VÍNCULOS
ENERO-JUNIO DE 2011
VOLUMEN 8 ▪ NÚMERO 1

Aplicación de técnicas de minería de datos en la construcción de un inventario de maguey papalote, en el estado de Guerrero[1]

*Application of data mining techniques to build an inventory of maguey kite in the state of Guerrero**

Lorena Alonso, Dante Covarrubias***, Juan Carlos Medina****

Fecha de recepción: 15 de abril del 2011
Fecha de aceptación: 16 de junio del 2011

Líneas de investigación
H.2 [Minería de datos]: Muestreo espacial. Estadística multivariada.

Resumen

En este trabajo, se presenta un comparativo de técnicas del manejo de datos aplicadas para determinar el patrón espacial del maguey papalote, destacando el uso de minería de datos como ténica de análisis de datos, ya que, con la construcciónón de redes de interacción biótica, permite representar una herramienta potente que permita predecir factores para la producción y del eclinamiento del maguey papalote. Otra técnica es el análisis de componentes principales, la cual es una técnica en estadística multivariada muy utilizada para reducir variables, sin la pérdida de informaciónón; los resultados de la investigación son utilizados en Desarrollo de un sistema de inventario y monitoreo de maguey papalote (Agave cupreata Trel & Berger)" (SIMMP), en la aplicaciónón de las dos técnicas en la elaboración de un inventario de maguey papalote en el estado de Guerreo.

Palabras clave: framewors, funcionalidad, algoritmos, usuario, estrategias de muestreo, SIG y ecología del agave.

* Otras versiones de este documento se encuentran en: 2do encuentro Iberoamericano de Biometría, (mayo de l2010). Inventario de maguey papalote de la región centro del Estado de Guerrero. Boca de Rio, Veracruz, México.www.uv.mx/eib. 1er Encuentro Internacional del Medio Ambiente (noviembre del 2010). Muestreo por conjuntos Ordenados (Ranked Set Sampling) y su aplicación en Poblaciones de maguey silvestre. Puebla, Puebla; México. Recuperado de: www.eima.todoencomputacion.htm. Primer Congreso Internacional de Computación UDA México-Colombia XII Jornada Académica en Inteligencia Artificial.

** Centro de Investigación y Desarrollo Tecnológico Unidad Académica de Matemáticas, Av. Lazaro Cárdenas, C.U., Guerrero, México. Correo electrónico: alonso@uagro.mx.

*** Universidad Autónoma de Guerrero Unidad Académica de Matemáticas, Av. Lazaro Cárdenas, C.U., Guerrero, México. Correo electrónico: dcova@uagro.mx.

**** Universidad Autónoma de Guerrero, Unidad Académica de Ingeniería, Av. Lazaro Cárdenas S/N, C.U., Guerrero, México. Correo electrónico: jcmedina74@yahoo.com.mx.

Introducción

Una preocupaciónón creciente en los estados del sur de México es la producción de maguey papalote (de origen silvestre), producto esencial para la fabricación del mezcal, bebida típica de la región y representante de México en el mundo. Por esto, la necesidad de conocer el número de cabezas de maguey en el estado de Guerrero dio pie a la búsqueda de modelos matemáticos que permitan estimar dicho número, pero, más aún, conocer las variables que mejor expliquen su comportamiento en los diferentes tipos de modelos; para la caracterización de este comportamiento, los investigadores realizan varias mediciones a la población del maguey papalote, lo cual, normalmente, genera una gran cantidad de datos, lo que nos lleva a implementar un enfoque de análisis de datos, geográficos basados en la aplicación de inteligencia artificial, en combinación con enfoques estadísticos multivariados y de simulación numérica. Esta unión de métodos se conoce como minería de datos, datamining o aprendizaje estadístico (Hastie, Tibshirani y Friedman, 2001), que forma parte de un proceso mayor denominado "descubrimiento de conocimiento", en el que se busca generar conocimiento mediante el aprendizaje de patrones presentes en los datos (Vallejos, 2006). La minería de datos (MD) tiene como paradigma de análisis el trabajo con un gran volumen de datos observacionales y su exploración sistemática utilizando enfoques diversos, ya que a priori se desconocen tanto las características de la distribución de valores, como los métodos que puedan resultar adecuados y, en particular, permiten analizar relaciones no lineales entre variables, sin embargo, se debe tener presente que debe ser un proceso controlado y planeado, en el que se deben tomar en cuenta las necesidades y las sugerencias de los usuarios o clientes quienes, finalmente, son los que están relacionados con el funcionamiento de la ecología del maguey papalote (Medina, s.f.; Vallejos, 2006; Alonso, Covarrubias y Maradiaga 2009).

Es importante mencionar que la MD es una herramienta particularmente apropiada para generar hipótesis cuando no hay información disponible a priori o esta es muy limitada. En análisis espacial, el procesamiento digital de imágenes utiliza enfoques de minería de datos desde hace mucho tiempo, pero en aplicaciones de inventarios ecológicos ha cobrado interés recientemente, destacándose su utilización para estudios de biodiversidad, invasiones biológicas y modelización de distribuciones de especies (Stockwell y Peters, 1999, pp. 143-158; White y Sifneos, 2002, pp. 600-614; Lawler, White, Sifneos y. Master, 2003, pp. 875-882).

