UNIVERSIDAD POLITÉCNICA DE VALENCIA

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

MÁSTER UNIVERSITARIO EN INGENIERÍA DEL SOFTWARE,
MÉTODOS FORMALES Y SISTEMAS DE INFORMACIÓN

CURSO ACADÉMICO 2013-2014

ANÁLISIS DE MINERÍA DE DATOS PARA LA CLASIFICACIÓN DE IMÁGENES AÉREAS

TRABAJO FIN DE MÁSTER

PRESENTADO POR:
PABLO CRESPO PEREMARCH

TUTORA:
Prof. Dra. MARÍA JOSÉ RAMÍREZ QUINTANA

COTUTOR:
Prof. Dr. LUÍS ÁNGEL RUÍZ FERNÁNDEZ

VALENCIA, SEPTIEMBRE 2014

Resumen

Este proyecto trata de mejorar la clasificación de los usos del suelo con imágenes aéreas utilizando la minería de datos. Para ello se comparan diferentes técnicas de minería de datos en clasificaciones de zonas urbanas y periurbanas, utilizando la segmentación de imágenes para poder separar los diferentes objetos a clasificar. A partir de los resultados obtenidos, se estudia el problema de la agregación y selección de atributos para este tipo de imágenes, y se analiza el comportamiento de los modelos generados cuando se aplican a otras zonas geográficas con características similares.

Palabras Clave: *Minería de datos, Teledetección, Clasificación imágenes aéreas, LiDAR, Selección atributos, Test Wilcoxon, Weka.*

Abstract

This project tries to improve the land cover classification of aerial images using data mining. A comparison of some data mining techniques is done so as to classify urban and peri-urban areas, using the image segmentation in order to differentiate the objects that we want to classify. The next step is to analyse if adding and selecting attributes, the classification accuracy can be improved. Finally, we investigate whether the models that have been generated can classify a new geographical area with similar features.

Key Words: *Data mining, Remote sensing, Aerial image classification, LiDAR, Attribute selection, Wilcoxon test, Weka.*

Contenido

Índice de tablas

Índice de figuras

1. Introducción

La historia de la cartografía y la representación de mapas comenzó en los pueblos primitivos que tuvieron una cierta forma de cartografía rudimentaria, expresada muchas veces por lo que se podría llamar cartografía efímera: meros trazos momentáneos en la arena, en tierra húmeda u otros elementos. Esos trazos no eran más que una simple flecha indicadora entre dos puntos, pero son considerados como los primeros esbozos cartográficos.
Tras estos primeros trazos se pasó a dibujar en tablillas babilónicas, hasta llegar a su impresión en papel, o más actualmente a su representación digital, ofreciendo infinidad de posibilidades, así como poder compartir la información con otros usuarios. Los mapas constituyen hoy una fuente importantísima de información y una gran parte de la actividad humana está relacionada de una u otra forma con la cartografía.

La adquisición de los datos para poder realizar los mapas también ha sufrido importantes avances a lo largo de la historia. Si en la antigüedad únicamente era necesaria la presencia del ojo humano para dibujar esbozos, más tarde aparecieron los instrumentos que permitían medir ángulos y posteriormente también distancias, para de esta manera introducir la geometría en el cálculo y representación cartográfica.
Esas técnicas han llegado hasta nuestros días, pero la necesidad del ser humano de extenderse hasta el cielo y el espacio han hecho que apareciesen otras técnicas como el GNSS, la fotogrametría o la teledetección.

En este proyecto nos centramos en la teledetección, o adquisición de información de los objetos sin estar en contacto con ellos [2], que permite clasificar elementos del terreno a partir de imágenes, las cuales no sólo proporcionan información visible, sino también de otras zonas del espectro electromagnético (infrarrojo cercano, térmico, etc.), facilitando así la clasificación. Al no ser necesario estar en contacto con los objetos, la teledetección es de gran utilidad para el cartografiado de otros planetas o zonas no accesibles.
Como resultado se pueden obtener mapas de usos del suelo o detección de objetos presentes sobre el terreno.

