**Universidad Central de Venezuela**
**Facultad de Ciencias**
**Escuela de Computación**

***Lecturas en Ciencias de la Computación***

***ISSN 1316-6239***

# Minería de datos y agricultura inteligente: Una aplicación en la generación de mapas de cobertura de la tierra a partir de imágenes multiespectrales

Luis Pernía, Sergio Gamarra, Manuel Maneiro, Enrique Egaña, Esteban Álvarez, María Yépez, Yeiremi Freites, Nora Montaño, Haydemar Núñez

**RT 2019-01**

Centro de Investigación en Ingeniería de Software y Sistemas
Caracas, junio de 2019

# Minería de datos y agricultura inteligente: Una aplicación en la generación de mapas de cobertura de la tierra a partir de imágenes multiespectrales

Luis Pernía[4,5], Sergio Gamarra[1,4,5], Manuel Maneiro[4,5], Enrique Egaña[5], Esteban Álvarez[2,5], María Yépez[5], Yeiremi Freites[3,5], Nora Montaño[1,5], Haydemar Núñez[1,5]

[1]Centro de Ingeniería de Software y Sistemas, Escuela de Computación, Universidad Central de Venezuela
[2]Centro de Física Teórica y Computacional, Escuela de Física, Universidad Central de Venezuela
[3]Escuela de Matemáticas, Facultad de Ciencias, Universidad Central de Venezuela
[4]Instituto de Geografía, Maestría en Análisis Espacial y Gestión del Territorio, Universidad Central de Venezuela
[5]Fundación Agrícola Agronerds

**Resumen**. El estudio de los cambios de la cobertura de la tierra y del uso del suelo es de vital importancia para una apropiada gestión de este recurso. Los sistemas de teledetección han resultado una herramienta muy útil para esta tarea, ya que posibilitan el análisis espacial de la cobertura terrestre a partir de imágenes. Sin embargo, el procesamiento e interpretación de estos datos es un proceso complejo, que demanda mucho tiempo y dedicación. Con el fin de apoyar a los profesionales en esta área, en los últimos años se han venido utilizando técnicas de inteligencia artificial, en particular la minería de datos, que se basa en la aplicación de algoritmos de aprendizaje para la adquisición automática de conocimiento. En este trabajo se presenta una aplicación de estas técnicas para la generación de mapas de cobertura de diferentes zonas de Venezuela a partir de imágenes de satélite. Los resultados obtenidos son prometedores, y muestran la factibilidad en cuanto a la utilización de estas técnicas inteligentes para el desarrollo de aplicaciones que puedan apoyar los procesos de toma de decisiones en la planificación y gestión del suelo, de una manera eficiente y oportuna.

**Palabras clave**: Gestión de suelos, teledetección, imágenes mutiespectrales, minería de datos, inteligencia artificial.

## 1. Introducción

A lo largo de los últimos años ha aumentado la necesidad de disponer de información confiable y actualizada sobre los usos y coberturas del suelo, con el fin de elaborar inventarios a partir de los cuales desarrollar programas de planificación, ordenación y gestión de este valioso recurso. En este sentido, la teledetección espacial, técnica que permite adquirir imágenes multiespectrales desde sensores remotos, es una herramienta muy útil para la gestión del territorio, ya que posibilita el análisis espacial de la cobertura terrestre para determinar la dinámica de la población, actividades económicas, evaluación ecológica de la naturaleza, entre otras aplicaciones. Sin embargo, el procesamiento e interpretación de estas imágenes requiere de un profundo conocimiento de técnicas y herramientas de la geografía, la física, computación, estadística y agronomía. En Venezuela, la experiencia en el uso de imágenes multiespectrales es de reciente data y su manipulación se realiza con programas especiales como Erdas Imagine, Envi, entre otros. Todos estos aplicativos tienen muchas bondades, pero

también algunas limitaciones, como el empleo de un gran número de profesionales del área geográfica y respuesta muy lenta para superficies de gran tamaño.

Por otra parte, la gran cantidad de aplicaciones potenciales que tiene la teledetección en el sector agrícola, ha motivado el uso de técnicas inteligentes para tratar la complejidad de los procesos relacionados con esta tecnología, con el fin de construir sistemas de apoyo a la toma de decisiones en este sector. Una vertiente de trabajo y de investigación está relacionada con la aplicación de las técnicas de minería de datos para la adquisición automática de conocimiento a partir de imágenes multiespectrales, con el fin de facilitar la generación de mapas de uso y de cobertura de la tierra, que puedan ser utilizados para la planificación y gestión de las áreas agrícolas, de una manera más eficiente y oportuna.

En este trabajo, se presenta una aplicación del proceso de minería de datos en la generación de mapas de cobertura del suelo de varias zonas del país, a partir de imágenes del sensor Landsat 8 de la Agencia Espacial Norteamericana. El objetivo principal es mostrar las posibilidades que ofrecen las técnicas basadas en inteligencia artificial junto con las tecnologías asociadas a la teledetección, para apoyar la gestión y supervisión del uso de la tierra en Venezuela. El documento se encuentra estructurado de la siguiente manera: en la siguiente sección se presentan los conceptos básicos relacionados con la minería de datos, tipos de tareas y técnicas de aprendizaje, y se mencionan algunas aplicaciones en el sector agrícola que hacen uso de estas técnicas. Luego, en la Sección 3, se describe la metodología de minería de datos que fue utilizada en este trabajo. La Sección 4 presenta en detalle cómo se aplicó esta metodología para la generación de los mapas de cobertura. Por último, se presentan las conclusiones y trabajos futuros.

## 2. Minería de datos

La minería de datos se define como el proceso no trivial de extracción de conocimiento válido, novedoso, potencialmente útil y comprensible, a partir de grandes volúmenes de datos. Con las técnicas de minería de datos es posible identificar, de manera automática, patrones, relaciones, reglas, asociaciones, tendencias y regularidades que puedan resultar útiles para apoyar la toma de decisiones (Aggarwal, 2015; Roiger, 2017; Witten y Frank, 2017). Es un campo de investigación muy activo, debido a la calidad de los resultados que se han reportado en la resolución de problemas complejos y a la diversidad de áreas donde puede ser utilizada.

Los modelos de conocimiento que pueden ser generados con la minería de datos, pueden ser de dos tipos: modelos predictivos y modelos descriptivos. La principal diferencia entre ellos está relacionada con la información de supervisión que reciben los algoritmos durante la fase de modelación de datos (ver Figura 1). Hay varias categorías, las principales son: aprendizaje supervisado y aprendizaje no supervisado.

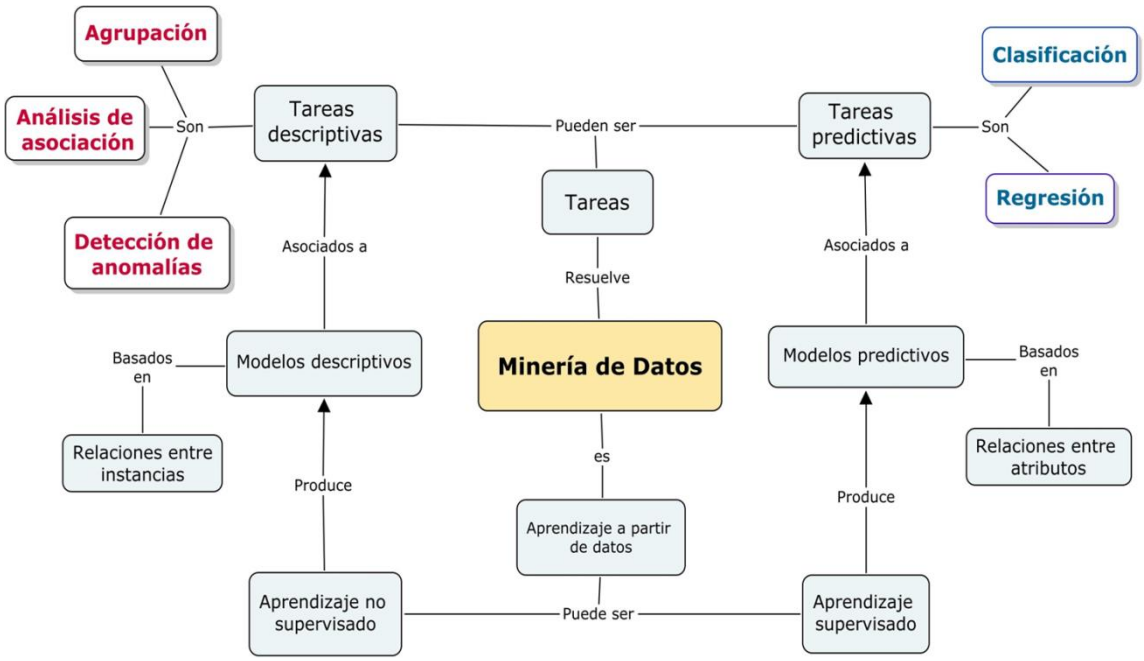



Figura 1. Tareas de la minería de datos.