En el análisis de componente principales (ACP), el cual es una técnica estadística de síntesis de la información o reducción de la dimensión (numero de variables), cuyo objetivo principal es reducir a un menor número perdiendo la menor cantidad de información posible. Los nuevos componentes principales o factores serán una combinación lineal de las variables originales y, además, serán independientes entre sí.

Así al elegir un método adecuado, como es la minería de datos y el análisis de componentes principales, para el análisis de datos es de enorme ayuda para estos estudios de ecología y biodiversidad el comportamiento de la población, ya que los productores de mezcal requieren información de calidad en escalas múltiples para determinar como está cambiando el estado del recurso maguey presentes en algún hábitat y lograr con esto un mejor entendimiento del comportamiento (Maradiaga, 2004; Medina, s.f.; Alonso, Covarrubias y Maradiaga, 2009).

Planteamiento del problema

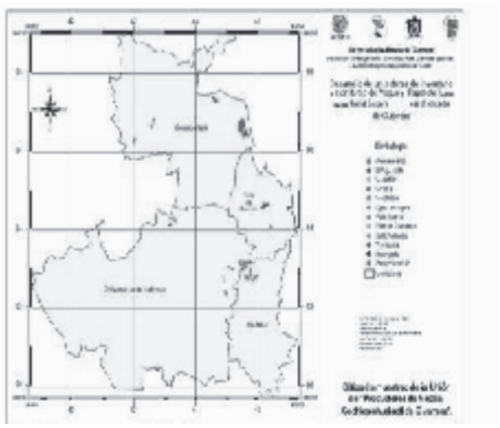
Hasta hace pocos años, la industria del mezcal se desarrolló bajo el supuesto de que disponían de un recurso autr-renovable. Sin embargo, poco a poco, los productores de mezcal se han percatado de la declinación de las poblaciones naturales de maguey papalote (Agave cupreata Trel. & Berger) y, por tanto, han adquirido conciencia de la necesidad de implementar algunos proyectos que permitan recuperar áreas y poblaciones de maguey papalote, que garanticen la sustentabilidad de la materia prima en la industria (Maradiaga, 2004; Alonso, Covarrubias y Maradiaga, 2009).

Una parte importante son los componentes principales que requiere una aplicación para un inventario ambiental de maguey papalote, en el que se tenga un registro de los productores, así como de los predios en donde hay maguey silvestre y reforestado, con información sobre el año en que se sembró y cuánto se sembró. Esta información se tomará de un sistema de información geógrafa (SIG) y se realizarán los cálculos de existencias de materia prima, que serán la base, junto con los registros que se llevan en cada fábrica mezcalera y en las asambleas de planeación de temporada, para proceder a una buena planeación de análisis de datos, por lo cual estos sistemas deben garantizar la contabilidad de los datos y de la información producida (Maradiaga, 2004; Medina, s.f.; Moreno, Quintales, Penalvo y Martin, 2006).

Materiales y métodos

Datos: el maguey papalote (Agave cupreata Trel. & Berger) es una especie típicamente de la cuenca del río Balsas, concentrándose en la porción correspondiente a la vertiente interior de las sierras Madre del sur y norte en el estado de Guerrero, aunque también ha sido colectada en el estado de Michoacán y se estima su presencia en áreas aledañas de los estados de Puebla, Oaxaca y Morelos, en México (Maradiaga, 2004).

Figura 1. Mapa de la unión de productores de mezcal xochicalehuactl de Guerrero
Fuente: Maradiaga (2004).



A la fecha, solo existe un inventario estatal detallado y de cobertura estatal completa (Maradiaga, 2004), que puede ser tomado como base, el reporte técnico cuyo nombre es *Desarrollo de un sistema de inventario y monitoreo de maguey papalote (Agave cupreata Trel. & Berger)*, (SIMMP), en el 2004, a fin de realizar una simulación partiendo de datos auténticos.

Dicho inventario consideró seis etapas de desarrollo del maguey papalote: plántula, juvenil medio, juvenil mayor las siguientes tres categorías corresponden a etapas en las cuales la planta ha entrado en un proceso de maduración, momento en el puede ser preparada para aprovecharse o, por el contrario, llegarán a la floración-fructificación. La primera de ellas, Velilla, calehual y Capón. Para este estudio se considero como clase objetivo solo la categoría "maduración" propiamente dicha y como situación de ausencia a la clase "no madura".

Variables predictoras

En el presente trabajo se propone la utilización de análisis de componentes principales y técnicas de minería de datos, los árboles de decisión, para obtener un conjunto de reglas que conduzca a una clasificación de especie eficiente y que, a su vez, permita reducir la complejidad computacional requerida en la clasificación. La reducción de complejidad computacional se da por medio de la reducción de atributos requeridos para realizar la clasificación que se produce mediante la utilización de los árboles de decisión –atributo que no aporta nada a la clasificación, no es considerado en el árbol– y reducir factores al observar la relación entre variables iníciales.
Esta investigación retoma la base de datos proveniente del reporte técnico (SIMMP), figura 1. proporcionada por el Instituto de Investigación Científica Área Ciencias Naturales de la Universidad Autónoma de Guerrero. La base de datos presenta un detalle espacial adecuado para caracterizaciones regionales del orden del 1: 5000. Dicha recopilación de variables se presenta en la tabla 2 y abarca variaciones geomorfológico-topográficas y de cobertura de la vegetación, que se espera sean indicadoras de diferencias en procesos de generación y mantenimiento de las áreas potenciales para la producción magueyera en el estado de Guerrero.

