Aplicación de la minería de datos en la evaluación de la aptitud física de las tierras para el cultivo de la caña de azúcar

Saddys Segrera
Dept. de Informática
Instituto Nacional de Investigaciones de la Caña de Azúcar
19390 Ciudad de la Habana
saddys@inica.edu.cu

María N. Moreno
Dept. de Informática y Automática
Facultad de Ciencias
Univ. de Salamanca
37008 Salamanca
mmg@usal.es

Luis A. Miguel
Dept. de Informática y Automática
Facultad de Ciencias
Univ. de Salamanca
37008 Salamanca
lamq@usal.es

Resumen

La evaluación de tierras es una herramienta para la planificación estratégica que ayuda a determinar el comportamiento de la tierra bajo usos determinados, en términos de beneficios, costos, y efectos ambientales. De ahí que sin clientes que utilicen los resultados de estos estudios, no tenga sentido hacerlos. El objetivo de este trabajo es tomar como caso de estudio los datos procedentes de las variables empleadas en la emisión de las categorías de aptitud física de las tierras del cultivo de la caña de azúcar en Cuba para aplicar técnicas de minería de datos que, a través de clasificadores, permitan predecir qué factores del suelo, del clima y agrícolas, presentes en las parcelas estudiadas, podrían ser más determinantes en este proceso. En la realización de este estudio se ha empleado la herramienta *WEKA* que se encuentra disponible de forma gratuita en Internet, creada por especialistas de la Universidad de Waikato y también *Mineset* de *Silicon Graphics Inc*.; el primer sistema fue utilizado para probar varios clasificadores con distintos inductores y con el segundo, se evaluó un clasificador de árbol de decisión y se realizó la evaluación de dicho clasificador con el apoyo de su potencial visual.

1. Introducción

El proceso de evaluación de tierras permite realizar estudios con el fin de valorar si el uso dado a una unidad agrícola es el más adecuado, apoyándose para tales decisiones en factores agroclimáticos que inciden en su comportamiento. [2] presentó los requisitos de datos de suelos para lograr evaluaciones cuantificadas de tierras que contribuyesen a un acercamiento más preciso de la aptitud de la tierra para determinado uso.

En Cuba el Instituto Nacional de Investigaciones de la Caña de Azúcar (INICA), por primera vez realizó un trabajo de evaluación de tierras que abarcó todas las áreas cañeras del país, que se elaboró de manera participativa en el que intervinieron productores e investigadores, como premisa para establecer estrategias para la planificación del desarrollo del área agrícola del Ministerio del Azúcar [15].

Por su parte, la aplicación de técnicas de minería de datos podría contribuir a mejorar los resultados del proceso de evaluación de tierras.

Aunque hasta hace unos años las técnicas utilizadas por la minería de datos no se habían convertido en un tema de primer orden, estas han sido utilizadas desde los años 80. Esta tecnología emergente combina los análisis estadísticos, la máquina de aprendizaje y la gestión de las bases de datos para extraer información de voluminosas tablas de datos [14].

Según [8] haciendo referencia al uso de esta tecnología en el área de los negocios, la minería de datos es un nuevo proceso de análisis de apoyo a las decisiones para encontrar conocimiento oculto en datos corporativos y entrega la comprensión del análisis a profesionales de los negocios.

Por su parte, [5] la describen como un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos.

La agricultura es una fuente de información constante ya que a través de distintos factores: suelo, fertilizantes, temperatura, precipitaciones, entre otros, vinculados todos a las cosechas de cultivos.

Estas variables a su vez, son mediciones imprescindibles para la realización de investigaciones relacionadas directamente con la

toma de decisiones para la mejora de tecnologías agrícolas. Por este motivo, muchos han sido los esfuerzos para la aplicación de técnicas de minería de datos en esta actividad con el fin de descubrir relaciones ocultas entre las distintas variables y características relacionadas con la agricultura, que al ser identificadas y estudiadas podrían aportar iniciativas para el beneficio de esta rama de la economía [1], [4], [10].

La clasificación de las técnicas de la minería de datos se divide en 2 categorías: supervisadas y no supervisadas [7], [16]. Las primeras predicen los valores de un atributo etiqueta u objetivo con la ayuda de los valores de otros atributos, por lo que van a estar dirigidas a la clasificación y a los sistemas de predicción. En el caso de las técnicas no supervisadas a partir de un conjunto de datos disponible se persigue encontrar relaciones entre los atributos, patrones habituales de comportamiento, desconocidos antes del análisis, de ahí a que este tipo de técnicas también se les llame de descubrimiento del conocimiento.

En este estudio, el trabajo está encaminado a comparar algoritmos supervisados a través de clasificadores. En este caso, se probaron distintos inductores (árboles, reglas y funciones) y en cada caso se analizaron los clasificadores que se obtenían como resultado de la aplicación de distintos algoritmos.

