matriz de confusión se muestra en la Figura 9. Para la variable Rendimiento se obtuvo un coeficiente de correlación de 0.97.

```
 a  b  c   <-- classified as
38  0  2 |  a = buena
 0 27  3 |  b = mala
 4  1 25 |  c = regular
```

Figura 10: Matriz de confusión algoritmo Perceptrón Multicapas ajustado.

Después de realizar varios ajustes al algoritmo para ambas variables se obtuvieron mejoras en su aplicación. Se obtuvo un nivel de acierto del 90%, un 3% superior a su aplicación anterior, el resultado se evidencia en la matriz de confusión mostrada en la Figura 10.

Al ajustarlo para la variable Rendimiento se alcanzó un coeficiente de correlación de 0.98, mejorando el resultado obtenido anteriormente, y con un error absoluto medio del 61.8%, un error cuadrático relativo del 75.3%, un error relativo absoluto del 58.4% y un error cuadrático relativo del 60.8%.

Después de los ajustes realizados la arquitectura de la red neuronal para la variable Calidad quedo de la siguiente forma:

- Cantidad de capas ocultas: 10
- Cantidad de neuronas por capa: 11
- Velocidad de aprendizaje: 0.1
- Impulso: 0.6
- Función de activación: Sigmoide

- Función de salida: y=x

Para la variable Rendimiento después de los ajustes realizados quedo de la forma siguiente:

- Cantidad de capas ocultas: 2
- Cantidad de neuronas por capa: 13
- Velocidad de aprendizaje: 0.1
- Impulso: 0.4
- Función de activación: Sigmoide
- Función de salida: y=x

3.4.4. Máquinas de soporte vectorial

El algoritmo de máquinas de soporte vectorial, solo puede ser aplicado a variables nominales, se destinó únicamente a la variable Calidad.

```
 a  b  c   <-- classified as
38  0  3 |  a = buena
 0 27  3 |  b = mala
 3  3 24 |  c = regular
```

Figura 11: Matriz de confusión algoritmo Maquinas de Soporte Vectorial.

Después de aplicado en algoritmos se obtuvo un nivel de acierto de 88.1% (Ver Figura 11), se apreciaron dificultades para clasificar los elementos pertenecientes a todas las clases.

3.5. Comparación de los algoritmos

Para definir cuál era el algoritmo más adecuado a aplicar fueron examinados varios factores, para el caso de la variable Calidad, el nivel de precisión del algoritmo respecto a la muestra, en la matriz de confusión la correcta clasificación de las clases buena y regular. Para la variable Rendimiento el factor más importante era el coeficiente de correlación, aunque se tomaron en cuenta otros factores.

Tabla 3: Comparación entre algoritmos.

Algoritmos	Calidad	Rendimiento
ZeroR	40%	-0.36
ConjunctiveRule	70%	0.79
J48graft	96%	-
DecisionStump	-	0.78
Perceptrón Multicapas	90%	0.98
Máquina de soporte vectorial	88.1%	-

La Tabla 3 compara los resultados obtenidos con la aplicación de los diferentes algoritmos, proporcionó la base para realizar el proceso de selección. Se decidió, finalmente, seleccionar el árbol de decisión J45graft para la variable Calidad, por tener el mejor índice de acierto y obtener buenos resultados en la clasificación de las clases “buena” y “regular”.

Tabla 4: Comparación entre algoritmos de regresion

Algoritmos	ConjunctiveRule	DecisionStump	Perceptrón Multicapas
Coeficiente de correlación	0.79	0.78	0.98
Error absoluto medio	62.6	61.8	16.1
Error cuadrático medio	74.2	75.3	20.3
Error relativo absoluto	59.1	58.4	15.2

Error cuadrático relativo	60.0	60.8	16.4

Aunque el Perceptrón Multicapas presenta el coeficiente de correlación más alto también se tomaron en cuenta otros aspectos, en la Tabla 4 es mostrada una comparación entre los algoritmos probados. Basándose en la tabla de comparación y en los resultados obtenidos, se seleccionó para la variable Rendimiento el algoritmo Perceptrón Multicapas porque, en cada uno de los criterios de comparación alcanza mejores resultados que obtenidos por los otros algoritmos.