En un contexto de aprendizaje supervisado se conoce la respuesta correcta o valor de salida para cada instancia del conjunto de datos. Si la salida está descrita en términos cualitativos se define una tarea de clasificación, pero si la salida es una variable numérica se define entonces una tarea de regresión. Muchos problemas se pueden adaptar a esta formulación como, por ejemplo: determinar la aptitud física de la tierra para un determinado cultivo, predecir la cantidad de lluvia a corto plazo a partir de datos climatológicos, clasificar especies a partir de sus características morfológicas, diagnóstico de enfermedades de plantas, pronóstico de ventas de un producto agrícola, entre muchos otros.

Sin embargo, no siempre se conoce la salida para las instancias de un conjunto de datos. En este caso, se puede resolver el problema mediante el aprendizaje no supervisado, donde sólo se toman en cuenta las características o variables de entrada. La tarea podría consistir entonces en describir cómo están relacionados o agrupados los datos. Por ejemplo, es posible encontrar grupos naturales en el conjunto de aprendizaje mediante técnicas de agrupación o determinar cuáles son las asociaciones más frecuentes entre las variables, tarea que se conoce como análisis de asociación. También se incluye en esta categoría la detección de patrones que presentan diferencias significativas con respecto a la mayoría de los datos a través de una detección de anomalías. Ejemplos de situaciones donde esta formulación puede ser aplicada son: descubrimiento de patrones de comportamiento en diferentes contextos agrícolas, identificación del tipo de paisaje, identificación de variedades de semillas, determinación de los productos que se compran juntos más frecuentemente en establecimientos agropecuarios, identificación de patrones inusuales a partir de datos derivados de sensores de monitoreo de cultivos, entre otros.

## 2.1. Técnicas de aprendizaje

Se han propuesto una diversidad de técnicas para el aprendizaje a partir de datos, cada una está dirigida a un lenguaje de representación de conocimiento particular e implanta un procedimiento de optimización con el objetivo de determinar el modelo que mejor se ajuste a los datos. A continuación, se describen brevemente algunas de estas técnicas:

- *Técnicas basadas en reglas de decisión:*

Estas técnicas están dirigidas a tareas de clasificación y el lenguaje de representación que utilizan son reglas del tipo *Si condición entonces Clase,* las cuales identifican las relaciones entre los atributos de un conjunto de datos y la etiqueta de clase. En general, utilizan algoritmos que dividen el espacio de entrada en sub-espacios más pequeños, buscando que todas las instancias que pertenecen a cada subconjunto estén asociadas a una sola clase; estos algoritmos se conocen como *algoritmos de cobertura*. Entre ellos, uno de los más conocidos es RIPPER, el cual utiliza una métrica de evaluación que da preferencia a reglas con una alta cobertura y exactitud. Construye una lista de reglas de decisión y puede ser aplicado a conjuntos de datos con atributos nominales y numéricos.

- *Técnicas basadas en árboles de decisión*

Estas técnicas están dirigidas a tareas de clasificación y regresión. El lenguaje de representación son árboles de decisión, los cuales son estructuras jerárquicas consistentes de nodos y arcos dirigidos, donde el nodo raíz y los nodos internos están asociados a condiciones o test de atributos que permite separar registros con características similares. Los nodos hojas contienen una etiqueta de clase o valor numérico, que se utiliza para generalizar una instancia que satisface las condiciones especificadas en la rama del árbol asociada

En el marco de clasificación, estos algoritmos construyen un árbol de decisión de manera recursiva y, en cada iteración, dividen el espacio de entrada en conjuntos más puros, donde la pureza está determinada por la distribución de las clases en el nodo actual. Para esto, ejecutan un procedimiento de optimización local basado en una medida de calidad con el fin de seleccionar el atributo que será utilizado como test de partición.

Hay muchas medidas que pueden utilizarse para determinar la mejor manera de dividir los registros (entropía, índice GINI, error esperado, entre otras), las cuales se basan en la probabilidad observada de cada una de las clases en el sub-conjunto de registros que llegan al nodo actual. Una característica que distingue a los diferentes algoritmos que se han propuesto es precisamente la medida calidad para seleccionar el test; por ejemplo al algoritmo C4.5, uno de los más populares, utiliza la entropía.

- *K vecinos más cercanos.*

El algoritmo K-vecinos más cercanos (K-NN), está dirigido a tareas de clasificación y regresión. Se basa en el aprendizaje basado en instancias o por analogía, es decir, la predicción o generalización de una etiqueta de clase o un valor numérico a un nuevo dato, toma en cuenta

las salidas asociadas a las instancias más cercanas a éste en el espacio de entrada; estas se conocen como los "vecinos más cercanos". Un aspecto a resaltar es que estos algoritmos no construyen un modelo global a partir del conjunto de datos antes de realizar cualquier predicción; más bien, almacenan este conjunto y realizan una aproximación local al dato a generalizar en el momento en que este se recibe. Es por esto que también se les conoce como métodos perezosos o *lazy*.

La "cercanía" es definida en términos de una métrica de distancia (como la distancia euclídea). Para determinar el conjunto de vecinos más cercanos, es necesario calcular los valores de proximidad del dato a procesar con cada instancia del conjunto de datos. Una vez determinados los K vecinos, se utiliza algún criterio de decisión para determinar la salida a asignar. Por ejemplo, en el ámbito de clasificación, se puede seleccionar la clase más frecuente en el conjunto de vecinos (lo cual se conoce como criterio de decisión por mayoría).

- *Combinación de clasificadores*

Los métodos basados en la combinación (*ensembles*) de clasificadores construyen un modelo compuesto con el fin de mejorar el rendimiento de predicción (Figura 2). A partir de los datos de aprendizaje se realiza un muestreo para construir el conjunto que utilizará cada clasificador; esto con el fin de crear diversidad entre los diferentes modelos. La forma como se realiza este muestreo, junto con los tipos de algoritmos utilizados para aprender los diferentes clasificadores y la manera como se integran las predicciones de estos para producir una única respuesta dan lugar a diferentes esquemas de combinación. Un ejemplo de ellos es *Random Forest*, el cual realiza un muestreo con reemplazo para la construcción de los subconjuntos de datos y una selección aleatoria de los atributos (candidatos a test de división), a partir de los cuales se construyen los diferentes modelos basados en árboles de decisión. Para determinar el valor de salida de la combinación se utiliza un esquema de votación simple y se selecciona la clase que tenga más votos.

- *Técnicas de agrupación*.

Los algoritmos de agrupación dividen el conjunto de datos en subconjuntos o grupos (*cluster*s) de objetos que comparten características comunes. Los datos son agrupados utilizando sólo la información que los describe, por lo que estas técnicas pertenecen al conjunto de paradigmas que utilizan aprendizaje no supervisado. La idea es que los objetos dentro de un grupo sean similares (o estar relacionados), y diferentes (o no relacionados) a objetos de otros grupos. Para la construcción de los grupos se utilizan medidas o criterios de similitud, que dependen del tipo de variable presente en el conjunto de datos. Es importante resaltar que un grupo puede ser considerado como una clase implícita; por lo tanto, la agrupación puede ser visualizada como una forma de clasificación, pero no supervisada.