El objetivo de este trabajo es presentar un estudio a través de técnicas de minería de datos que permitan determinar, a través de un clasificador, el grado de aptitud de parcelas de tierras para el cultivo de caña de azúcar. La importancia que tiene conocer la aptitud física de las tierras es que posibilita realizar un uso correcto de las mismas. De manera que si no es recomendable su uso para este cultivo pudiera emplearse para otras actividades, evitándose así la degradación del suelo en las áreas no aptas y a la vez concentrar los recursos en aquellas parcelas con las condiciones más adecuadas para alcanzar los mayores rendimientos en caña de azúcar.

2. Materiales y métodos

En este trabajo se utilizó la herramienta *WEKA* (*Waikato Environment for Knowledge Analysis*) 3.4 de la Universidad de Waikato, software que se encuentra de manera gratuita en el sitio oficial de esta institución en Internet y contiene múltiples algoritmos para la aplicación de técnicas supervisadas y no supervisadas. También se empleó *Mineset* de *Silicon Graphics, Inc*., debido a la capacidad de visualización más detallada e ilustrativa de este sistema en comparación a *WEKA*, aunque el número de inductores para obtener los clasificadores es más reducido.

Durante el desarrollo de este estudio se siguieron los pasos del proceso de minería de datos definidos por [3].

La base de datos utilizada posee 3879 registros que corresponden a parcelas cultivadas con caña de azúcar. Las variables que los componen son 12 factores asociados al suelo, al clima y características agrícolas (pendiente del terreno, pedregosidad, rocosidad, salinidad, acidez del suelo, capacidad de intercambio catiónico, drenaje, compactación, precipitaciones, profundidad efectiva, agrupamiento agroproductivo del suelo y categoría de aptitud de la tierra).

Inicialmente, la base de datos contenía el atributo rendimiento agrícola (`rendim`) pero no fue considerado en el modelo con vista a conocer la predicción a partir de otros atributos de la tierra, ya que el conocer esta variable implica saber si es adecuada la tierra para este cultivo. Esto conllevaría a obtener un modelo casi perfecto, por lo cual fue eliminado este atributo..

El valor del atributo conocido a predecir en este trabajo está representado en la etiqueta `eval`, perteneciente a la aptitud física de las tierras para el cultivo de la caña de azúcar.

En la figura 1 se muestra a través del *Explorer* de *WEKA* la composición de la base de datos y el número de registros por categoría de aptitud de la tierra. Esta variable posee los valores A1 para las áreas sumamente aptas, A2 para las moderadamente aptas, A3 para las marginalmente aptas y N como áreas no aptas para el cultivo de la caña de azúcar. Por su parte, en la figura 2 se ilustra el número de registros para las distintas variables en función de las categorías de la aptitud de las tierras.

Los datos del estudio proceden del Instituto Nacional de Investigaciones de la Caña de Azúcar (INICA), institución adjunta al Ministerio del Azúcar de Cuba, responsable de las investigaciones que desde 1999 se desarrollan para la diversificación de la agricultura cañera, tomando como base la evaluación de la aptitud física de las tierras cultivadas con este fin.

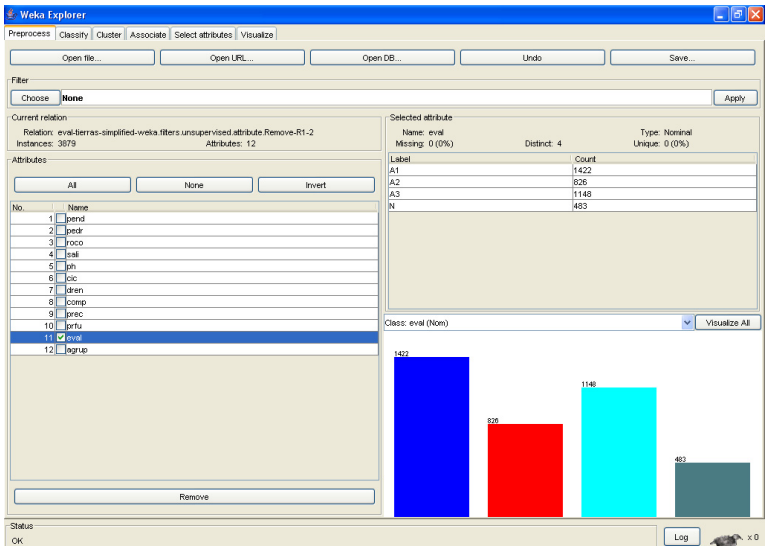


Figura 1. Composición de la base de datos de estudio a través de WEKA y visualización del número de registros en función de la categoría de aptitud de las tierras.