Objetivo

El objetivo fundamental de este trabajo es proporcionar a los investigadores patrones ambientales, para con ellas en fases posteriores, puedan obtener indicadores ambientales y de sostenibilidad del maguey paplote.

Organización del documento

La organización de este documento está conformada por cinco apartados: en el primero, se cuenta con información básica sobre las técnicas aplicadas análisis de componentes principales y árbol de decisión; en el siguiente, se describe la aplicación y la simulación del estudio; de igual forma, en el que le sigue. Más adelante, siguen las conclusiones de la investigación, las referencias bibliográficas y, por último, las tablas analizadas en esta investigación de los dos métodos desarrollados.

Técnicas aplicadas: componentes principales y árboles de decisión

En minería de datos (Vallejos, 2006), se utilizan varias técnicas de clasificación algunas de ellas son los arboles de decisiones, clúster, redes neuronales y métodos estadístico. Los árboles de clasificación o de decisión se caracterizan por su sencillez, su campo de acción abarca diversas áreas.
En minería de datos, la tarea de clasificación tiene como objetivo predecir la asignación de una instancia a una clase dada a partir de una muestra con ejemplos positivos y negativos de esta, y por eso también se le denomina aprendizaje supervisado. Los principales métodos de clasificación son:

- Los enfoques bayesianos basados en las probabilidades a priori de los datos.
- Las funciones, entre las cuales se encuentran los métodos de regresión lineal y no lineales tal como las redes neuronales multicapa y las máquinas de soporte vectorial (VSM).
- Los de aprendizaje perezosos entre los que se encuentran vecino más cercano.
- Reglas basadas en predicados lógicos.
- Árboles de decisión.
- Los meta clasificadores o clasificadores por votación de conjuntos.

Los arboles de decisión trabajan con dos opciones de entradas, las cuales pueden ser un objeto o una situación descrita por medio de un conjunto de atributos y a partir de esto devuelve una respuesta que, al final, es una decisión que es tomada a partir de las entradas. Los valores que pueden tomar las entradas y las salidas pueden ser valores discretos o continuos. Se utilizan más los valores discretos por simplicidad, cuando se utilizan valores discretos en las funciones de una aplicación se denomina clasificación y cuando se utilizan los continuos se denomina regresión (White y Sifneos, 2002, pp. 600-614).

Una de las utilidades de los arboles de decisión es que se pueden descubrir patrones en los datos, estos datos se recogen y se organizan en modelos que se utilizaran posteriormente, los modelos pueden ser descritos como gráficos o árboles. Empezamos por explicar la técnica estadística: el análisis de componentes principales, para solucionar este problema de clasificación.

Análisis de componentes principales (ACP)

Estas técnicas fueron inicialmente desarrolladas por Pearson a finales del siglo XIX y posteriormente fueron estudiadas por Hotelling en la década de los treinta del siglo XX. Sin embargo, hasta la aparición de las computadoras se empezaron a popularizar (Renchner,1995).

Los ACP pertenecen a una serie de técnicas estadísticas multivariantes, eminentemente descriptivas, se utiliza en grandes masas de datos, su principal objetivo es reducir la dimensionalidad de los datos, transformando el conjunto de p variables originales en otro conjunto de m variables incorrelacionadas llamadas componentes principales (Renchner,1995).

Este análisis nos permite trabajar con dos opciones: usar la matriz de correlaciones o bien, la matriz de covarianzas. En la primera opción, se le está dando la misma importancia a todas y a cada una de las variables; esto puede ser conveniente cuando se considera que todas las variables son igualmente relevantes.
La segunda opción se puede utilizar cuando todas las variables tengan las mismas unidades de medida y considerando también su grado de variabilidad (Renchner,1995; White y Sifneos, 2003, pp. 600-614).

Una de las formas de utilizar los ACP es realizar combinaciones lineales de las variables originales, de manera que se ordenen en función del porcentaje de varianza. En este sentido, el primer componente será el más importante, porque es el que explica el mayor porcentaje de la varianza de los datos. Además. este estudio se realiza en el espacio de las variables, en forma dual, en el espacio de los individuos. En resumen:

- Las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.
- De modo ideal, se buscan $m < p$ variables que sean combinaciones lineales de las p originales y que estén no correlacionadas, recogiendo la mayor parte de la información o variabilidad de los datos.
- Si las variables originales están no correlacionadas, entonces, no tiene sentido realizar un análisis de componentes principales.
- El análisis de componentes principales es una técnica matemática que no requiere la suposición de normalidad multivariante de los datos, aunque si esto último se cumple se puede dar una interpretación más profunda de dichos componentes.

Árbol de decisión

La minería de datos cuenta con tres componentes: clustering o clasificación, reglas de asociación y análisis secuenciales. En el clustering o clasificación se analizan los datos y se generan conjuntos de reglas que agrupen y clasifiquen los datos futuros; se debe tener en cuenta que en la minería de datos se busca obtener reglas que particionen los datos en clases definidas, esto se torna complicado cuando existe una gran cantidad de atributos y millones de registros, un algoritmo implementado en estadística. Kass (1980, pp. 119-127) introdujo un algoritmo recursivo de clasificación no binario, llamado CHAID (Chi-square automatic interaction detection). Más tarde, Breiman, Friedman, Olshen y Stone (1984) introdujeron un nuevo algoritmo para la construcción de árboles y los aplicaron a problemas de regresión y clasificación. El método es conocido como CART (classification and regression trees) por sus siglas en ingles.