Uno de los algoritmos más utilizados es K-medias, el cual utiliza un prototipo o centro para representar a los grupos, los cuales son exclusivos (un objeto sólo puede pertenecer a un grupo). Sin embargo, ha sido extendido para encontrar grupos difusos; este algoritmo se conoce como *Fuzzy K-medias.*

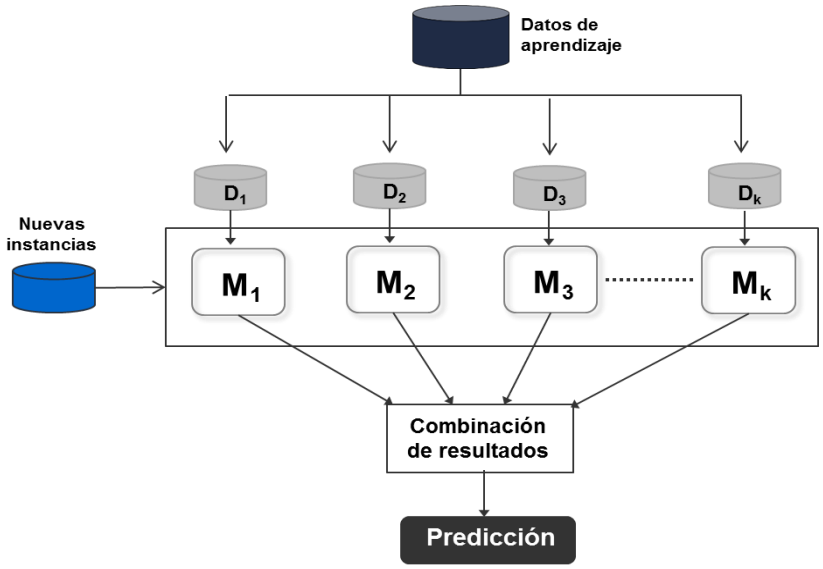



Figura 2. Esquema de combinación de clasificadores

- *Redes neuronales artificiales*

Paradigma que trata de emular la estructura y características del sistema nervioso, con el fin de alcanzar una funcionalidad similar en la resolución de problemas. Representan, más que una sola técnica de aprendizaje, una familia de modelos que han sido utilizados en una amplia variedad de problemas de clasificación, regresión y agrupación.

Las redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (neuronas artificiales), con capacidad de adaptación en respuesta a las entradas externas; la interconexión es por capas, como se observa en la Figura 3.

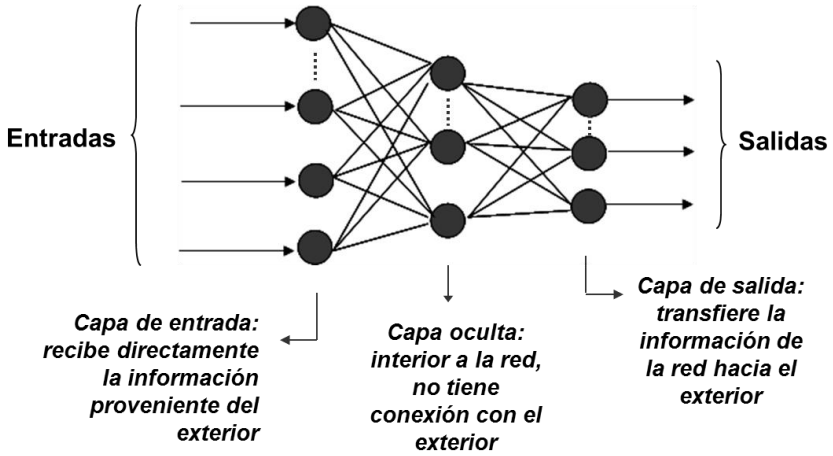



Figura 3. Arquitectura de una red neuronal

En el contexto supervisado, una de los modelos más utilizados es el Perceptrón Multicapa (MLP), junto con el algoritmo de Retropropagación del Error (*Backpropagation*)

para el entrenamiento de la red. Para el caso no supervisado se puede citar a los Mapas Autoorganizativos de Kohonen, modelo con una gran aplicabilidad práctica en agrupación y reconocimiento de patrones, visualización y reducción de la dimensionalidad.

- *Redes de aprendizaje profundo (Deep learning):*

El aprendizaje profundo (*Deep learning*, DL) es un tipo de aprendizaje que construye modelos compuestos por una jerarquía de funciones que son aprendidas a partir de los datos [Aggarwal, 2018; Goodfellow *et al*., 2016]. ¿Cuál es la idea que subyace en este paradigma? Muchas tareas de aprendizaje pueden resolverse definiendo el conjunto correcto de características, las cuales se le proporcionan al algoritmo de aprendizaje. Sin embargo, para tareas complejas, es difícil saber cuáles características deberían ser extraídas. Una solución a este problema es usar el algoritmo no sólo para descubrir el mapeo fundamental de los datos, sino también para aprender la representación en sí misma. Esto se conoce como *aprendizaje de representación*.

El ejemplo por excelencia de DL es el Perceptron Multicapa (MLP), en el cual cada capa representa una nueva representación de los datos. Así, el MLP implanta una composición o "cadena" de funciones y la longitud de esta "cadena" proporciona la profundidad. Y es a partir de esta terminología que surge el nombre de "redes profundas" (ver Figura 4). Entre las redes profundas se pueden citar a las redes convolucionales, utilizadas principalmente para aplicaciones de visión artificial y procesamiento de imágenes, las redes recurrentes, muy utilizadas en el procesamiento del lenguaje natural, los auto-codificadores (*autoencoder*s), las redes de creencia profundas, entre otras.

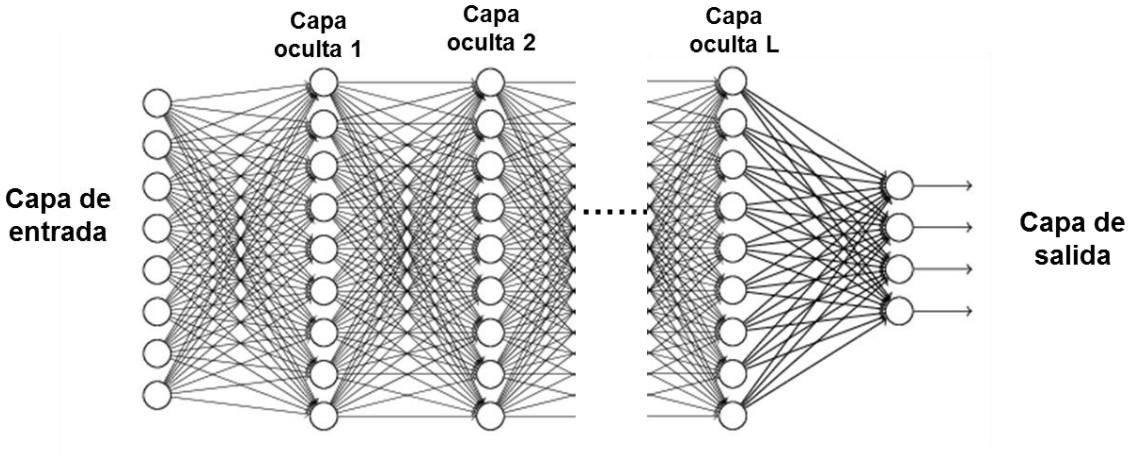



Figura 4. Arquitectura de una red profunda.

## 2.2. Minería de datos y uso de la tierra

La minería de datos se ha venido aplicando ampliamente, y cada vez con mayor fuerza, para el desarrollo de sistemas que faciliten y apoyen la toma de decisiones en la gestión, uso y preservación de la tierra (Mucherino *et al*, 2009). Una tendencia es combinar la minería de datos con las tecnologías y métodos asociados a la percepción remota, utilizando como fuente de datos las imágenes satelitales de la zona de estudio, por ejemplo, para la detección de

cambios en la cobertura vegetal (Antunes, 2018; Martínez *et al*, 2017; Yirsaw *et al*, 2017), clasificación del tipo de cobertura vegetal y uso de la tierra (Borràs *et al*, 2017; Helbert *et al*, 2017; Riegler-Nurscher *et al*, 2018; Yang *et al*, 2018), identificación de tipos de tierra raras (Bogner *et al*, 2018), estudio y monitoreo de la desforestación (Lu *et al*, 2012), determinación de futuros usos de la tierra (Sopan *et al*, 2017), entre muchas otras aplicaciones. Las técnicas de aprendizaje utilizadas en el análisis de las imágenes van desde los métodos tradicionales como algoritmos de árboles de decisión, basados en reglas, algoritmos de agrupación, como k-medias y *Fuzzy* k-medias, redes neuronales, tanto supervisadas como no supervisadas y, como última tendencia, redes profundas (*Deep Learning*), en particular redes profundas convolucionales (Kamilaris *et al*, 2018).

## 3. Marco metodológico

En este trabajo se utilizó la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) para minería de datos. Esta metodología surgió como un proyecto de un consorcio europeo bajo la iniciativa de financiación ESPRIT. Actualmente, es promovida por IBM[1] y utilizada en su suite de minería de datos IBM® SPSS Modeler.