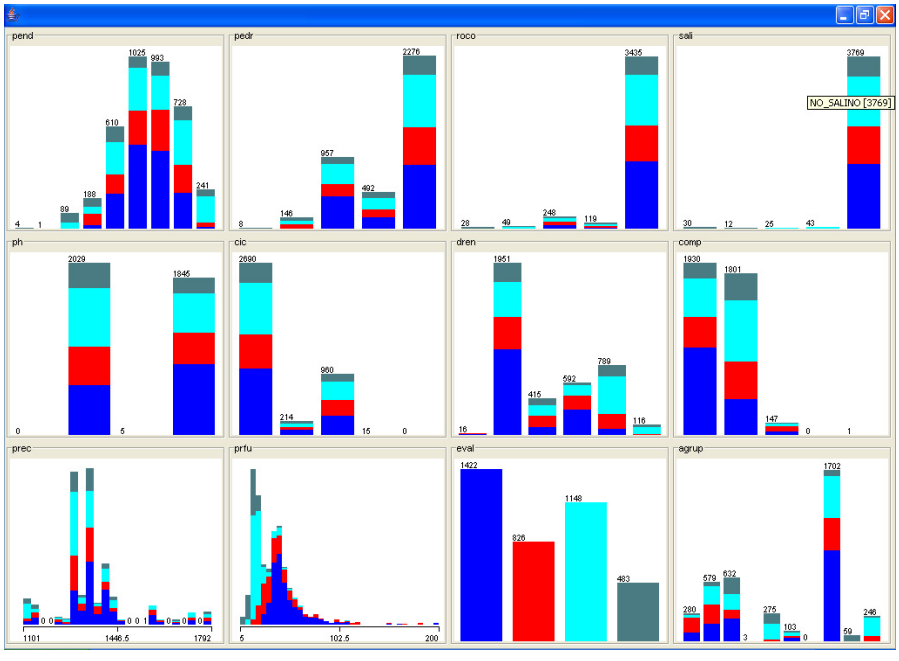


Figura 2. Visualización del número de registros de cada variable en función de la caracterización de la aptitud de las tierras

En el caso de *Mineset* fue necesario construir el archivo con extensión `.schema`, según las indicaciones que ofrece la Guía del usuario de este software [13]. De modo que se construyeron 2 archivos: `eval_ms.schema` donde se declaró la naturaleza de las variables, el separador empleado entre los datos de las distintas columnas y el nombre del archivo con los datos (Figura 3), el segundo archivo `eval_ms.data` contiene cada uno de los artículos ordenados con el separador señalado en el archivo `.schema` y según el orden de aparición expuesto de las columnas, lo cual no fue difícil de elaborar ya que se tomó como base el archivo `.arff` de *WEKA*. De modo que al ser introducidos los datos en *Mineset*, se obtuvo la vista que aparece en la figura 4.

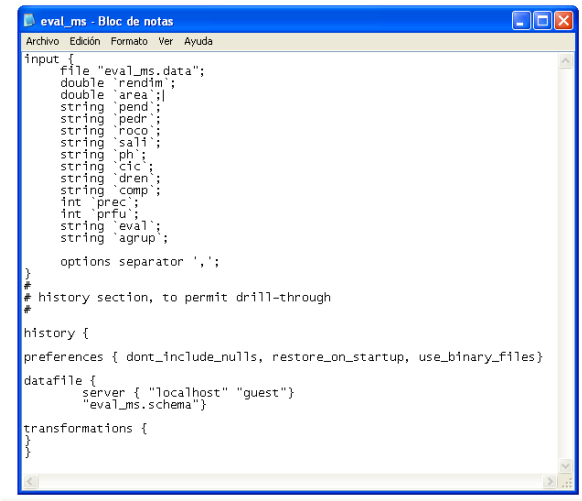
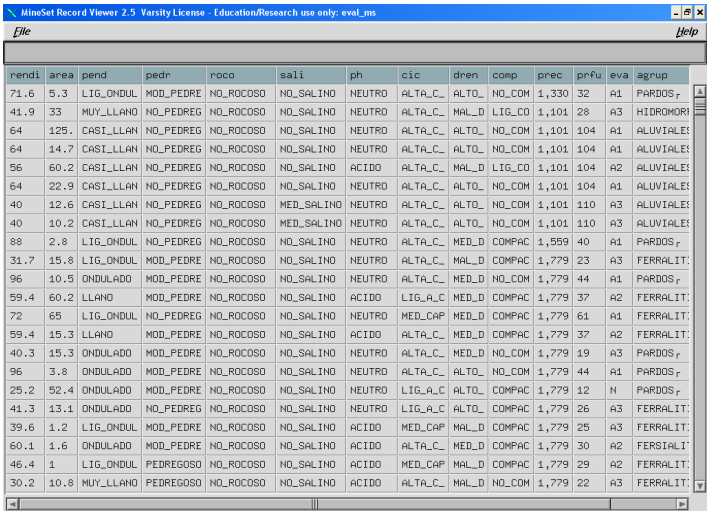