Estos algoritmos utilizados en árboles de decisión es una técnica de predicción que se emplea en el campo de inteligencia artificial, en el cual, a partir de una base de datos, se construyen diagramas de construcción lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren en forma repetitiva para solución de un problema (Hastie, Tibshirani y Friedman, 2001, Medina, s.f.; Vallejos, 2006).

En esta técnica de construcción de arboles de clasificación, se toma la próxima partición de manera optima en el conjunto del árbol, esto evita la confusión combinatoria en cuanto numero de decisiones futuras por considerar, por esto se elige la medida justa por optimizar en cada corte, para facilitar las próximas divisiones. Los pasos por seguir de esta técnica son los siguientes:

1. Aprendizaje: consisten en la construcción del árbol a partir de un espacio muestral , este paso es el más complejo, y de él depende el resultado final.
2. Clasificación: en este paso se realiza el etiquetado de un patrón *W*, independiente del conjunto de aprendizaje, en el que se tratan de responder los cuestionamientos asociados a nodos interiores, utilizando un parámetro de patrón *W*; este proceso se repite desde la raíz , hasta alcanzar una hoja, siendo el camino impuesto por los resultado de cada evaluación.

La técnica de arboles de decisiones permite trabajar con patrones, sin embargo, su eficiencia está estrechamente relacionada con la calidad de los patrones que se utilicen. Las ventajas del árbol de decisión son:

- Puede ser aplicado a cualquier tipo de variables predictoras: continuas y categóricas.
- Los resultados son fáciles de entender e interpretar.
- No tiene problema de trabajar con datos perdidos.
- Hace automáticamente selección de variables.
- Es invariante a transformaciones de las variables predictoras.
- Es robusto a la presencia de *outliers*.
- Es un clasificador no parametrito, es decir, que no requiere suposiciones.
- Toma en cuenta las interacciones que puede existir entre las variables predictoras.
- Es rápido de calcular.

Desventajas del árbol de decisión:

- El proceso de selección de variables es sesgado hacia las variables con más valores diferentes.

- Dificultad para elegir el árbol óptimo
- La superficie de predicción no es muy suave, ya que son conjuntos de planos.
- Requiere un gran número de datos para asegurarse que la cantidad de observaciones en los nodos terminales es significativa.
- Ausencia de una función global de las variables y como consecuencia perdida de la representación geométrica.
-

Además las técnicas estadísticas, son fundamentales a la hora de validar hipótesis y analizar datos, por lo cual la estadística desempeña un papel muy importante, para cuantificar adecuadamente la incertidumbre; la utilización de estas técnicas de clasificación nos permite ratificar e identificar prematuramente los datos más significativos de las muestras tomadas a la población de maguey papalote, lo que admite intervenir prematuramente a disminuir la extinción de la planta (White y Sifneos, 2002, pp. 600-614; Lawler, White, Sifneos y Master, 2003, pp. 875-882; Hastie, Tibshirani y Friedman, 2001; Alonso, Covarrubias y Maradiaga, 2009).

Simulación

Se generó una base de datos de 8000 puntos para extraer los valores de los predictores de la base de datos SIMMP. De acuerdo con Maradiaga (2004), las áreas naturales de maguey papalote ocupan cerca del 78% del territorio nacional, pero al distribuir puntos al azar esta representatividad queda reducida a valores cercanos al 40%. Para mejorar las posibilidades de predicción se consideró un diseño balanceado, localizando un número de muestras al azar (muestreo por conjuntos ordenados de rangos) similar en dos conjuntos, maduración y no madura, mediante un mapa que se obtuvo del Sistema de Información Geográfica (SIG), tomando el criterio de la existencia de poblaciones silvestres de maguey papalote, que fuesen áreas de aprovechamiento para la producción de mezcal.

Evaluación de los resultados de análisis de componentes principales

Se dispone de una muestra de doce localidades de Guerrero, México en las que se midieron diferentes variables relacionadas con la producción del maguey papalote; las variables se muestran en la tabla 2 e interesa investigar la relación entre la etapa del maguey papalote y el resto de las variables; esto para conocer la producción del maguey papalote, y con esto poder orientar la reforestación de dicha población.

Se utilizó el paquete estadístico SPSS versión 17.0 para realizar el análisis de componentes principales, (subsección 2.1). Antes de emplear la técnica multivariada se realizó un contrate de esfericidad de Barlett y de Medida de KMO para determinar si hay correlación entre las variables objeto de estudio y para determinar si la técnica de ACP es aplicable; en este caso tabla 1.

Al observar los resultados en la tabla 1, el estadístico KMO tiene un valor de 0,617 que lo acerca a la unidad, lo que indica que los datos se adecuan para efectuar un ACP y el contraste de Bartlett con p-valor 0,000 indica que se rechaza la hipótesis nula de que las variables iníciales no están correlacionadas, por tanto, se puede efectuar un análisis factorial ACP.

En consecuencia, se continúa con la matriz de correlaciones para observar cómo se comporta cada variable frente a las otras y para observar su determinante, el cual debe ser muy pequeño para poder decir que el grado

de intercorrelación entre las variables es muy alto (tabla 3).