La clasificación de imágenes por parte de un experto es una tarea larga y costosa. Estos problemas son bien conocidos en el ámbito de las bases de datos en los que el enorme volumen de información de que disponen las organizaciones ha motivado la aparición de la minería de datos. Estos datos guardan una valiosa información que puede ser utilizada para la toma de decisiones que a veces es difícil de extraer por los expertos humanos (pudiendo pasar por alto algunos detalles), e incluso por las técnicas estadísticas clásicas de análisis de datos.
Es por eso que la alternativa es la utilización de la minería de datos, una disciplina que propone el uso de técnicas de aprendizaje automático para la extracción de conocimiento en bases de datos.

A pesar de la utilidad de la minería de datos en la clasificación de imágenes, muchas veces ésta no se tiene en cuenta y se utilizan clasificadores basados en otras técnicas que vienen por defecto en los diferentes software relacionados con la teledetección.

La finalidad de este proyecto es poder realizar un estudio más completo de las posibilidades que nos ofrece la minería de datos para la clasificación de imágenes aéreas o satélite.
La zona de estudio se centrará en una zona urbana y periurbana costera con elementos geográficos típicos de la zona mediterránea.
Para dividir la imagen en diferentes objetos o grupo de píxeles, donde cada uno de ellos corresponderá con una instancia, se usará un análisis orientado a objetos de la imagen, más concretamente se aplicará una segmentación, donde la única información introducida es la propia imagen y un coeficiente de escala que según su valor irá unificando o dividiendo los diferentes segmentos.
Se trabajará con diferentes resoluciones espaciales (0.5, 1 y 2 metros) y en cada una de ellas se analizará qué clasificador realiza una mejor clasificación.

Por consiguiente, los objetivos concretos del proyecto son:

- Estudiar qué atributos son más significativos para la clasificación de imágenes, para ello:
    - Definiremos nuevos atributos que aporten una información extra a los atributos iniciales obtenidos por la segmentación, con la finalidad de mejorar la precisión de la clasificación.
    - Realizaremos una selección de atributos con objeto de reducir el ruido, teniendo en cuenta la correlación entre los mismos con el ánimo de disminuir el tiempo de cálculo y dar robustez a los modelos.
- Estudiar y comparar diferentes técnicas de aprendizaje automático para determinar, si es posible, qué clasificador realiza una mejor clasificación.
- Analizar la adaptabilidad del modelo aprendido, es decir, estudiar si el entrenamiento realizado en una zona A puede ser utilizado en una zona B de características similares - para trabajar en zonas parecidas con un modelo ya construido y que puede ser actualizado, ahorrándose el trabajo de selección de la muestra de entrenamiento.

Con este proyecto se pretende conseguir el clasificador más preciso en este tipo de entornos mediterráneos, mejorando así lo máximo posible los resultados de clasificación de la zona de estudio con unas clases básicas.

Las aplicaciones de toda esta información clasificada pueden ser muy variadas, siendo la fundamental la de actualizar la cartografía temática de media y alta resolución, pudiéndose emplear conjuntamente con otra información como la del Catastro para detectar automáticamente construcciones ilegales, o darle la posibilidad a un

ayuntamiento de hacer estudios de la superficie utilizada para cada uno de los usos, o simplemente disponer de la imagen con los diferentes usos del suelo.