Figura 3. Contenido del archivo eval_ms.schema para procesar datos en Mineset



rendi	area	pend	pedr	roco	sali	ph	cic	dren	comp	prec	prfu	eva	agrup
71.6	5.3	LIG_ONDUL	MOD_PEDRE	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	ALTO_	NO_COM	1,330	32	A1	PARDOS_r
41.9	33	MUY_LLANO	NO_PEDREG	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	MAL_D	LIG_CO	1,101	28	A3	HIDROMORF
64	125.	CASI_LLAN	NO_PEDREG	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	ALTO_	NO_COM	1,101	104	A1	ALUVIALES
64	14.7	CASI_LLAN	NO_PEDREG	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	ALTO_	NO_COM	1,101	104	A1	ALUVIALES
56	60.2	CASI_LLAN	NO_PEDREG	NO_ROCOSO	NO_SALINO	ACIDO	ALTA_C_	MAL_D	LIG_CO	1,101	104	A2	ALUVIALES
64	22.9	CASI_LLAN	NO_PEDREG	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	ALTO_	NO_COM	1,101	104	A1	ALUVIALES
40	12.6	CASI_LLAN	NO_PEDREG	NO_ROCOSO	MED_SALINO	NEUTRO	ALTA_C_	ALTO_	NO_COM	1,101	110	A3	ALUVIALES
40	10.2	CASI_LLAN	NO_PEDREG	NO_ROCOSO	MED_SALINO	NEUTRO	ALTA_C_	ALTO_	NO_COM	1,101	110	A3	ALUVIALES
88	2.8	LIG_ONDUL	NO_PEDREG	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	MED_D	COMPAC	1,559	40	A1	PARDOS_r
31.7	15.8	LIG_ONDUL	MOD_PEDRE	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	MAL_D	COMPAC	1,779	23	A3	FERRALIT
96	10.5	ONDULADO	MOD_PEDRE	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	MED_D	NO_COM	1,779	44	A1	PARDOS_r
59.4	60.2	LLANO	MOD_PEDRE	NO_ROCOSO	NO_SALINO	ACIDO	LIG_A_C	MED_D	COMPAC	1,779	37	A2	FERRALIT
72	65	LIG_ONDUL	NO_PEDREG	NO_ROCOSO	NO_SALINO	NEUTRO	MED_CAP	MED_D	COMPAC	1,779	61	A1	FERRALIT
59.4	15.3	LLANO	MOD_PEDRE	NO_ROCOSO	NO_SALINO	ACIDO	ALTA_C_	MED_D	COMPAC	1,779	37	A2	FERRALIT
40.3	15.3	ONDULADO	MOD_PEDRE	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	MED_D	NO_COM	1,779	19	A3	PARDOS_r
96	3.8	ONDULADO	MOD_PEDRE	NO_ROCOSO	NO_SALINO	NEUTRO	ALTA_C_	ALTO_	NO_COM	1,779	44	A1	PARDOS_r
25.2	52.4	ONDULADO	MOD_PEDRE	NO_ROCOSO	NO_SALINO	NEUTRO	LIG_A_C	ALTO_	COMPAC	1,779	12	N	PARDOS_r
41.3	13.1	ONDULADO	NO_PEDREG	NO_ROCOSO	NO_SALINO	NEUTRO	LIG_A_C	ALTO_	COMPAC	1,779	26	A3	FERRALIT
39.6	1.2	LIG_ONDUL	MOD_PEDRE	NO_ROCOSO	NO_SALINO	ACIDO	MED_CAP	MAL_D	COMPAC	1,779	25	A3	FERRALIT
60.1	1.6	ONDULADO	MOD_PEDRE	NO_ROCOSO	NO_SALINO	ACIDO	ALTA_C_	MED_D	COMPAC	1,779	30	A2	FERSIALI
46.4	1	LIG_ONDUL	PEDREGOSO	NO_ROCOSO	NO_SALINO	ACIDO	MED_CAP	MAL_D	COMPAC	1,779	29	A2	FERRALIT
30.2	10.8	MUY_LLANO	PEDREGOSO	NO_ROCOSO	NO_SALINO	ACIDO	ALTA_C_	MAL_D	NO_COM	1,779	22	A3	FERRALIT

Figura 4. Visualización de los datos en Mineset.

Las fases posteriores del proceso de minería de datos forman parte del próximo epígrafe.