El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre éstas. El ciclo de modelado de CRISP-DM contiene seis fases (ver Figura 5), con flechas que indican las dependencias más importantes y frecuentes entre ellas. La secuencia de las fases no es estricta, además, el modelo se puede personalizar y adaptar fácilmente a los objetivos del problema que se intenta resolver. Cada fase está comprendida por un conjunto de tareas, con una descripción de los productos y reportes que deben ser generados. A continuación, se describen brevemente cada una de estas fases:

- Fase I. Comprensión del Negocio. Esta fase inicial se enfoca en la comprensión de los objetivos y requisitos del proyecto, desde el punto de vista de la organización. En general, conlleva un proceso de ingeniería del conocimiento para conocer la situación actual y las perspectivas de solución, así como conocimiento del dominio a partir de expertos. Este conocimiento luego se traslada a uno o más objetivos desde el punto de vista de la minería de datos (enfoque analítico) y se identifican las fuentes para la recopilación de los datos. Por último, se diseña un plan preliminar para alcanzar los objetivos, en el cual se especifica, entre otros, el software y recursos a utilizar, capacidades de cómputo necesarias, tomando en cuenta las características del conjunto de datos (tamaño y dimensionalidad), así como otros requerimientos, estimación de costos y tiempos de ejecución.

- Fase II. Comprensión de los datos. Esta fase comienza con la recolección de los datos a partir de las fuentes que se hayan identificado en la fase anterior, y continúa con las actividades que permiten familiarizarse con estos, la verificación de la calidad de los datos y la aplicación de técnicas de análisis exploratorio, con el fin obtener un conocimiento básico de estos

[1] https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html

(características, tendencias, correlaciones, entre otros). Esto ayuda a determinar si el conjunto de datos es adecuado para resolver el problema y permite identificar errores.

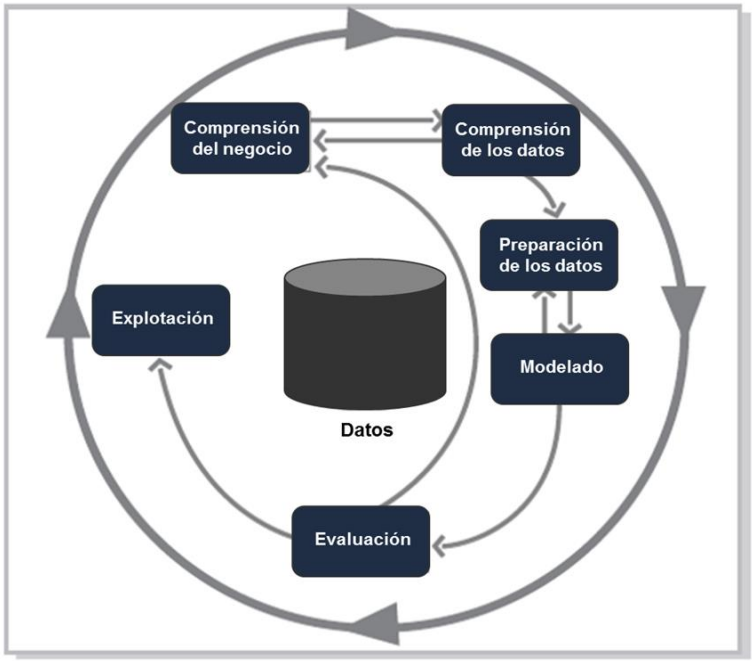



Figura 5. Fases de la metodología CRISP-DM

- Fase III. Preparación de datos. Cubre todas las actividades necesarias para construir el conjunto final de datos (el que será utilizado para la construcción de los modelos predictivos o descriptivos). Las tareas incluyen la limpieza de los datos, la selección de datos y de atributos, la ingeniería de características, así como la aplicación de técnicas de transformación según sea el caso (por ejemplo, para cambiar la forma de los datos y mejorar su representación para el algoritmo de aprendizaje). Hay que tomar en cuenta que si los datos son complejos, se debe buscar una representación estructurada.

- Fase IV. Modelado. En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema y se calibran sus parámetros a valores óptimos.

- Fase V. Evaluación. Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, y comparar el modelo obtenido con los objetivos iniciales. Un aspecto clave es determinar si hay alguna cuestión importante que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso del análisis de datos. Es importante contrastar el conocimiento adquirido con cualquier conocimiento previo que esté disponible y realizar la verificación con los expertos.

- Fase VI. Explotación. La creación del modelo no es generalmente la etapa final del proyecto. Incluso si el objetivo del modelo es aumentar el conocimiento de los datos, el conocimiento

obtenido tendrá que ser organizado y presentado de modo que pueda ser utilizado por los usuarios finales. Dependiendo de los requerimientos, el resultado de esta fase puede culminar con el desarrollo de un software especializado; por ejemplo, sobre el modelo final pueden desarrollarse aplicaciones que automaticen y/o apoyen los procesos de toma de decisiones. También, se deben desarrollar los mecanismos de monitorización que permitan determinar la vigencia del modelo en el tiempo.

## 4. Generación de los mapas de cobertura de la tierra

A continuación, se explica cómo se aplicó la metodología para desarrollar mapas de cobertura de la tierra de algunas zonas del país.

### 4.1. Comprensión del negocio.

La determinación de mapas de cobertura y uso actual de la tierra que permitan una mejor planificación y gestión de este recurso, es un objetivo estratégico que posee características de gran dificultad, al ser necesario recurrir a numerosos profesionales especializados y acceder a información de imágenes de los sensores como Landsat 8, de la Agencia espacial Norteamericana y Sentinel 2, de la Agencia espacial Europea y en el futuro, los productos del satélite Sucre de la Agencia Espacial Bolivariana de la República Bolivariana de Venezuela. Debido a la complejidad inherente al procesamiento e interpretación de las imágenes para la obtención del conocimiento necesario para una toma de decisiones adecuada y oportuna, se consideró utilizar técnicas de la inteligencia artificial, en particular la minería de datos, para realizar el análisis de las imágenes y generar, de manera automática, mapas de cobertura de diferentes zonas del país. En una primera instancia se utilizarán imágenes del sensor Landsat 8 y el software Weka, para la construcción de los modelos de conocimiento, así como el sistema de información geográfica ArcGis.

Una imagen de satélite por lo general tiene múltiples bandas que representan distintas longitudes de onda desde las porciones ultravioleta hasta las visibles e infrarrojas del espectro electromagnético. En el caso de Landsat 8, el cual es un satélite de observación terrestre que posee 2 sensores (OLI y TIRS), las imágenes son datos recopilados desde once bandas distintas; es decir, posee una resolución espectral de 9 + 2 bandas, con una longitud de onda que varía desde los 0.433 μm a los 1.390 μm para el sensor OLI, y de 10.30μm a 12.50 μm para el sensor TIRS. La resolución espacial para las bandas de 1 a 7 y 9 es de 30 metros, para la banda 8 (pancromática) es de 15 metros y para las bandas 10 y 11 es de 100 metros.

Cada banda está representada en una escala de grises que representan los niveles digitales disponibles para mostrar los detalles de la imagen, los cuales se expresan en términos de dígitos binarios. Para el caso del Landsat 8 se tiene una resolución radiométrica de 16 bits, es decir, la escala de grises se extiende de 0 a 65.535 para cada pixel.

La resolución temporal del satélite Landsat 8 es de 16 días, lo que permite obtener información actualizada de una misma porción de la tierra aproximadamente 2 veces por mes.

Como sistema de referencia usa el WGS84 o Sistema Geodésico Mundial, ya que, al contrario de muchos sistemas de referencia, este permite su uso en cualquier parte del planeta.

### 4.2. Comprensión de los Datos.

Un raster, es cualquier tipo de imagen digital representada en una cuadricula, que divide el espacio en celdas regulares donde cada una de ellas representa un único valor asociado a una banda (Figura 6). Estas celdas de igual tamaño distribuidas en toda la imagen reciben el nombre de pixel. Para ser utilizado en un sistema de información geográfica (SIG), un raster debe poseer información de referencia espacial que la correlacione con el espacio real. Esta información espacial viene expresada en:

- Un sistema de referencia, expresado por la proyección y el geoide utilizado por la imagen para conseguir una posición univoca en la esfera terrestre.
- La ubicación, representada por el valor de latitud y longitud de la esquina superior izquierda de la imagen.
- La resolución espacial, referida al área terrestre representada por cada píxel.
- La resolución radiométrica, representada por el rango de valores del pixel, que determina el nivel de detalle de la imagen.
- Además de estas características, los raster provenientes de sensores remotos poseen otras particularidades como la resolución espectral, la cual se refiere al número y ancho de las bandas espectrales registradas por un sensor; y la resolución temporal, que viene dada por la periodicidad con que el sensor adquiere imágenes de una misma porción de la superficie terrestre.