Para poder llevar a cabo este trabajo se ha dividido la memoria en 9 capítulos, siendo el primero de ellos esta introducción.
En el Capítulo 2 se revisa el estado del arte presentando aquellas aproximaciones más relacionadas con nuestra propuesta como son los temas relacionados con datos LiDAR, la teledetección, y la minería de datos utilizada en la clasificación de imágenes.
En el tercer capítulo se define el plan de trabajo y la metodología seguida en este TFM.
En el Capítulo 4 se nombran algunos conceptos previos utilizados en el TFM y se realiza una descripción del marco de trabajo con el fin de que el lector pueda comprender todo el proyecto.
El Capítulo 5 describe todos los pasos realizados para el preprocesado de los datos cartográficos y la posterior generación de las vistas minables.
El sexto capítulo explica el proceso que se ha seguido para la generación de los diferentes modelos de minería de datos y la utilización del test estadístico para hacer un ranking con los diferentes clasificadores empleados.
El Capítulo 7 se muestran los resultados obtenidos tras realizar el estudio sobre los clasificadores y los atributos, y la utilización del test estadístico como se explicaba en el capítulo anterior.
El Capítulo 8 trata sobre la utilización de los modelos generados en el capítulo anterior para clasificar una nueva zona con características diferentes, y comparar esta clasificación con la efectuada por un software comercial como es ENVI; que utiliza los clasificadores kNN y SVM. En este capítulo se presentan la explicación del proceso y los resultados obtenidos.
En el Capítulo 9 se muestran las conclusiones a las que se ha llegado tras haber realizado el trabajo, los problemas que han aparecido durante su realización, se da respuesta a los objetivos planteados en la introducción y se mencionan unos posibles trabajos futuros relacionados con éste que podrían ser interesantes.

2. Antecedentes

Son muchos los trabajos que tienen en cuenta el estudio de las técnicas de minería de datos dentro del campo de la cartografía, ya sea para clasificar imágenes aéreas/satélite u otro tipo de información cartográfica, como es la clasificación de datos LiDAR (Light Detection And Ranging- que son sensores que miden la distancia entre el propio sensor y un objetivo situado sobre la superficie por medio de un pulso láser [4] desde aviones, obteniendo millones de puntos) tomando como atributos los datos resultantes de estudiar cada punto con su entorno más cercano [5].

En [6] y [7] se indica cómo los clasificadores convencionales estadísticos, que se han utilizado durante estas dos últimas décadas, no son apropiados para la clasificación del suelo a partir de la combinación de imágenes satélite con información geográfica, y únicamente se obtienen buenos resultados cuando los datos no tienen ruido o están normalizados. Sin embargo las técnicas de minería de datos mejoran la clasificación de las imágenes satélite [7], y sobre todo con la aparición de las imágenes hiperespectrales, que son imágenes con gran cantidad de bandas (pueden tener más de 200 bandas), ya que en este caso es necesario trabajar con gran cantidad de datos [8].

A las imágenes hiperespectrales también se le suma una nueva generación de sensores SAR (Synthetic-Aperture Radar), que a partir de una antena emite y recibe la información en la longitud de onda del radar, obteniendo una información diferente según el objeto con el que impacta. Estos sensores hacen que haya gran cantidad de imágenes en diferentes frecuencias, polarizaciones y con diferentes resoluciones, haciendo aumentar aún más la importancia de la minería de datos [8].

En [9] se efectúa una comparación mediante la clasificación de imágenes Landsat 5, que tienen una resolución espacial de 30 metros, de un método basado en técnicas estadísticas convencionales como es el de máxima verosimilitud con otros relacionadas con la minería de datos, como son las redes neuronales artificiales, Random Forest o máquinas de vectores soporte (SVM). Los resultados obtenidos corroboran claramente las afirmaciones expuestas en los párrafos anteriores, donde se indica que la minería de datos obtiene mejores resultados que los métodos estadísticos clásicos.

Una vez está demostrada la eficiencia e importancia de la minería de datos para la clasificación de imágenes aéreas, ciertos artículos como [10] y [11] realizan una comparación entre diferentes clasificadores para estudiar sus resultados, y cuál de ellos se comporta mejor. Las conclusiones obtenidas son que los árboles de decisión son los que mejoran los resultados, frente a las redes neuronales u otro tipo de clasificadores. En los artículos encontrados, la comparación entre clasificadores se hace directamente analizando la precisión obtenida, pero en ningún momento se aplica

un test estadístico, como se realiza en este TFM, para poder afirmar que las diferencias entre clasificadores son significativas y no son fruto de la aleatoriedad.