3. Resultados y discusión

A través de los algoritmos de clasificación de *WEKA* se probaron varios algoritmos de distintos inductores para seleccionar aquel que con un menor error construyese un clasificador para la predicción de la categoría de la aptitud física de las tierras y determinar qué parcelas son aptas (A1, A2 y A3) y cuáles no (N).

En este estudio el trabajo estuvo encaminado a comparar algoritmos supervisados a través de clasificadores, de los que existen diferentes métodos, en este caso, se comenzó el análisis con el uso del inductor de árboles de decisión que ofrece *Mineset* (Figura 5). Los árboles de decisión son una forma de representación sencilla, muy usada entre los sistemas de aprendizaje supervisado, para clasificar ejemplos en un número finito de clases. Se basan en la partición del conjunto de ejemplos según ciertas condiciones que se aplican a los valores de los atributos. Su potencia descriptiva viene limitada por las condiciones o reglas con las que se divide el conjunto de entrenamiento; por ejemplo, las reglas pueden ser simplemente relaciones de igualdad entre un atributo y un valor, o relaciones de comparación, etc. [6]. Sin embargo, en el estudio de estos datos no se obtuvieron resultados precisos, ya que el error del clasificador fue de 12.84% y tampoco mejoró este valor los inductores de evidencia y de tablas de decisión, que completan los inductores que ofrece *Mineset*. Por lo que fue necesario realizar el análisis con otros inductores. *WEKA* ofrece más de 20 algoritmos que podrían mejorar estos resultados y hallar un clasificador que indujera mejor a la etiqueta de la aptitud física de las tierras.

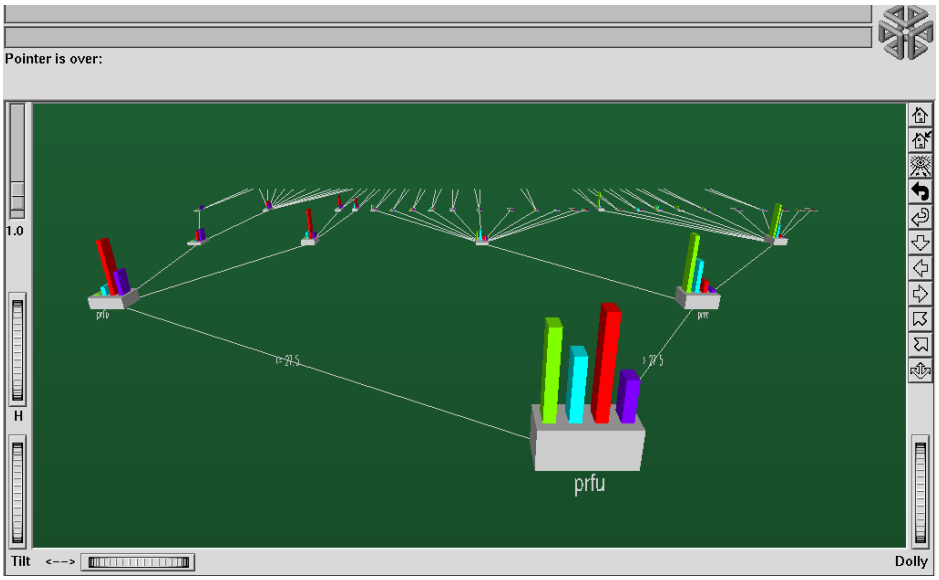


Figura 5. Visualización en Mineset de un modelo de inducción de árboles de decisión

En *WEKA*, primeramente, y siguiendo la idea de que los árboles de decisión constituyen un modelo simple, se realizaron análisis con distintos algoritmos que allí existen para árboles de decisión, que al igual que en *Mineset*, no obtuvieron valores de tasa de errores inferiores al 10%. Posteriormente, se probaron otros inductores, siendo los de mejores resultados los inductores de reglas. El algoritmo *Nnge* (Figura 6), del mismo inductor mejoró los resultados anteriores al obtener un clasificador con un error de 5.76%, donde el 88.4761% de las instancias eran clasificadas de forma correcta. El algoritmo *Nnge* es conocido como el algoritmo del vecino más cercano a través del uso de ejemplares generalizados no anidados (que son hiperrectángulos que pueden ser vistos como reglas del tipo *if then*) [9], [12]. Se utilizó, para todos los casos, el 66% de los datos como conjunto de entrenamiento o *training set* y el resto como conjunto de prueba.

Se generaron 584 reglas en las que intervinieron todos los atributos y fueron reducidas a 183.

La matriz de confusión ha sido utilizada para la evaluación del clasificador mostrando el tipo de predicciones correctas e incorrectas, en la diagonal aparecen los aciertos y fuera de ella las desacertadas. En la figura 7 se refleja el comportamiento del clasificador obtenido, donde se observan, simbolizadas con rectángulos, aquellas parcelas no clasificadas de manera correcta con respecto a la variable profundidad efectiva.