3.6. Descripción del modelo final

El modelo propuesto tuvo como punto de partida la aplicación de un pre-procesamiento a los datos, de los cuales se seleccionaron solo las variables independientes influyentes en las variables dependientes del problema. Se les aplicaron transformaciones a los valores de las variables para mejorar los resultados obtenidos de la aplicación de los algoritmos.

Para clasificar los parámetros de entrada en función de la variable Calidad se aplicó un árbol de decisión J45graft, que permitió realizar una correcta clasificación en el 96% de la muestra.

En la predicción del valor de la variable Rendimiento se usó la red neuronal Perceptrón Multicapas, porque aportó un coeficiente de correlación de 0.98 y minimizo el error. El modelo matemático obtenido clasifica como modelo empírico por los tipos de algoritmos empleados.

3.7. Prueba del modelo seleccionado

Para probar el modelo obtenido se desarrolló un prototipo de aplicación, utilizando para ello el lenguaje C++, algunas funcionalidades de acceso a ficheros y trabajos con cadenas del Framework QT y la biblioteca Waffles que tiene implementado los algoritmos utilizados en el modelo, liberando a la aplicación de las limitaciones del Weka como software académico.

Waffles es una colección de herramientas de línea de comandos para los investigadores de aprendizaje automático, minería de datos, y otros campos relacionados (Gashler, et al.). Las relaciones entre la librería, el Famework y la aplicación son representadas en el diagrama de clases de la Figura 12.