En la tabla 3, se observa que el determinante indica un grado de intercorrelación entre las variables, además, la edad estimada del maguey papalote con el número de líneas están correlacionadas entre sí; asimismo el número de líneas y altura total presentan la mayor relación "0,771", lo que indica un factor común en relación con la producción del maguey papalote.

Al aplicar el método de extracción de componentes principales en el análisis factorial con rotación varimax se obtuvo los siguientes resultados vistos en la tabla 4. Es así como los valores propios, también conocidos como *eigenvalores,* para cada componente se encuentran en la columna "total" y en la siguiente columna se observa el porcentaje de varianza explicada con el método de extracción. Sin embargo, al aplicar la rotación de los ejes se ve cómo el porcentaje de explicación particular varía, pero el acumulado sigue siendo el mismo. Esto se debe a que en el momento de realizar la rotación algunas variables cambian de componente, pero el objetivo sigue siendo el mismo, el cual es minimizar las distancias entre cada grupo perdiendo la mínima información posible, a la vez, que se aumenta la relación de las variables que quedan en cada factor, por lo que se puede concluir que en la técnica de ACP se pasa de siete variables observables a dos "ficticias" con las cuales se explica el 71,158% de la variación total.

La tabla 5 contiene las proyecciones de cada una de las variables sobre cada uno de los factores encontrados mediante el método de componentes principales; estas proyecciones reciben el nombre de saturaciones. Al sumar el cuadrado de cada saturación para cada componente "Factor" se obtiene su eigenvalor, mencionado en la tabla 3, por tanto, para el primer factor será: 0:7922+0:7292+0:7572+0:8562+ 0:1932 0:4552 0:5442 = 4:17. De igual manera, para el factor 2 será: 0:1062 + 0:5312 + 0:5352 0:732 0:6042 + 0:8162 + 0:6012 = 2:183.
Esto significa que las siete variables son explicadas suficientemente con dos factores, en los cuales el primer factor agrupa las variables: edad estimada, altura total, número de líneas y cobertura aérea, mientras que el segundo factor contiene la variable exposición (mN).

Adicionalmente, con la tabla 6, se puede obtener las transformaciones lineales que relacionan los componentes con las variables y, por lo tanto, encontrar el resultado de las dos nuevas variables "ficticias" para cada registro, con lo cual se podrán utilizar estos valores en análisis posteriores (regresión, clúster, etc.), ya que estas variables sustituyen las variables iníciales que las resumen en virtud del ACP que se acaba de realizar.
De esta manera, las fórmulas para las nuevas transformaciones lineales son:

C1 = 0:756 - Edadestimada + 0:890 * Alturatotal + 0:917 – No.delineas + 0:733 * Coberturaárea - 0:098 * mE - 0:042 * mN - 0:218 * Altitud (1)

C2 = -0:256 *Edadestimada + 0:149 * Alturatotal + 0:140 *No.delneas - 0:448 *Coberturaaérea - 0:626 *mE + 0:9:33 *mN + 0:781 *Altitud (2)

Con estas dos ecuaciones, se pueden encontrar las dos nuevas variables sustitutas que se encuentran relacionadas con las variables observables del maguey papalote.

Evaluación de los resultados del árbol de decisión

En el caso de minería de datos (MD), la primera etapa consistió en construir la base de datos (BD) y almacenar la data para obtener su integridad, validez, relevancia y confiabi-

lidad; se realizó la depuración y validación, luego se procedió con la preparación de los datos. Este paso es muy importante, porque en él se debe decidir que hacer con los valores perdidos (*missing values*).

El uso de muestreadores puntuales permitió trabajar con bases geográficas de distinta proyección, modelo de datos o tamaño de celda sin necesidad de reestructuración de estas; al mismo tiempo, facilitó su exploración posterior de manera directa con distintos software de minería de datos

Moreno, Quintales, Penalvo y Martin, finalmente, sustituyeron los datos con valores nulos para alguna de las variables aplicando el vecino más cercano. En otra etapa de MD, para situaciones como nuestro caso de estudio, los árboles de decisión son una alternativa. Se utilizó el software estadístico R, por su amplio contenido de librerías útiles para el proyecto y además es libre. Por último, se procedió al acoplamiento de las diferentes herramientas. El árbol que se desea construir, cuenta con las mismas variables utilizadas en ACP (ver tabla 2).

El árbol de decisión (figura 2) no presenta ninguna característica en especial; de cada nodo sale una rama por valor del atributo que se está siendo probada y así sucesivamente, hasta llegar a las hojas que indican la clase. El árbol de decisión presentado ha sido previamente podado (figura 4) y es exhaustivo, de ramas mutuamente excluyentes.

Adicionalmente, ya conociendo el árbol de decisión, se analizan los errores predichos en cada una de las ramas del árbol generado para analizar si es conveniente simplificarlo. En este caso, el error total predicho para el árbol estará dado en la tabla 8 y, asimismo, el error después de podar (figura 5).

A pesar de esta diferencia entre los modelos obtenidos en uno y otro caso, observamos que la proporción de error en ambos casos es baja. Con lo anterior, aunque la ganancia favorezca a los atributos con mayor cantidad de valores posibles (tabla 7), no podemos afirmar que esto influya en gran medida en el análisis sobre los datos de prueba.