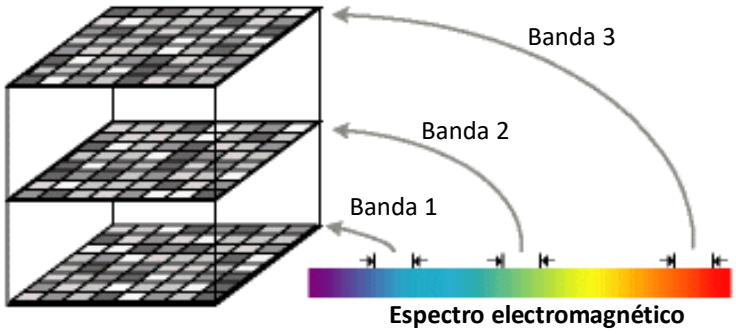



Figura 6. Rásteres con múltiples bandas

Cada banda de los sensores del Landsat 8 proporciona información diferente de acuerdo a los intereses del estudio. La banda 1 (azul profundo) es útil para estudios costeros y aerosoles. Las bandas del espectro visible, azul, verde y rojo, bandas 2, 3 y 4 respectivamente, sirven para identificar cuerpos de agua, vegetación, suelos desnudos, fuego, área urbana, entre otros, mientras que las bandas del infrarrojo cercano y medio (bandas 5, 6 y 7) permiten identificar características no visibles asociadas al estado de vigor de la vegetación y a los suelos desnudos. La banda 9 es útil para la detección de cirrus. Las bandas 10 y 11,

provenientes del sensor TIRS, muestran la temperatura a nivel del suelo y de la atmósfera respectivamente, y se utiliza para estudios de actividad volcánica, estudios urbanos y predicción del tiempo.

### 4.3. Preparación de los Datos.

Para poder aplicar los algoritmos de aprendizaje es necesario obtener una representación estructurada a partir de las imágenes. Esta representación es una matriz o tabla de dos dimensiones, en la cual cada banda representará un atributo (columna) y cada pixel será una fila, caracterizada por los valores de las diferentes bandas. Además, en este trabajo se consideró utilizar un aprendizaje supervisado, lo que conlleva atribuir a cada pixel una clasificación (tipo de cobertura), que será una columna adicional. La tabla debe ser construida de tal forma que el pixel de cada banda se encuentre en la misma ubicación, más el pixel del raster que contenga la clasificación muestreada.

Para obtener el raster que contenga esta clasificación se debe partir de la construcción de este a partir de la observación de la realidad y de la selección de las muestras de clases por un experto. La mejor manera de hacer visibles las características reales de la imagen de satélite es haciendo una combinación de bandas con un software especializado, para el caso del sensor OLI la combinación 4-3-2, proporciona el color verdadero, sobre esta imagen se deben extraer una muestra de las clases, conocidas como áreas de interés (ROI), que para el caso de este estudio serán: cuerpos de agua, área urbana, suelos desnudos, vegetación alta, vegetación baja, nubes y sombra de nubes (Figura 7).

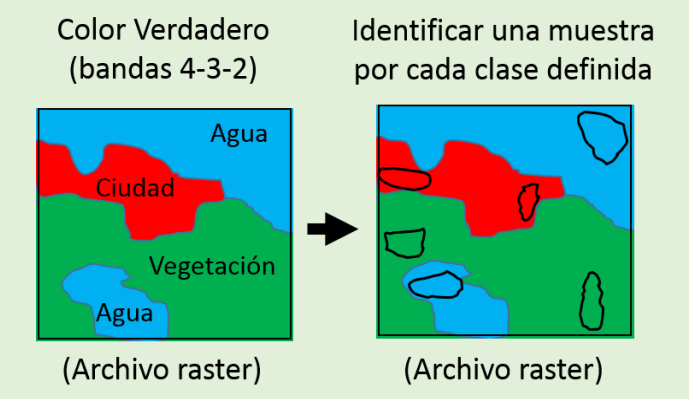



Figura 7. Identificación de muestras por clase.

Una vez definida esta muestra, en un software especializado en manejo de información geográfica (SIG), debe crearse un archivo vectorial en el cual se asocia el valor de la clase a cada registro creado (Figura 8). La Figura 9 indica cómo se generó la muestra, utilizando este procedimiento, para una de las zonas de estudio.

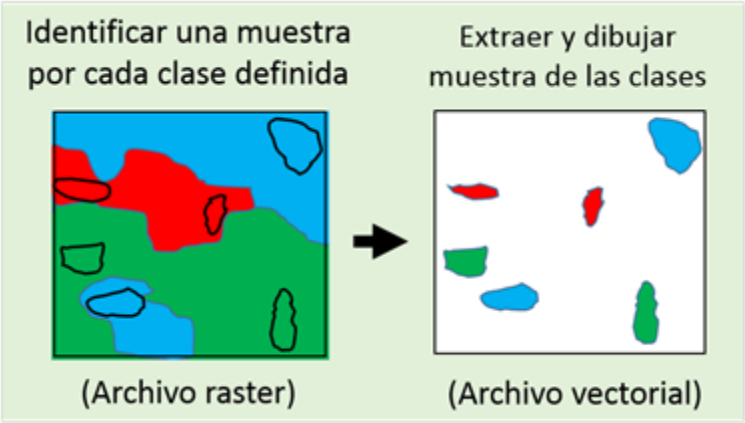



Figura 8. Archivo vectorial de muestras de las clases

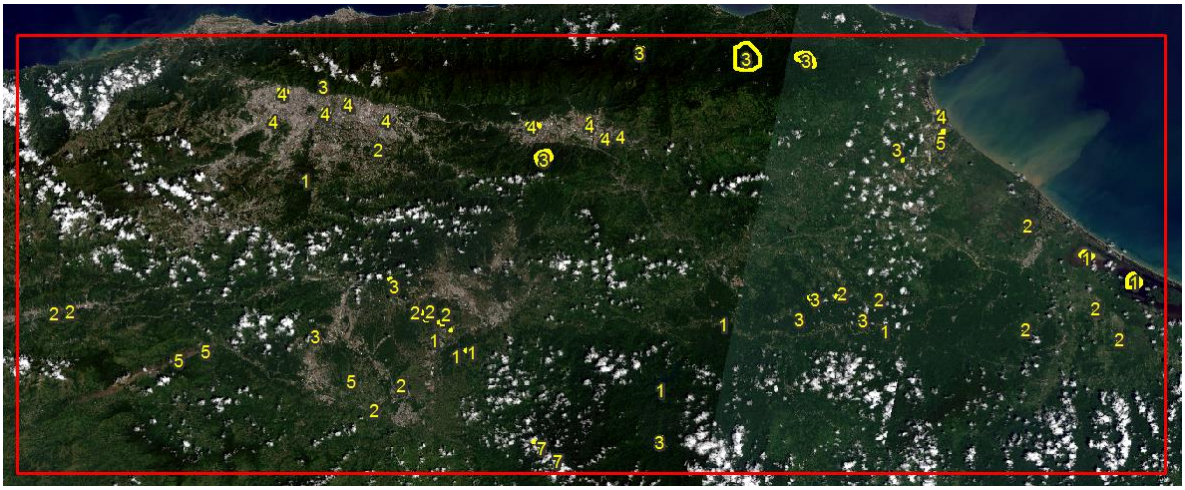


Figura 9. Generación de la muestra en una de las zonas de estudio.

El archivo vectorial obtenido se transforma a formato raster, almacenando como valor del pixel el valor entero asignado nominalmente a la clase (ver Figura 10), conservando las mismas características de resolución, ancho, largo y posición de la escena correspondiente al sensor OLI, de tal manera de hacer posible su comparación con las bandas de dicho sensor.

Una vez se tienen todos los atributos (bandas + clases) estructurados de tal forma que se puedan correlacionar espacialmente (mismo tamaño y ubicación), se hace necesario acceder de forma eficiente a los datos (valores de los pixeles), es por ello que se procede a transformar todas las imágenes de formato raster a ASCII (Figura 11). Esto se realiza con funciones de transformación propias del software especializado en tratamiento de información geográfica.