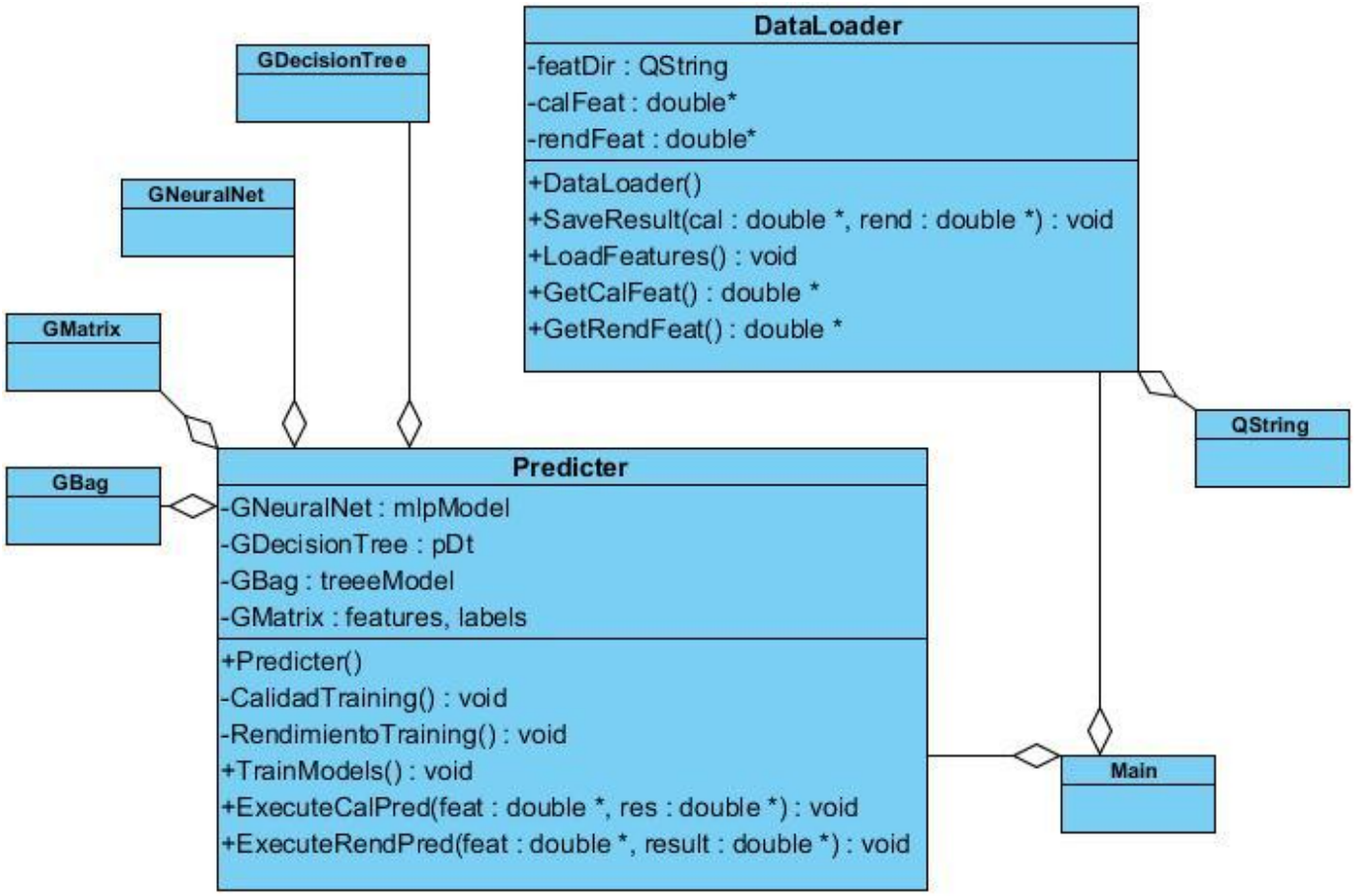

Figura 12: Diagrama de clases de la aplicación.

Las funcionalidades de la aplicación fueron separadas en dos clases fundamentales, Predicter y DataLoader. La clase DataLoader es responsable del trabajo con ficheros apoyándose en las funcionalidades que brindan las clases del Framewor Qt.

Mediante el método LoadFeatures se seleccionó del conjunto de variables independientes de entrada, en un fichero, las más influyentes definidas en el epígrafe 4.3, y las separó para cada variable dependiente, además se utilizó el método SaveResult guarda el resultado obtenido en un fichero de salida.

El modelo fue implementado en la clase Predicter, para ello se utilizó las clases de la librería Waffles. Mediante el método TrainModels se entrena el modelo de cada variable con un conjunto de datos de entrenamiento guardados en un fichero arff. Con el empleo de los métodos ExecuteCalPred y ExecuteRendPred se ejecutan los modelos para las variables Calidad y Rendimiento respectivamente.

Concluida la aplicación se realizaron pruebas, sobre la base de tres juegos diferentes de datos, con el objetivo de observar el comportamiento del modelo ante estos valores de las variables de entrada. En las Tablas 5 y 7 se muestran los valores tomados para las variables de entrada.

Tabla 5: Prueba realizada con el modelo para la variable Calidad

Variables	Prueba 1	Prueba 2	Prueba 3
Profundidad suelo	11 cm	14 cm	18 cm
Estado semillas	Bueno	Malo	Malo
Forma siembra	Directa	Bulbillos	Bulbillos
Riego	3500 m3/Ha	2600 m3/Ha	2100 m3/Ha
Abono/Fertilizante	Urea	Superfosfato triple	Fosfato de amonio
Calidad	**Buena**	**Mala**	**Regular**

Tabla 6: Resultados de la prueba.

Calidad	% Acierto Prueba 1	% Acierto Prueba 2	%Acierto Prueba 3
Buena	0,58%	0,14%	0,30%
Mala	0,23%	0,50%	0,25%
Regular	0,17%	0,34%	0,45%

Los resultados de las pruebas realizadas a la variable Calidad se muestran en la Tabla 6. Después de aplicar el algoritmo el resultado obtenido es un índice probabilístico para cada una de las posibles clases, se tomándose como valor final la clase que tenga mayor nivel de probabilidad.