En la figura 2, las variables no son presentadas por un grupo grande de muestras, esto lleva a que los grupos más pequeños sean mal clasificados. Para pasar a la siguiente etapa de experimentos, se podría pensar en elegir la configuración al mínimo tamaño para Split de 2 y mínimo tamaño para hoja de 1 (tabla 8 y figuras 4 y 5), que da el menor error de clasificación. Sin embargo, si observamos detenidamente este árbol (figura 4), podemos notar que algunos registros son clasificados según la edad del maguey papalote, lo que nos da como resultado un árbol unido con el conjunto para el que fue entrenado y que sería incapaz de generalizar nuevos registros, por lo que elegimos la configuración mínimo tamaño para Split de 2 y mínimo tamaño para hoja de 1, la cual, aunque tienen un *performance* similar al de la etapa pasada (figura 2), es más simple, lo que trae múltiples ventajas entre ellas mayor velocidad al momento de clasificar.

Los resultados obtenidos son muy interesantes. Si analizamos los árboles, vemos que, en primer termino, el atributo "Edad" es el que más información brinda tanto utilizando la ganancia como la proporción de ganancia como medidores de información. Pero, una vez en el caso de las variables que tienen edad=6 años, la ganancia considera que el atributo Exposición es el que brinda más información, mientras que la proporción de ganancia considera que el atributo altitud brinda más información que los demás. Si se analiza el archivo de lo generado por el

programa en R, para el caso de la edad estimada del maguey papalote. En la tabla 7, se observa la diferencia en la elección de atributos de división para ambos medidores de información.

Si analizamos las características de los datos, vemos que el atributo "Edad estimada" toma nueve valores distintos, mientras que el atributo "Altura total" toma tres valores distintos. Recordemos que la ganancia favorece a los atributos con más valores y esa es la razón por la que se comenzó a utilizar la proporción de ganancia, que promedia o normaliza, el calculo de la ganancia de información en un conjunto de datos (ver tablas 7, 8 y 9).
El árbol de decisión obtenido utilizando la ganancia como criterio de decisión es de mayor tamaño que el obtenido utilizando la proporción de ganancia. Esta diferencia se origina por la preferencia de la ganancia por atributos con más cantidad de valores. Veamos que en la rama "Altura total", el método que utilizó la ganancia dividió los datos según el atributo de 30 a 167 cm, que toma cuatro valores distintos. En este caso, el hecho de que un árbol sea de mayor tamaño no favorece su resultado; el segundo árbol, más pequeño, tuvo un mejor resultadp, en los casos de prueba, ya que clasificó solo siete de ellos erróneamente, mientras que el árbol generado con la ganancia clasificó veinticinco erróneamente. A pesar de esta diferencia, la estimación del error sobre futuros casos es muy buena para los dos árboles: del 53,9% para el generado utilizando la ganancia y del 53,9% para el generado utilizando la proporción de ganancia (ver tabla 10).

A manera general, el árbol elegido como nuestro modelo de clasificación después de los experimentos anteriores es el de la figura 4, que tiene la configuración mostrada en la tabla 9.

Conclusiones

Se detectaron patrones de respuesta agronómica del maguey papalote, en relación con las variedades estudiadas, lo cual mostró que las técnicas de minería de datos son aplicables en este tipo de estudios, siempre que se cuente con los datos necesarios. Cuando el analista se enfrenta a una cantidad de variables, estas no presentan una relación aparente; por otra parte, es importante complementar la estadística multivariante con *dataming*, ya que esta le proporciona herramientas de contraste claras para observar una realidad que no se nota a simple vista.

En muchas ocasiones, el análisis multivariante permitirá reducir considerablemente el tiempo de gestión o desarrollo del estudio final, gracias a las alternativas de reducción, con respecto a las variables o a los registros.
El método de análisis de componentes principales es una técnica multivariante muy fuerte para aplicarla en la reducción de la dimensión del estudio dado; sin embargo, se debe tener cuidado al tratar con *dataming* y componentes principales, ya que se pueden aplicar como herramientas estadísticas por separado, lo cual traerá resultados diferentes o vincularlas utilizando el análisis con método de extracción de los factores; por eso, se le recomienda al lector profundizar en las diferencias y similitudes de estas dos técnicas.
El análisis de componentes principales efectuado se convierte en un procedimiento útil antes de interpretar la producción obtenida del maguey papalote, cuando se realiza una composición de la edad estimada. Junto a la altura del maguey papalote, en el resultado de los componentes principales se pueden diferenciar más elementos que cuando se utilizan todas las variables independientes; además, se puede concluir que los dos variables están altamente correlacionadas, por tanto, con los componentes seleccionados se conserva la variabilidad.

Con la técnica de *dataming*, se pudo reducir la dimensión de veinte variables originales a una variable sustitutas con la que se podrán realizar estudios en dos dimensiones siendo más práctico para el analista.

Por último, cabe mencionar que, aunque existen diversas herramientas de minería de datos que nos permiten realizar las diferentes etapas de la extracción de conocimiento de manera automática, es necesaria la supervisión de un usuario que guíe este proceso y verifique la congruencia de los resultados. La técnica de arboles de decisión es un buen clasificador y predictor que se puede aplicar cuando se conocen de antemano las clases de un conjunto de entrenamiento (aprendizaje supervisado).

Autores adicionales: C. N. Bouza (Universidad de la Habana, email: bouza@matcom.uh.cu) y F. Madariaga (Universidad Autónoma de Guerrero, email: fmaradiaga@hotmail.com).