Aparte del estudio entre clasificadores también es muy importante saber cómo se va a obtener la información a partir de la cual se van a entrenar los clasificadores, es decir, los datos pueden extraerse de una imagen ya sea píxel a píxel o por objetos, que correspondería con una agrupación de píxeles; éste último método se conoce como orientado a objetos y dentro de éste los objetos se pueden extraer a partir de información vectorial (como puede ser con la información parcelaria del Catastro) o por segmentación, en el que se va haciendo una agrupación de los píxeles según sus valores radiométricos y un factor de escala dado.
En algunos artículos la clasificación se realiza a nivel de píxel [12], en otros se hace una comparación entre la anterior y una orientada a objetos [14], pero en la mayoría de los artículos consultados no se menciona si la clasificación se hace a nivel de píxel o es orientada a objetos, aunque en algunos casos como en [13] se puede deducir de qué tipo es por los atributos que se utilizan: a nivel de píxel se pueden extraer valores puntuales como la elevación, mientras que orientado a objetos se obtendrán valores medios de una región [14]. En [8] se menciona que las clasificaciones a nivel de píxel hacen que el resultado sea poco legible, ya que le dan un efecto *salt and pepper*, donde los píxeles de cada clase están muy intercambiados; por el contrario, las orientadas a objetos clasifican grupos de píxeles dando un resultado más acorde con la realidad.
Las clasificaciones orientadas a objetos también hacen que la clasificación sea mucho más rápida, ya que al no trabajar con cada uno de los píxeles sino con agrupaciones de éstos, el número de instancias disminuye en gran medida, haciendo posible trabajar con zonas más extensas o con más atributos.

Otro punto importante es conocer qué atributos se van a utilizar para poder aprender un modelo y posteriormente realizar una clasificación. En la mayoría de los casos no se citan los atributos utilizados, pero sin lugar a duda el que se utiliza más comúnmente aparte de los originales es el NDVI [13][19]. Este índice (Normalized Difference Vegetation Index) es usado para estimar la cantidad, calidad y desarrollo de la vegetación con base a la medición [26], por lo que es un índice que aparte de poder mostrar el estado de la vegetación, es capaz de discriminarla muy eficazmente de otro tipo de objetos.
En [9] también se menciona la importancia de la información textural, que representa la variación espacial del brillo de una imagen, y la utilización de este tipo de información resulta interesante en paisajes mediterráneos, debido a la gran variedad y fragmentación de patrones espaciales.
En [14] también se menciona la importancia de utilizar datos con información espacial como pueden ser datos GIS (Geographic Information Systems) como información extra para mejorar la clasificación. En este caso a lo que se refieren es a añadir atributos como pueden ser el área o la forma de un objeto, ya que esto podría diferenciar el agua de un río (con una forma alargada) del agua de un lago (con una forma más redondeada y ancha). No se hace mención en ninguno de los artículos, pero

una información extra que sería posible obtener con herramientas GIS y podría ser de gran utilidad sería el estudio de vecindad de cada objeto, es decir, añadir por ejemplo como información que el “Mar” puede colindar con otro objeto igual o con otro de clase “Playa”.