Tabla 7: Prueba realizada con el modelo para la variable Rendimiento.

Variables	Prueba 1	Prueba 2	Prueba 3
Estado semillas	Bueno	Malo	Malo
Fecha de recolección	4/Dic	12/Ene	13/Feb
Forma siembra	Directa	Bulbillos	Bulbillos
Temperatura día/noche	25,9 / 14,1	24,5 / 20,1	35 / 14,9
Personal	5	1	5
Abono/Fertilizante	Urea	Superfosfato triple	Fosfato de amonio
Rendimiento	**411**	**67**	**354**

Tabla 8: Resultados de la prueba.

Rendimiento	Resultado	Real	Desviación
Prueba 1	411	425	3,1%
Prueba 2	67	61	9,8%
Prueba 3	354	382	7,2%

En la Tabla 8 se muestran los resultados obtenidos de aplicar la red neuronal Perceptrón Multicapas, así como el nivel de desviación correspondiente a cada una de las pruebas realizadas. Los resultados

obtenidos fueron satisfactorios pero el modelo, en algunos casos pudo ser afectado por cuestiones de azar. Esta afectación del azar es producida por inicializar los nodos de las estructuras formadas por los algoritmos con valores aleatorios. Para determinar la posible afectación del azar sobre el modelo, se realizaron pruebas sin modificar los valores de las variables, con el objetivo de obtener la diferencia entre los resultados.

Tabla 9: Resultados de las pruebas con la aplicación.

Variables	Iteración 1	Iteración 2	Iteración 3	Iteración 4	Iteración 5	Media
Calidad	Regular	Regular	Mala	Regular	Regular	Regular
Buena	0.30%	0.20%	0.02%	0.16%	0.26%	0.18%
Mala	0.25%	0.24%	0.45%	0.23%	0.17%	0.26%
Regular	0.45%	0.54%	0.43%	0.59%	0.55%	0.51%
Rendimiento	354.7	374.5	357.9	396.7	384.0	373.5
Desviación	7.20%	2.07%	6.40%	3.80%	0.52%	2.30%

Los resultados observados en la Tabla 9 fueron obtenidos usando los parámetros de la Prueba 3 mostrados en la Tabla 5 y 7. Como se puede apreciar los valores varían de una iteración a la siguiente, a pesar de que presentan alteraciones en algunas de ellas, como se puede apreciar una mala clasificación en la Iteración 3 y la desviación más alta en la Iteración 1, en general tienden a un valor más confiable. Para atenuar la afectación del azar y reducir los posibles errores, basados en los resultados de las pruebas realizadas, se ejecutó el modelo diez veces y se dio como salida la media obtenida, que es una valor de tendencia más próximo al valor real, y con un error más atenuado.

3.8. Resultados alcanzados

En el trabajo se realizó un análisis de correlaciones y se determinaron cuáles eran las variables más influyentes en los modelos, dándole simplicidad a estos.

Se seleccionaron los algoritmos de mejor resultados, basándose en diversos criterios de comparación, quedando para la variable Calidad el algoritmo J48graft y para la variable Rendimiento la red neuronal Perceptrón Multicapas.

La aplicación de un árbol de decisión, para los parámetros de entrada en función de la variable Calidad permitió realizar una correcta clasificación en el 96% de la muestra.

El uso de la red neuronal Perceptrón Multicapas fue muy efectivo en la predicción del valor de la variable Rendimiento, porque aportó un coeficiente de correlación de 0.98.

Se desarrolló un modelo matemático ajustado al problema planteado capaz de predecir el comportamiento de las variables Calidad y Rendimiento.

Las pruebas efectuadas permitieron discriminar los modelos teleonómicos sugeridos por la literatura y demostraron el desempeño eficiente de los modelos empíricos para la solución del problema estudiado.