Referencias

Alonso, L., Covarrubias, D. y Maradiaga, F. (2009). *Estrategias de muestreo para el inventario de maguey papalote (Agave Cupreatra Trel & Berger)*. Tesis para obtener el grado de Maestría en Ciencias Área Estadística aplicada, No publicada, Guerrero, México: Unidad Académica de Matemáticas, UAGro.

Breiman, L., Friedman, J.H., Olshen, R.A. y Stone, C. J. (1984). *Classication and Regression Trees*. Belmont, Ca.: Wadsworth International.

Hastie, T., Tibshirani, R. y Friedman, J. (2001). *The Elements of Statistical Learning: Datamining, Inference and Prediction*. New York: Springer Verlag.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29* (2),119-U127.

Lawler, J.J., White, D., Sifneos, J. y Master, L. (2003). Rare species and the use of indicator groups for conservation planning. *Conservation Biology, 17* (5),875-882.

Maradiaga, F. (2004, nov.). *Desarrollo de un sistema de inventario y monitoreo de maguey papalote (agave cupreata trel. & berger) en el estado de guerrero*. Fundación PRODUCE Guerrero A.C., Programa de Recursos Biológicos Colectivos (CONABIO) e Instituto de Investigación Científica Área Ciencias Naturales de la UAGro.

Medina, J.C. (s.f.). *Análisis comparativo de técnicas, metodologías y herramientas de ingeniería de requerimientos*. Tesis para obtener el grado de Maestría en Ciencias en la especialidad de Ingeniería Eléctrica Opción Computación, México, D.F.: Centro de Investigación y de Estudios Avanzados del IPN.

Moreno, M.N., Quintales, L.A.M., Penalvo, F.J.G., y Martin, M.J.P. (2006). Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software. España: Universidad de Salamanca. Departamento de Informática y Automática, Salamanca.

Renchner, A.C. (1995). *Methods of multivariate Analysis*. New York: Wiley.

Stockwell, D. y Peters, D. (1999). The garp modeling system: Problems and solutions to automated spatial prediction. *Geographical Information Science,13* (2),143-158.

Vallejos, S.J. (2006). Minería de Datos. Argentina: Universidad Nacional del Nordeste Facultad de Ciencias Exactas, Naturales y Agrimensura.

White, D. y Sifneos, J. (2002). Regression tree cartography. *Journal of Computational and Graphical Statistics, 11* (2), 600-614.