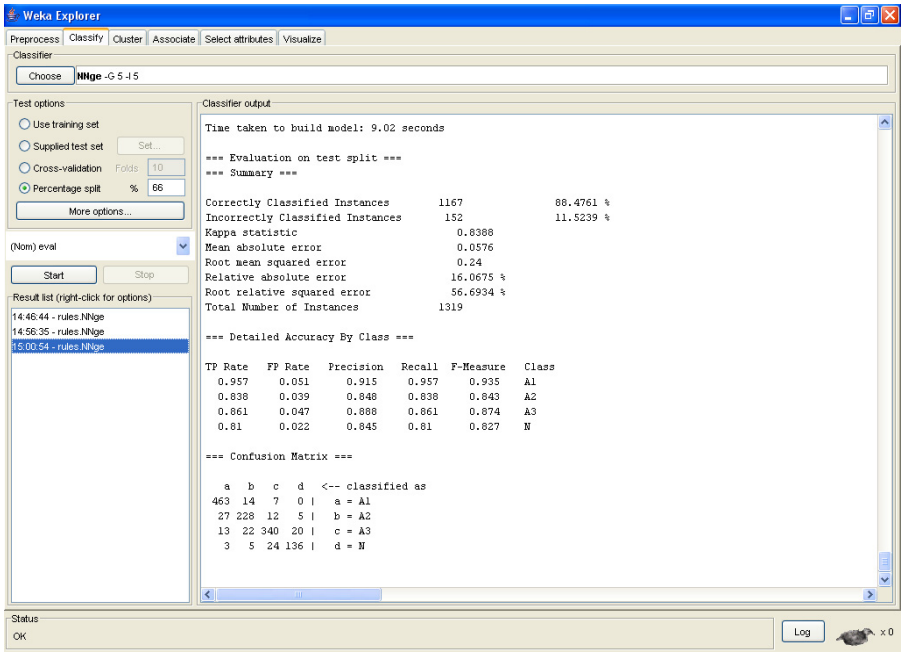


Figura 6. Clasificador obtenido por el algoritmo Nnge.

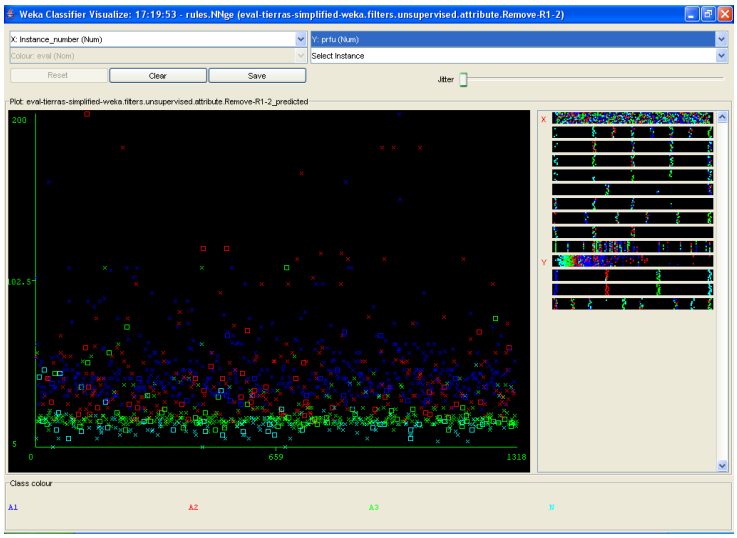


Figura 7. Visualización de los errores del clasificador por el algoritmo Nnge.

3.1. Evaluación del clasificador por árbol de decisión en Mineset

El gráfico de la curva de aprendizaje describe la relación de la precisión *versus* el tamaño del conjunto de entrenamiento, frecuentemente se usa para determinar el más pequeño conjunto de entrenamiento para adicionar instancias extras, sin que aumente el error y mejorar la precisión [11], [17].

Aprovechando las facilidades que brinda *Mineset* se realiza la evaluación del clasificador de inducción del árbol de decisión. En la figura 8 se muestra la curva de aprendizaje obtenida con los datos del caso de estudio en *Mineset*, donde se observa que se utiliza casi la totalidad de los registros para obtener el menor error de aprendizaje posible, y que en este caso, no coincide con un valor de error bajo, sino que representa aproximadamente un 11%. Antes de construir las curvas de eficacia y las curvas de retorno de la inversión en *Mineset* se creó una nueva columna nombrada `cultivar`. Se agruparon en una clase todas las parcelas con categorías A1, A2 y A3, es decir, las parcelas aptas para el cultivo de la caña de azúcar, en mayor o en menor grado, el análisis se realizaría sólo con 2 categorías, parcelas aptas (A1+A2+A3) y las parcelas no aptas (N). Las parcelas aptas se corresponden con categoría "Sí" de la nueva columna, es decir, sí están aptas para cultivar y para el caso de las no aptas se ha transformado su categoría a "No" (no están aptas para cultivar). Se obtuvieron posteriormente los gráficos que aparecen en las figuras 9 y 10 para la clase "No" correspondientes a la curva de eficacia y la de retorno de la inversión para esa clase, respectivamente.