Se desarrolló una aplicación con los modelos implementados la cual fue probada con varios juegos de datos para verificar su efectividad (Tablas de la 5 - 9), además los resultados expresados en la Tabla 9 permitieron demostrar que se logró mejorar la calidad del resultado en la aplicación. El prototipo de aplicación desarrollado independiente de la herramienta de trabajo es capaz de realizar todo el proceso descrito y reducir el nivel de error, aproximadamente de un 4% en el caso peor, del modelo. seleccionado.

Conclusiones

Luego de culminada la investigación se arribaron a las siguientes conclusiones:

- El modelo matemático obtenido puede ser aplicado con efectividad a otros cultivos de la familia de las Liliaceae.
- El modelo desplegado se ajusta al problema planteado, es capaz de predecir el comportamiento de la Calidad con un 96% de acierto y el valor del Rendimiento con un coeficiente de correlación de 0.98.
- Los resultados obtenidos demostraron la efectividad de los modelos empíricos para la solución del caso estudiado.
- La realización e implementación de una aplicación demostrativa fue válida, permitió observar y atenuar los errores producidos por el azar en la aplicación de los algoritmos, y proporcionó la posibilidad de examinar el comportamiento de los algoritmos fuera del software de prueba.
- El modelo matemático de carácter empírico obtenido permite predecir con una efectividad superior al 90% el comportamiento de los datos en el sistema DSerp Agro en relación con el cultivo de cebolla.

Recomendaciones

Teniendo como punto de partida los resultados obtenidos se recomienda:

- Ajustar el modelo empleando un mayor volumen de información en la base de datos.
- Probar la aplicación del modelo en otras especies de la familia Liliaceae.
- Convertir el estudio realizado en un módulo para ser integrado en el software DSerp Agro.

Índice de tablas y figuras

Bibliografía

A. Acosta, J. Gaveola. 1989. *Manual de Producción de Semilla de Cebolla.* Santiago de Chile : FAO, 1989.

Bigus, J. 1996. *Data Mining with neural networks.* USA : Mc GrawHill, 1996.

Bouckaert, Remco R. 2009. *Manual WEKA* . Hamilton : s.n., 2009.

Brassard, Gilles and Bratley, Paul. 1997. *Fundamentos de Algoritmia.* Madrid : Prentice Hall, 1997.

David Hand, Heikki Mannila, Padhraic Smyth. 2003. *Principles of Data Mining.* s.l. : The MIT Press, 2003.

Erlbaum, Lawrence. 2003. *THE HANDBOOK OF DATA MINING.* Arizona : Nong Ye, 2003.

Gashler, Mike and Moyer, Eric. Waffles. [Online] [Cited: 1 14, 2012.] http://waffles.sourceforge.net/.

Giaconi, V., Escaff , M. 2004. *Cultivo de hortalizas.Decimoquinta edición.* Santiago de Chile : Universitaria, 2004.

Goudriaan, J. and Laar, H. Van. 1994. *Modelling Potential Crop Growth Processes.* Dordretch, Netherland : Kluwer Academic Publishers, 1994.

Han, Jiawei and Kamber, Micheline. 2006. *Data Mining: Concepts and Techniques.* Boston : Morgan Kaufmann, 2006.

Ian H. Witten, Eibe Frank. 2005. *Data Mining Practical Machine Learning Tools and Techniques.* Hamilton : Morgan Kaufmann, 2005.

J. Hernandez Orallo, M.J. Ramirez Quintana, C.Ferri Ramirez. 2004. *Introducción a la Minería de Datos.* s.l. : Pearson Prentice Hall, 2004.

LAROSE, DANIEL T. 2005. *Discovering Knowledge in Data An Introduction to Data Mining.* Connecticut : A JOHN WILEY & SONS, 2005.

Lipo Wang, Xiuju Fu. 2005. *Data Mining with Computational Intelligence.* New York : Springer, 2005.