Otro aspecto que se tiene en cuenta en este TFM es la selección de atributos, la cual ha sido uno de los mayores problemas en el campo del análisis de patrones [15]. El número de atributos puede reducirse sin llegar a perder información, lo que conlleva una reducción del tiempo empleado para realizar la clasificación, pudiendo incluso mejorar la precisión de la misma. También puede ocurrir que cuando un clasificador es utilizado en un problema con una gran cantidad de variables, se pueda observar el efecto Hughes, que consiste en que la precisión de la clasificación disminuye cuando el número de atributos supera un límite dado, por lo que una reducción de los atributos soluciona también este problema [16].
El método más conocido para la reducción de atributos es el análisis de los componentes principales, pero éste no mantiene los atributos originales tras la reducción, sino que calcula una combinación de ellos y no es adecuado para la clasificación de imágenes.
Otra manera de realizar la selección es con el error Bayes. Éste es un buen indicador para la selección de atributos, pero es difícil obtener una expresión explícita y analítica, por lo que aparecen las medidas de separabilidad, que obtienen resultados similares y evitan este problema [17].
Dentro de las medidas de separabilidad existe la Distancia Jeffreys-Matusita, que corresponde con el área no común de las curvas de distribución de probabilidad de cada clase, siendo un buen criterio para la selección de atributos en aplicaciones de teledetección [18]. Un problema es que existen dificultades a la hora de realizar una evaluación cuando la información no se distribuye de manera Gaussiana [17].
Aparte de estos métodos también existe una selección de atributos rápida y que está basada en una estrategia de optimización Greedy [17]. Este método no sólo se basa en la información mutua de los diferentes atributos como hacen las medidas de separabilidad, sino que también tiene en cuenta la complementariedad existente entre ellos [17].
En este proyecto se utilizará el método Greedy junto a otros como Race y Genetic a través del software Weka. Estos métodos proporcionarán un cálculo más rápido, obteniendo como resultado una selección de atributos originales, y sin restricciones en cuanto a la distribución de la información. Otra diferencia es que la reducción de atributos se ha utilizado básicamente en clasificaciones de imágenes hiperespectrales, donde se habla de imágenes con más de 200 bandas, mientras que en este caso la selección se lleva a cabo en una imagen de 3 bandas que componen cerca de 70 atributos.

Un punto importante y que no ha sido abordado en ninguno de los artículos consultados es la incorporación de nuevos atributos que sean combinación de otros originales. Esta opción la exploraremos en este TFM ya que como se observa en la

Figura 2.1, es posible que la combinación de dos variables ayude a discriminar mejor una clase del resto (en este caso la clase representada en rojo), donde se ve claramente que los valores bajos de la variable representada en el eje Y se corresponden únicamente con esa clase. Por lo tanto, la incorporación de este nuevo atributo podría mejorar considerablemente los resultados de la clasificación.

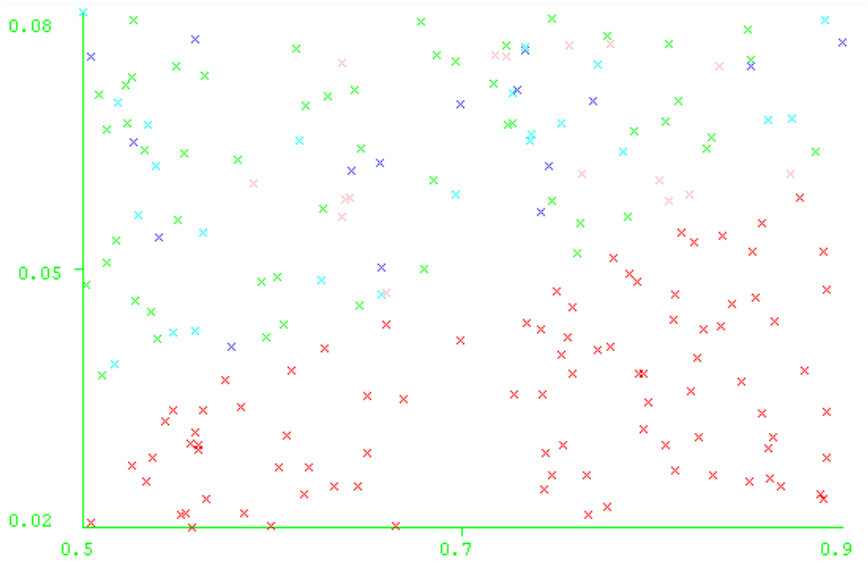


Figura 2.1. ***Distribución de las instancias con respecto a dos variables y representadas según la clase a la que pertenecen***

Ya se ha comentado anteriormente que sólo en [14] se hace una comparación entre una clasificación a nivel de píxel y otra orientada a objetos, pero en ningún caso menciona si cuando se hace la clasificación orientada a objetos las instancias se ponderan. Como se verá más adelante en este proyecto, ponderar las instancias cuando se hace una clasificación orientada a objetos es muy importante por el tamaño del objeto. Cuando se realiza una clasificación píxel a píxel todos tienen el mismo tamaño o peso, mientras que en la clasificación orientada a objetos cada objeto tendrá una superficie diferente y, por ejemplo, no se puede equiparar un objeto de $1m^2$ con uno de $1000m^2$, por lo que es muy importante hacer una ponderación con respecto al área.