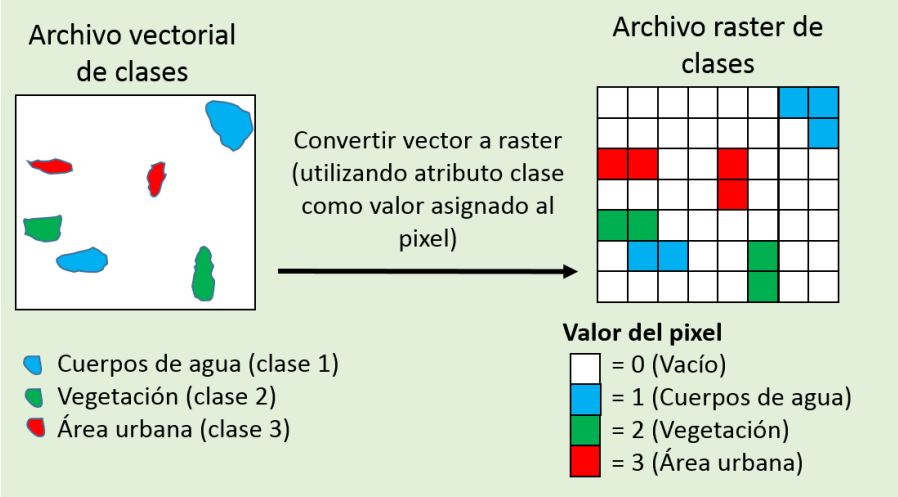



Figura 10. Transformación del archivo vectorial a raster.

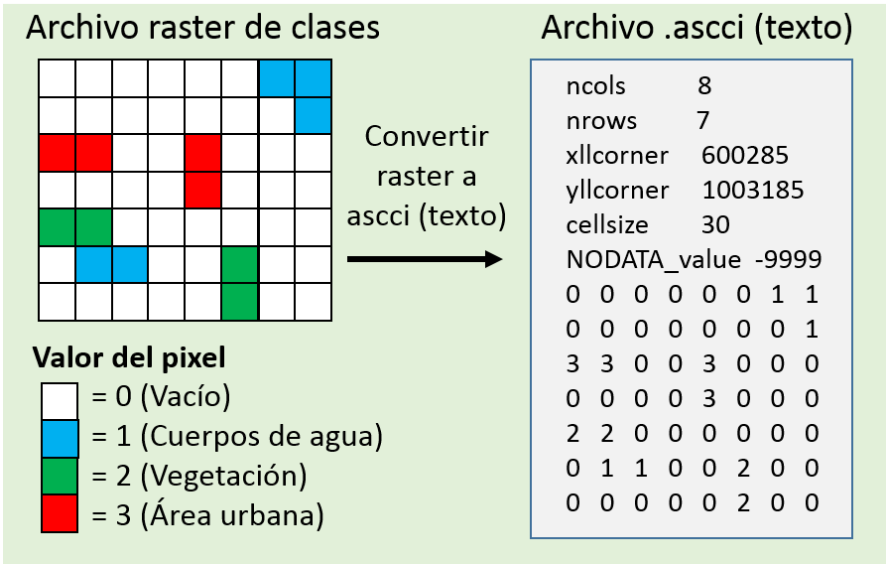



Figura 11. Valores de los píxeles transformados a texto.

El archivo ASCII consta de un encabezado que posee la información espacial de la imagen: número de columnas, número de filas, posición en $x, y$ de la esquina superior derecha, y la resolución espacial del pixel. La matriz de datos contendrá todos los atributos relevantes para el proceso de minería de datos; para el caso de la clasificación supervisada uno de estos campos será la clase, que representa el significado que le da el experto a la combinación de los atributos expuesta en cada registro. Como se dijo anteriormente, cada atributo vendrá representado por una banda del sensor, la combinación de estas bandas, expresado en el valor del pixel de cada celda, se debe vincular con el valor del pixel del atributo clase, cuidando de conservar la misma posición en la estructura de los archivos (ver Figura 12).

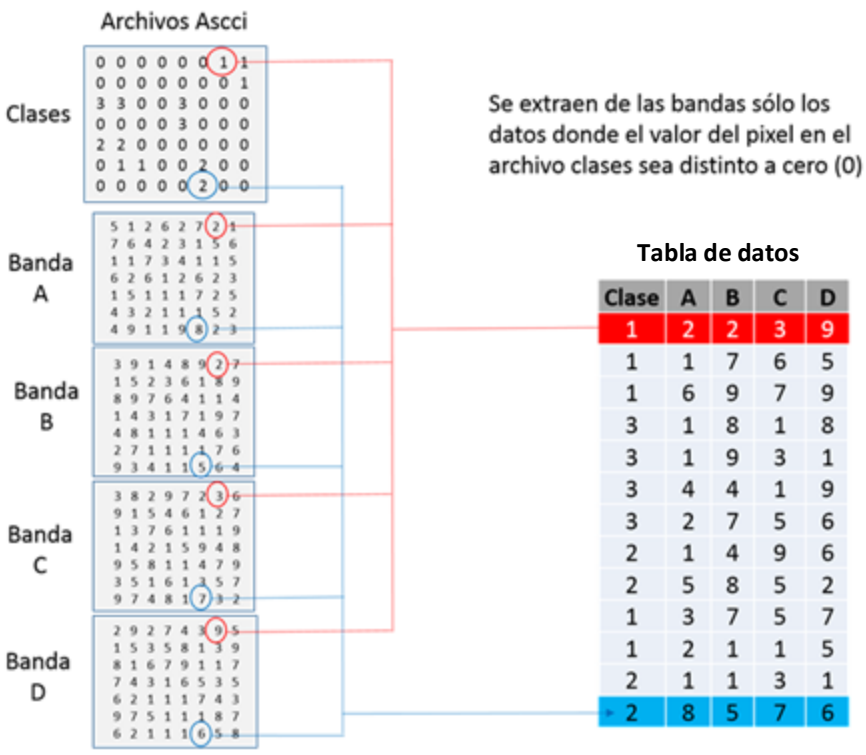



Figura 12. Construcción de la matriz de datos para imágenes de satélite

Es importante destacar que la construcción de esta matriz de datos para determinar cobertura de la tierra supone un ejercicio de correlación espacial y atributiva, vinculada a las clases. Una vez construida, esta pierde todo tipo de vinculación con el espacio y sólo muestra las relaciones que existen entre los valores de radianza del sensor y las clases identificadas. Para este trabajo se consideró la utilización de todas las bandas espectrales correspondientes al sector visible del sensor OLI y el modelo digital de elevación a 30 metros de resolución. Por otra parte, cada banda fue procesada mediante la aplicación ArcGis 10.2 exportando la matrix en código ASCII.

Aplicando este proceso de transformación a la muestra se logra construir un conjunto de datos (píxeles) caracterizado por 11 atributos (bandas) más un atributo que representa la clase (tipo de cobertura). La codificación que se utilizó para esta fue: cuerpos de agua (1), vegetación baja (2), vegetación alta (3), suelos desnudos (4), área urbana (5), nubes (6) y sombra de nubes (7). La Tabla 1 muestra un extracto de este conjunto.

### 4.4. Modelado.

Para el desarrollo del modelo de clasificación se utilizó el aplicativo Weka, (Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato») el cual es una plataforma de software para el aprendizaje automático y minería de datos escrito en Java y desarrollado en la Universidad de Waikato. En particular, se aplicó la técnica supervisada de minería de datos RIPPER, produciendo un

modelo basado en reglas de clasificación. Este algoritmo fue seleccionado ya que produce un modelo interpretable y es adecuado a la cantidad de datos disponible. La Figura 13 muestra un extracto del modelo obtenido.

Tabla 1. Extracto del conjunto de datos.