M. Berthold, D.J. Hand. 1999. *Intelligent Data Analysis: An introduction.* s.l. : Springer, 1999.

Marato, J. V. 1992. *Horticultura herbacia especial.3ra Edición.* Madrid : Mundi Prensa, 1992.

Mitchell, Tom M. 2003. *Machine Learning.* s.l. : Prentice Hall, 2003.

MODELOS DE SIMULACIÓN DE CULTIVOS. CARACTERÍSTICAS Y USOS. **Hernández, Naivy, Soto, F. and Caballero, A. 2009.** 1, La Habana, Cuba : Redalyc, 2009, Vol. 30.

Modelos matemáticos de hortalizas en invernadero: Trascendiendo la contemplación de la dinámica de los cultivos. **Cruz, I. L. Lopez. 2005.** 002, Chapingo : Redalyc, 2005, Vol. 11.

MOLINA, L.C. y RIBEIRO, S. 2001. *Descubrimiento conocimiento para el mejoramiento bovino usando técnicas de data mining.* Barcelona : Actas del IV Congreso Catalán de Inteligencia Artificial, 2001.

Molina, Luis Carlos. Data mining: torturando a los datos hasta que confiesen. [Online] [Cited: 1 20, 2012.] http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html.

Muñoz de Con, L., Prats Pérez, A. Brito Iglesias. 1991. *La Técnica de producción de Semilla de Cebolla.* Ciudad de la Habana : CIDA, 1991.

Nikhil R. Pal, Lakhmi Jain. 2004. *Advanced Techniques in Knowledge Discovery and Data Mining.* Xindong Wu : Lakhmi Jain, 2004.

Orozco, E. 2009. *Inteligencia Empresarial Qué y Como.* La Habana, Cuba. : IDICT, 2009.

Pérez Domínguez, C. 1988. *Cebolla, ajo y especies Afines: Conózcalas mejor*. Ciudad de La Habana : CIDA, 1988.

Piatetsky-Shapiro, G. KDnuggets. [Online] [Cited: 11 30, 2011.] http://www.kdnuggets.com.

Quinlan, J. R. 1993. *C4.5: Programs for Machine Learning.* s.l. : Morgan Kaufmann Publishers, 1993.

Rojas., R. 1996. *Neural Networks: A Systematic Introduction.* s.l. : Springer, 1996.

Russell, Stuart J. and Norvig, Peter. 2003. *Artificial Intelligence: A Modern Approach (2nd ed.).* New Jersey : Prentice Hall, 2003.

Smola., C. Burges B. Schölkopf y A. 1999. *Advances in kernel methods: Support vector machines.* Cambridge : MIT Press, 1999.

T. Hastie, R. Tibshirani, J. Friedman. 2001. *The elements of Statistical Learning.* s.l. : Springer, 2001.

Taniar, David. 2007. *Data Mining and Knowledge Discovery Technologies.* Monash : IGI PublIshInG, 2007.

Thornley, J. H. M. and Johnson, I. R. 2000. *Plant and crop modeling, A Mathematical Aproach to Plant and Crop Physiology.* New Jersey, USA : The BLackburn Press, 2000.

U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. 1996. *Advances in knowledge Discovery and Data Mining.* Cambridge : MIT Press, 1996.

Vapnik., V.Ñ. 1995. *The nature of statistical learning theory.* New York : Springer-Verlag, 1995.

VISSER, C. L. M. DE. 1994. *ALCEPAS, an onion growth model based on SUCROS87: Devlopment of the model.* s.l. : Journal of Horticultural Science, 1994.

Waikato, University. WEKA Wiki. [Online] http://weka.wikispaces.com/.

White., Colin J. 2001. *IBM Enterprise Analytics for the Intelligent e-Business.* USA : IBM Press, 2001.

ZhaoHui Tang, Jamie MacLennan. 2005. *Data Mining with SQL Server 2005* . Indianapolis : Wiley Publishing, 2005.