Anexo. Figuras y tablas

Figura 1. Árbol de decisión

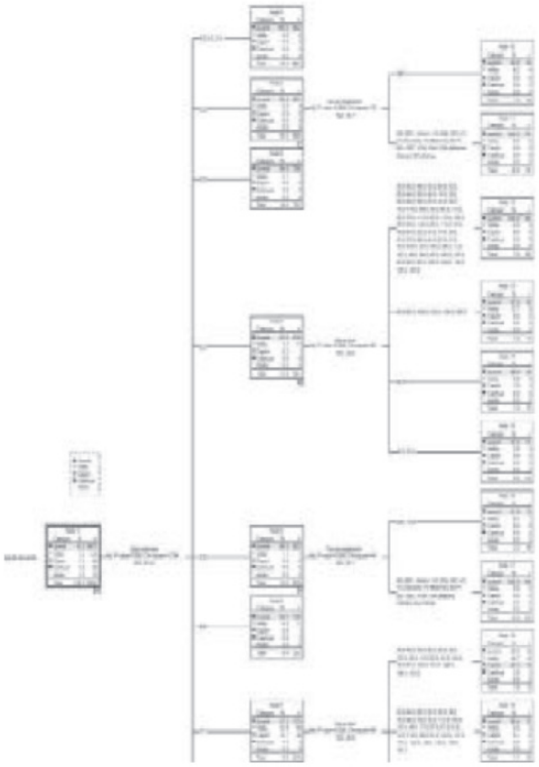


Figure 2: Árbol de decisión

Figura 1. Error al clasificar

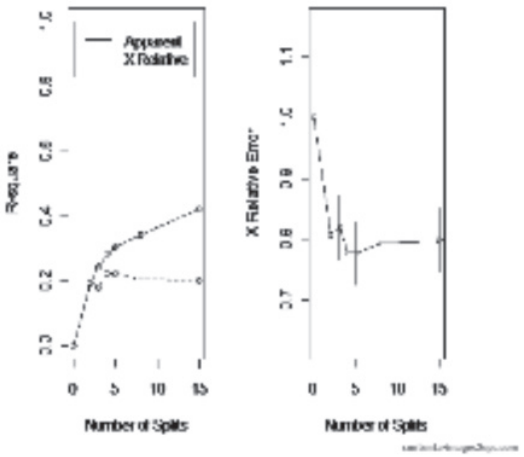


Figura 4. Poda del árbol de decisión

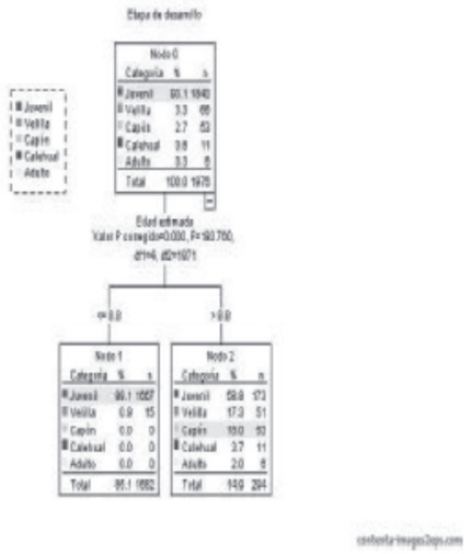


Figura 5. Poda del árbol de decisión

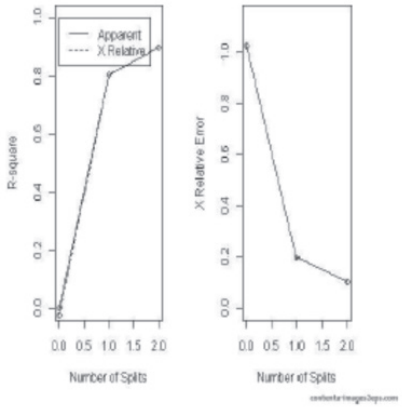


Figure 5: Error al clasificar

Table 1: KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin		.617
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	79.822
	gl	21
	Sig.	.000

Table 2: Variables predictoras evaluadas

Variables	Indicadores
Elaboró.	Nombre completo del técnico responsable de levantar la información
Fecha	Día en el cual se levantó la información. Día / Mes / Año.
Municipio	Nombre del Municipio en el cual se ubican los sitios de muestreo
Localidad	Nombre de la comunidad, ejido, población, o cualquier otro núcleo de población.
Propietario	El nombre completo del dueño, empezando por los apellidos
Nombre del predio	Cuando lo hubiese.
Tipo de tenencia	Si la propiedad es Ejidal, Pequeña propiedad o Comunal.
Tamaño del predio.	La superficie en hectáreas.
Forma del terreno	La forma del terreno se describe en base a la pendiente general.
Hidrología	Presencia de corrientes y cuerpos de agua superficiales.
Suelo	características más visibles de los suelos.
Manejo del predio	Información posible sobre la historia de manejo.
Precisión del GPS	Posicionamiento global.
GPS y Nombre del Punto	Cada organización dispone de dos equipos de georreferenciación.
Cuadrante.	Este puede ser A, B, C o, D.
Número de registro	Número de control por unidad de muestreo.
Altura total	Número de control por unidad desde la base de la planta hasta la punta del cogollo
Observaciones	Cualquier alteración que se observe en el espécimen o que afecte su desarrollo
ALTITUD(msnm)	Distancia vertical sobre el nivel del mar.
Núm. de individuos	Número de maguey papalote en el área.
COBERTURA aérea(m2)	Cantidad total del área muestreada.
Tipo de vegetación	Tipo de vegetación en el área muestrada.
Cobertura del dosel	Extención de las hojas de los agaves.
Etapa de desarrollo	Fase en la cual se encontro al maguey papalote.
Edad estimada	Edad en años de acuedo al número de línas de hojas formadas.
Exposición	Ubicación cartesiana.
Pendiente	Declive del terreno e inclinación, respecto a la horizontal del transecto.
Origen	Es la condición del maguey papalote, silvestre o plantación.
Paraje	Lugar de ubicación del área muestreada.

Table 3: Matriz de Correlaciones.

		Edad estimada	Altura total	No. de líneas
Correlación	Edad estimada	1.000	.497	.642
	Altura total	.497	1.000	.771
	No. de líneas	.642	.771	1.000
	Cobertura Aérea (m2)	.556	.574	.518
	Exposición (mE)	.306	-.172	-.175
	Exposición (mN)	-.169	.063	.062
	ALTITUD(msnm)	-.234	-.100	-.123
Sig. (Unilateral)	Edad estimada		.006	.000
	Altura total	.006		.000
	No. de líneas	.000	.000	
	Cobertura Aérea (m2)	.002	.001	.005
Determinante				0.18

Table 4: Varianza total explicada.

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción	
	Total	% de la varianza	% acumulado	Total	% de la varianza
1	3.004	42.919	42.919	3.004	42.919
2	1.977	28.239	71.158	1.977	28.239
3	.987	14.095	85.253		
4	.429	6.124	91.377		
5	.283	4.041	95.418		
6	.184	2.632	98.049		
7	.137	1.951	100.000		

Table 5: Matriz de componentes.

	Componente	
	1	2
Edad estimada	.792	.106
Altura total	.729	.531
No. de líneas	.757	.535
Cobertura Aérea (m2)	.856	-.073
mE	.193	-.604
mN	-.455	.816
ALTITUD(msnm)	-.544	.601

Table 6: puntuaciones en las componentes.

	Componente	
	1	2
Edad estimada	0.260	-0.070
Altura total	0.337	0.132
No. de líneas	0.347	0.129
Cobertura Aérea (m2)	0.238	-0.160
mE	-0.079	-0.302
mN	-0.49	0.353
ALTITUD(msnm)	-0.26	0.353

Table 7: Ganacia en los nodos.

Nodo	Ganacia	Proporción de ganacia
Edad estimada	64.0%	0.788
Altura total	36.0	0.376

Table 8: Error al clasificar, antes de podar.

split	error	Estimación
	1	2
406	1.0	10.2%
3769	0.194	19.5%
		53.9%

Table 9: Clasificación general.

Tipo	Clasificado como maduro	Clasificado como no maduro
	1	2
Maduro	1255	0
No maduro	2695	629

Table 10: Error al clasificar, despues de podar.

split	error	Estimación
	1	2
1	1.0	1.02%
2	0.194	0.636%
		53.9%