| 1: banda1 | 2: banda2 | 3: banda3 | 4: banda4 | 5: banda5 | 6: banda6 | 7: banda7 | 8: banda9 | 9: banda10 | 10: banda11 | 11: clase |
|---|---|---|---|---|---|---|---|---|---|---|
| Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Numeric | Nominal |
| 10561.0 | 9639.0 | 8095.0 | 6960.0 | 6588.0 | 6075.0 | 5796.0 | 5025.0 | 26998.0 | 24329.0 | 1 |
| 10618.0 | 9683.0 | 8150.0 | 7018.0 | 6640.0 | 6076.0 | 5819.0 | 5021.0 | 26998.0 | 24316.0 | 1 |
| 10621.0 | 9686.0 | 8158.0 | 7020.0 | 6627.0 | 6129.0 | 5848.0 | 5030.0 | 26999.0 | 24311.0 | 1 |
| 10580.0 | 9657.0 | 8152.0 | 7014.0 | 6606.0 | 6110.0 | 5844.0 | 5041.0 | 27000.0 | 24317.0 | 1 |
| 10601.0 | 9667.0 | 8157.0 | 7004.0 | 6618.0 | 6104.0 | 5856.0 | 5035.0 | 27002.0 | 24324.0 | 1 |
| 10611.0 | 9679.0 | 8149.0 | 7000.0 | 6598.0 | 6074.0 | 5825.0 | 5047.0 | 27003.0 | 24328.0 | 1 |
| 10604.0 | 9658.0 | 8162.0 | 7001.0 | 6567.0 | 6082.0 | 5800.0 | 5039.0 | 27004.0 | 24330.0 | 1 |
| 13120.0 | 12858.0 | 13884.0 | 14840.0 | 17403.0 | 13881.0 | 11092.0 | 5038.0 | 27992.0 | 24867.0 | 5 |
| 13802.0 | 13461.0 | 14311.0 | 14403.0 | 14477.0 | 10754.0 | 8921.0 | 5036.0 | 27946.0 | 24810.0 | 5 |
| 12304.0 | 11916.0 | 12932.0 | 13690.0 | 16464.0 | 15696.0 | 13118.0 | 5035.0 | 27877.0 | 24934.0 | 5 |
| 13206.0 | 12774.0 | 13500.0 | 13666.0 | 14079.0 | 11487.0 | 9990.0 | 5060.0 | 27790.0 | 24884.0 | 5 |
| 14381.0 | 14123.0 | 14341.0 | 13684.0 | 12325.0 | 7817.0 | 7008.0 | 5055.0 | 27691.0 | 24824.0 | 5 |
| 10136.0 | 9157.0 | 8300.0 | 7647.0 | 11111.0 | 10168.0 | 7996.0 | 5044.0 | 27534.0 | 24619.0 | 3 |
| ⋮ | | | | | | | | | | |
| 10523.0 | 9599.0 | 9542.0 | 8123.0 | 23868.0 | 15307.0 | 9477.0 | 5089.0 | 27070.0 | 24076.0 | 2 |
| 10449.0 | 9509.0 | 9488.0 | 7897.0 | 24466.0 | 15554.0 | 9558.0 | 5087.0 | 27066.0 | 24074.0 | 2 |
| 10454.0 | 9551.0 | 9454.0 | 8096.0 | 21826.0 | 15301.0 | 9621.0 | 5084.0 | 27067.0 | 24067.0 | 2 |
| 10485.0 | 9608.0 | 9490.0 | 8352.0 | 20238.0 | 15402.0 | 9862.0 | 5096.0 | 27102.0 | 24062.0 | 2 |
| 9400.0 | 8311.0 | 7300.0 | 6306.0 | 9943.0 | 6604.0 | 5597.0 | 5066.0 | 25895.0 | 23562.0 | 6 |
| 9362.0 | 8283.0 | 7245.0 | 6278.0 | 9551.0 | 6504.0 | 5550.0 | 5071.0 | 25854.0 | 23534.0 | 6 |
| 9357.0 | 8286.0 | 7225.0 | 6294.0 | 9267.0 | 6391.0 | 5516.0 | 5082.0 | 25824.0 | 23503.0 | 6 |
| 9363.0 | 8277.0 | 7223.0 | 6273.0 | 9753.0 | 6419.0 | 5509.0 | 5069.0 | 25798.0 | 23474.0 | 6 |
| 9339.0 | 8260.0 | 7193.0 | 6271.0 | 10047.0 | 6381.0 | 5481.0 | 5070.0 | 25776.0 | 23448.0 | 6 |
| ⋮ | | | | | | | | | | |
| 15496.0 | 14960.0 | 14412.0 | 14210.0 | 21340.0 | 17042.0 | 13830.0 | 5082.0 | 24410.0 | 22542.0 | 7 |
| 16071.0 | 15555.0 | 14673.0 | 14678.0 | 21745.0 | 17472.0 | 14264.0 | 5097.0 | 24408.0 | 22568.0 | 7 |
| 17934.0 | 17809.0 | 16917.0 | 16982.0 | 24335.0 | 19163.0 | 16027.0 | 5085.0 | 24420.0 | 22546.0 | 7 |
| 17818.0 | 17213.0 | 16720.0 | 16444.0 | 23411.0 | 18463.0 | 14932.0 | 5088.0 | 24448.0 | 22526.0 | 7 |
| 10220.0 | 9175.0 | 8551.0 | 7199.0 | 18693.0 | 11465.0 | 7721.0 | 5099.0 | 26833.0 | 23925.0 | 2 |
| 10201.0 | 9159.0 | 8507.0 | 7128.0 | 18050.0 | 11085.0 | 7544.0 | 5096.0 | 26847.0 | 23897.0 | 2 |
| 10316.0 | 9355.0 | 9006.0 | 7421.0 | 21306.0 | 13032.0 | 8355.0 | 5106.0 | 26874.0 | 23896.0 | 2 |
| 10466.0 | 9560.0 | 9508.0 | 7741.0 | 23628.0 | 14400.0 | 9003.0 | 5099.0 | 26906.0 | 23935.0 | 2 |

### 4.5. Evaluación.

A continuación, se presentan los resultados obtenidos de la evaluación del modelo utilizando la técnica de validación cruzada con 10 particiones:

Matriz de confusión:

```
   a    b    c    d    e    f    g   <-- classified as
3559    0    1    1    2    5    0 |    a = 1
   3 3534   17    4    8    1    1 |    b = 2
   1   26 3519    3    3    7    9 |    c = 3
   0   17    2 3485   55    1    8 |    d = 4
   7   20    3   60 3477    1    0 |    e = 5
   4    1   12    0    0 3551    0 |    f = 6
   0    0    1    4    2    2 3559 |    g = 7
```

Exactitud: 0.9883

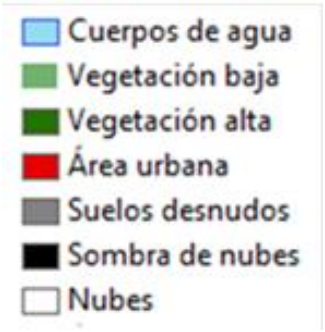
Se puede observar que el modelo obtiene un rendimiento de clasificación del 98,83. Además, de la matriz de confusión se desprende que el modelo tiende un poco a confundir suelos desnudos con las zonas urbanas y en mucho menos grado la vegetación alta con la baja.

```
Si (banda5 <= 7541) => clase=1
Si (banda6 <= 6022) and (banda10 >= 27138) and (banda10 <= 27514) => clase=1
.
.
(banda1 >= 15833) and (banda11 <= 23223) => clase=7
(banda6 <= 7813) and (banda1 >= 9258) and (banda4 <= 6974) => clase=6
.
.
(banda1 >= 15163) and (banda5 >= 25824) and (banda7 <= 28723) => clase=7
(banda7 <= 6248) and (banda1 >= 9210) and (banda11 <= 23443) => clase=6
.
.
(banda4 <= 9201) and (banda4 >= 7029) and (banda5 >= 18702) and (banda1 >= 9807) and (banda7 <= 9388) => clase=2
(banda4 >= 7029) and (banda6 >= 12776) and (banda4 <= 8774) and (banda9 <= 5058) => clase=2
.
.
(banda6 >= 17672) and (banda1 <= 11236) and (banda9 <= 5064) and (banda7 <= 16078) => clase=5
(banda6 >= 17186) and (banda1 <= 11291) and (banda7 <= 14727) and (banda5 <= 17754) => clase=5
.
.
(banda3 <= 8550) => clase=3
(banda10 <= 25055) => clase=3
.
.
De lo contrario clase=4
```

Figura 13. Extracto del modelo obtenido con RIPPER.

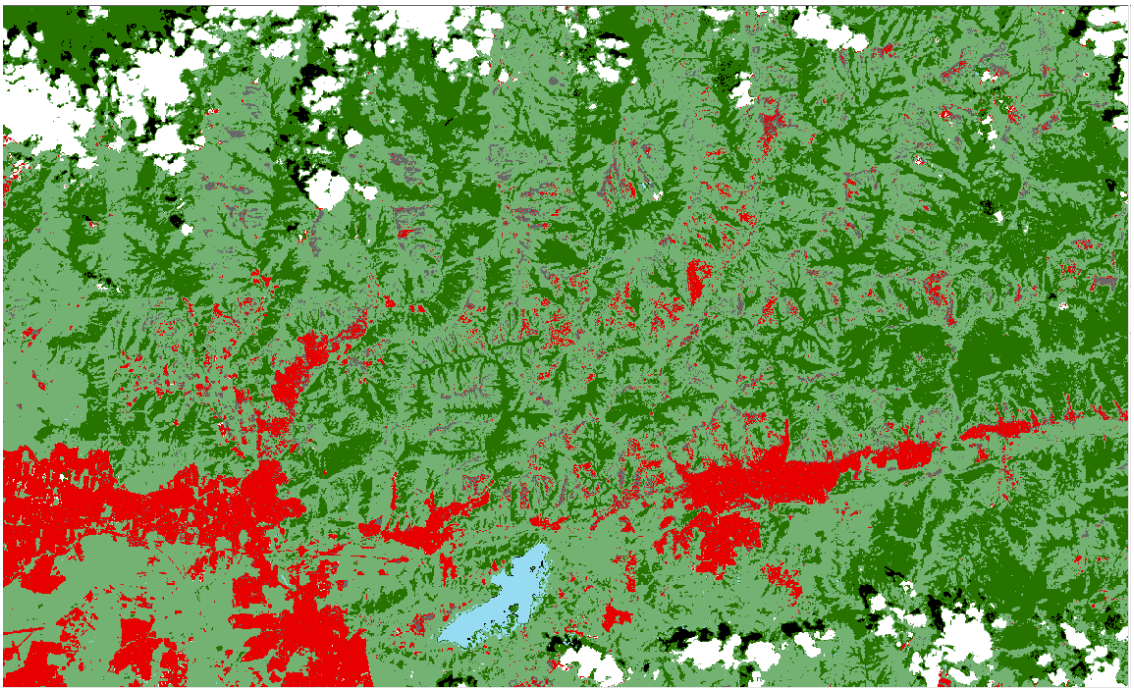
### 4.6. Explotación.