El análisis de la curva de eficacia obtenida nos indica que cuando se ordenan los registros en función del clasificador, interviniendo aproximadamente sobre el 14% del total de las parcelas (registros), se actúa sobre el 86% de las parcelas que con más probabilidad no tengan las condiciones necesarias para cultivar caña de azúcar en ellas. Esto permite que los estudios de diversificación posteriores dirigidos al uso de estas parcelas para otros fines para los que sí poseen aptitud suficiente se desarrollen para aquellas menos idóneas para este cultivo.

La curva representada en la figura 10 muestra que realizando una inversión sobre el 14% aproximadamente de las parcelas para la diversificación de las áreas donde actualmente se cultiva la caña de azúcar (entiéndase por diversificación a darle usos diferentes, que pueden ser para el cultivo de otras plantas, para la ganadería u otro uso de distinta índole), se obtiene el máximo retorno sobre la inversión.

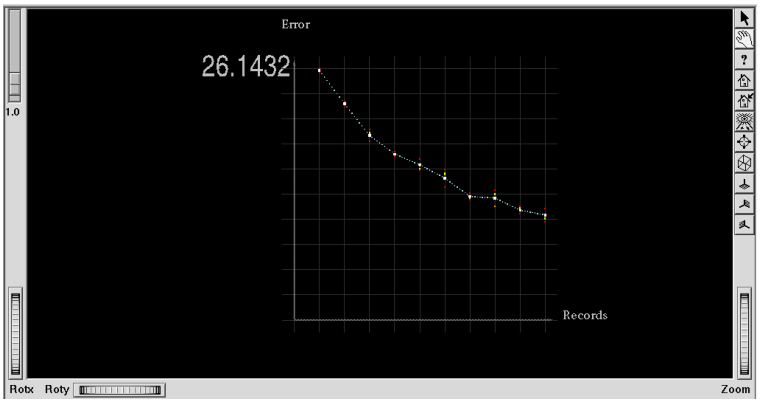


Figura 8. Curva de aprendizaje del clasificador en Mineset.

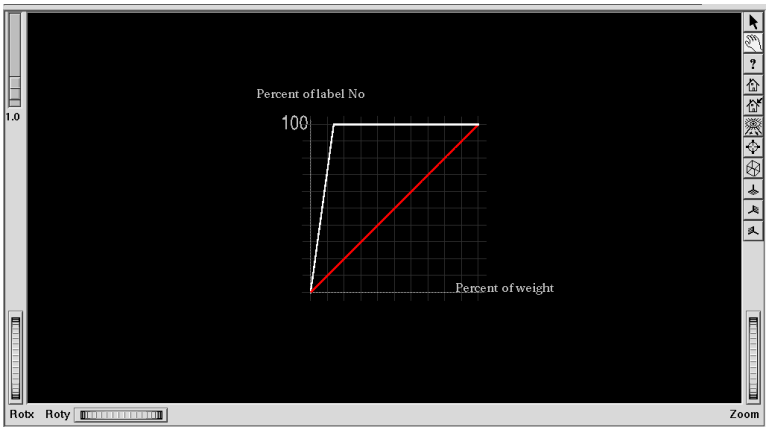


Figura 9. Curva de eficacia para la clase "No" (no son aptas para el cultivo de la caña de azúcar).

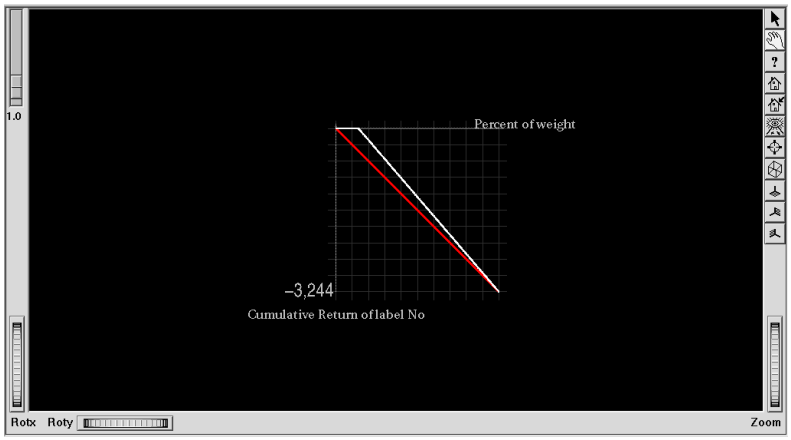


Figura 10. Curva de retorno para la clase "No" (no son aptas para el cultivo de la caña de azúcar).

Durante el estudio quedó demostrado que la preparación y transformación de los datos ocupa el 60% del tiempo en que se realiza una aplicación de minería de datos y que la naturaleza de los datos influye en gran medida en la precisión de un clasificador en dependencia del algoritmo utilizado.

4. Conclusión

A través del uso de la minería de datos se encontró un clasificador de inducción de reglas que predijera los valores de la aptitud de las tierras para parcelas cultivadas con caña de azúcar. De esta forma se contribuye al proceso de toma de decisiones en la agricultura cañera.

El uso de las herramientas informáticas *WEKA* y *Mineset* permitieron realizar un análisis amplio de los datos y de los resultados de su procesamiento. A pesar de que *WEKA* ofrece muchos algoritmos para la construcción de clasificadores carece de medios para su evaluación.

Referencias

[1] Abdullah, S. Brobst, I. Pervaiz and M.U. Azhar. Learning Dynamics of Pesticide Abuse through Data Mining. *Proc. of the Australasian Workshop on Data Mining and Web Intelligence (AMSM&WI2004)*, Dunedin, New Zealand, 2003.

[2] Bouma, J. Using soil survey data for quantitative land evaluation. In: *Advances in Soil Science* (ed. B.A. Stewart). Springer, New York, pp. 177-213, 1989.

[3] Cabena, P., P. Hadjinian, R. Stadler, J. Verhees and A. Zanasi, *Discovering Data Mining. From Concept to Implementation*, Prentice Hall, 1998.

[4] Christensen, W.F. and D. Cook. Data Mining Soil Characteristic Affecting Corn Yield, 1998. http://citeseer.ist.psu.edu/341770.html.

[5] Fayyad, U.M., G. Piatetsky-Shapiro and P. Smyth. From Data Mining to Knowledge Discovery. In Advances in Knowledge Discovery and Data Mining, Fayyad, Piatetsy-Shapiro, Smyth y Uthurusamy Eds., AAAI Press, Menlo Park, California, pp. 1-34, 1996. http://www.kdnuggets.com/gpspubs/ aimag-kdd-overview-1996-Fayyad.pdf

[6] Gómez, A.J. Inducción de conocimiento con incertidumbre en bases de datos relacionales borrosas. Capítulo 2: "Estado del arte", Tesis Doctoral, 1998. http://www.gsi.dit.upm.es/~anto/tesis/html/stateart.html

[7] Herschkowitz, D. and J. P. Nadal. Unsupervised and supervised learning: Mutual information between parameters and observations. Physical Review E, The American Physical Society, Volume 59, Number 3, March, pp. 3344-3360, 1999. http://www.menem.com/~ilya/digital_library/learning/hershkowitz-nadal.pdf

[8] Kleissner, C. Data mining for the enterprise. System Sciences, *Proceedings of the Thirty-First Hawaii International Conference on,* Volume 7, 6-9 Jan., pp. 295-304, 1998.

[9] Martin, B. Instance-Based learning : Nearest Neighbor With Generalization, Master Thesis, University of Waikato, Hamilton, New Zealand, 1995.

[10] Matsumoto, K. An Experimental Agricultural Data Mining System. Lecture Notes in Computer Science. *Proc. Discovery Science: First International Conference DS'98*, Fukuoka, 439-440, Springer-Verlag GmbH Publishers, Vol. 1532, 1998.

[11] Provost, F. Reviving the Learning Curve: A critical study of machine learning performance. In UT-Austin Data Mining Seminar Series Spring, 2005.

[12] Roy, S. Nearest Neighbor With Generalization,Unpublished, University of Canterbury, Christchurch, New Zealand, 2002.

[13] Silicon Graphics. Mineset User's Guide. Appendix A. Flat File Support for MineSet, 1998.

[14] Thuraisingham, B. A primer for understanding and applying data mining. IT Professional
Volume 2, Issue 1, Jan.-Feb., pp. 28-31, 2000.

[15] Villegas, R..,C. Balmaseda, D. Ponce de León, L. Benítez y R. Marín. Evaluación de la Aptitud Física de las Tierras dedicadas al Cultivo de la Caña de Azúcar: Primera Aproximación. Instit ut oNacional de Investigaciones de la Caña de Azúcar (INICA), La Habana, 2001.

[16] Weiss, S.M. and N. Indurkhyya. Predictive Data Mining. A Practical Guide. Morgan Kaufmann Publishers, San Francisco, 1998.

[17] Williams, G. The Data Mining Catalogue. In Data Mining Desktop Survival Guide, 2005. http://dmsurvivor.sarovar.org/L.html