Para generar el mapa de cobertura de una escena, se construyó un Scrip en Python que codifica las reglas obtenida con el algoritmo RIPPER. Para cada pixel de la imagen, el modelo predice una clase (tipo de cobertura) y es coloreado siguiendo la nomenclatura que se muestra a continuación:

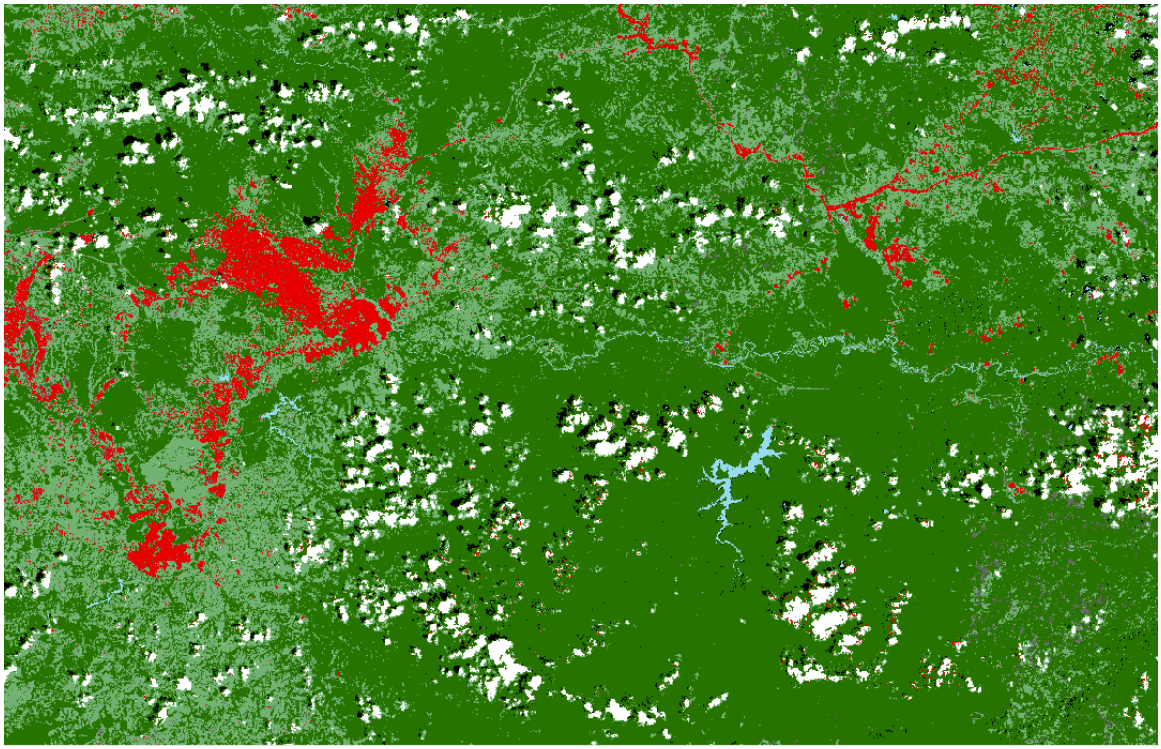


En lo que sigue, se muestran algunos de los mapas que fueron generados para las zonas de estudio consideradas (primero se muestra la imagen sin procesar, luego la imagen generada aplicando el procedimiento descrito):

a) Valles de Aragua



b) Valles del Tuy



## 5. Conclusiones

Mediante la técnica de minería de datos se logró, en un corto tiempo, generar mapas de cobertura de la tierra para una escena con gran precisión y rapidez. Es importante resaltar que la metodología presentada contribuye con una herramienta eficaz y confiable de gran apoyo al análisis del experto, con gran aplicabilidad para apoyar y promover la supervisión de las tierras agrícolas sin necesidad de utilizar ingentes recursos humanos y financieros. Como trabajo futuro se está considerando el análisis y generación de nuevos atributos que puedan resultar más informativos y que permitan distinguir mejor entre las clases que presentan confusión (por ejemplo, entre suelos desnudos y zonas urbanas). Además, los modelos son dependientes de la escena a partir de la cual se genera la muestra para la construcción del conjunto de datos. Se propone entonces realizar un muestreo más amplio que incluya otras escenas con el fin de obtener una mejor representación de las clases.

## Referencias


Adke, R., Johnson, J. (2017). *Predicting land use and atmospheric conditions from amazon rainforest satellite imagery*. the 2017 Stanford cs231n Poster Session will showcase projects in Convolutional Neural Networks for Visual Recognition.

Aggarwal, C. (2015). *Data Mining.* The Textbook. Springer

Aggarwal, C. (2918). *Neural Networks and Deep Learning*. A text book. Springer

Bogner, C., Seo, B., Rohner, D., Reineking, B. (2018). *Classification of rare land cover types: distinguishing annual and perennial crops in an agricultural catchment in south korea*. PloS one, 13(1). University of Maryland, USA.

Borràs, J., Delegido, J., Pezzola, A., Pereira, M., Morassi, G., Camps-Valls, G. (2017). *Land use classification from Sentinel-2 imagery*. Revista de Teledetección, 48, 55-66.

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. The MIT Press. Cambridge, Massachusetts

Jaramillo, L.V., Antunes, A.F. (2018). *Change detection in vegetation cover through interpretation of Landsat images by artificial neural networks (ANN). Case study: Ecuadorian Amazon Region*. Revista de Teledetección, 51, 33-46.

Kamilaris, A., Prenafeta-boldú, X. (2018). *A review of the use of convolutional neural networks in agricultura*. The journal of agricultural science, 1:11. Cambridge University Press.

Lu, D., Batistella, M., Li, Q., Moran, E., Hetrick, S. Da Costa, C., Vieira, L., Siqueira, S. (2012). *Land use/cover classification in the brazilian amazon using satellite images*. Pesq. Agropec. Bras., Brasília,.47(9):1185-1208.

Martínez-vega, M., Díaz, A., Nava, J., Gallardo, M., Echavarría, P. (2017). *Assessing land use-cover changes and modelling change scenarios in two mountain spanish national parks*. Environments, 4 (4):79.

Mucherino, A., Papajorgji, P., Pardalos, P. (2009). *Data minin in Agriculture*. Springer.

Patil, S., Gu, Y., Dias, F., Stieglitz, M., Turk, G. (2017*). Predicting the spectral information of future land cover using machine learning*. International Journal of Remote Sensing, 38(20):5592–5607.

Helber, P., Bischke, B., Dengel, A., Borth, D. (2017). *EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification*. IEEE International Geoscience and Remote Sensing Symposium (IGARSS-2018), Valencia, España.

Riegler-Nurscher, P., Prankl, J., Bauer, T. (2018). *Machine learning approach for pixel wise classification of residue and vegetation cover under field conditions*. Biosystems Engineering, 169:188-198. Elsevier.

Roiger, R. (2017). *Data Mining: A Tutorial-Based Primer*. CRC Press

Witten, I., Frank, E. (2017). *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufman Publishers. 4th Edition.

Yang, C., Rottensteiner, F., Heipke, C. (2018). *Classification of land cover and land use based on convolutional neural networks*. ISPRS annals of the hotogrammetry, remote sensing and spatial information sciences. Volume IV-3, Beijing, China.

Yirsaw, E., 1,2 , Wu , W., Shi, X., Temesgen, H., Bekele, B. (2017). *Land use/land cover change modeling and the prediction of subsequent changes in ecosystem service values in a coastal area of china, the su-xi-chang region*. Sustainability, 9(7).