También es importante hablar de las resoluciones espaciales utilizadas. A diferencia del proyecto que aquí se presenta, la mayoría de estudios realizados no trabajan con imágenes de alta resolución, sino que usan resoluciones espaciales comprendidas entre los 10 y 30 metros [13][14]; sólo en [12] trabajan con imágenes de 1,5 metros de resolución.
Cuando se trabaja con imágenes de alta resolución surge el problema contrario ya que a veces este tipo de imagen nos ofrece más detalles de los que realmente necesitamos, por lo que es interesante hacer un estudio para saber si algunas clases se clasifican mejor si la resolución disminuye, ya que por ejemplo las sombrillas presentes en la playa podrían homogeneizarse con la misma playa, mejorando así su clasificación.

Por otro lado, tanto en [13] como en [19] realizan, como se comentaba al principio de este apartado, una comparación entre diferentes clasificadores para un mismo conjunto de datos, salvo que en este caso utilizan el mismo software que se utilizará en este proyecto, conocido como Weka.
Ninguno de los artículos consultados entra en detalle de cómo se han obtenido los modelos estadísticos, salvo en [12] donde sí que se menciona que realizan 20 particiones para poder trabajar con 20 *datasets* para que los resultados obtenidos sean estadísticamente más robustos.

Como se puede observar hay muchos trabajos que relacionan la minería de datos con la clasificación de imágenes aéreas, incluso mostrando su importancia. En este proyecto se retomarán puntos presentes en los artículos consultados pero desde otra perspectiva. Se efectuará una comparación entre clasificadores y técnicas pero aplicando un test estadístico, con el fin de poder afirmar qué clasificador y técnica obtiene una mejor precisión en la clasificación, y que existe una diferencia significativa con el resto; también se añadirán atributos GIS como el área o la forma del objeto; y se incluirán otros aspectos no presentes en los artículos consultados como la ponderación de las instancias según su área, el añadir nuevos atributos combinación de los originales, hacer una selección de atributos más exhaustiva, trabajar con clases más generales, con imágenes de alta resolución (0,5 metros), y emplear el modelo obtenido para clasificar otra zona diferente pero con condiciones parecidas.

3. Plan de Trabajo y Metodología

En este apartado se indican las diferentes etapas de las que se compone el proyecto y qué orden siguen. El plan de trabajo está claramente diferenciado por tres etapas: en la primera se llevará a cabo el aprendizaje de diferentes modelos usando diversas técnicas de minería de datos. En la segunda etapa se intentarán mejorar los resultados obtenidos por los modelos, aplicando diversas técnicas de selección de atributos. Finalmente, en la última etapa, analizaremos cómo se comporta el modelo seleccionado aplicándolo a una imagen de test y compararemos los resultados obtenidos con los clasificadores kNN y SVM que proporciona una herramienta comercial como es ENVI.

Para desarrollar estas etapas nos hemos basado en una metodología estándar de proyectos de minería de datos. En este sentido, CRISP-DM (http://www.crisp-dm.org) (CRoss-Industry Standard Process for Data Mining) es una metodología propuesta por un consorcio de empresas (inicialmente bajo una subvención inicial de la Comisión Europea), incluyendo SPSS, NCR y DaimlerChrysler, para la gestión de proyectos de minería de datos. La difusión de este estándar ha sido altísima y, al ser independiente de la plataforma o herramienta, está siendo utilizado por cientos de organizaciones en todo el mundo como guía para implantar programas de minería de datos. En la Figura 3.1 se muestran las fases de dicho proceso para este TFM, que se diferencian ligeramente del diagrama original.

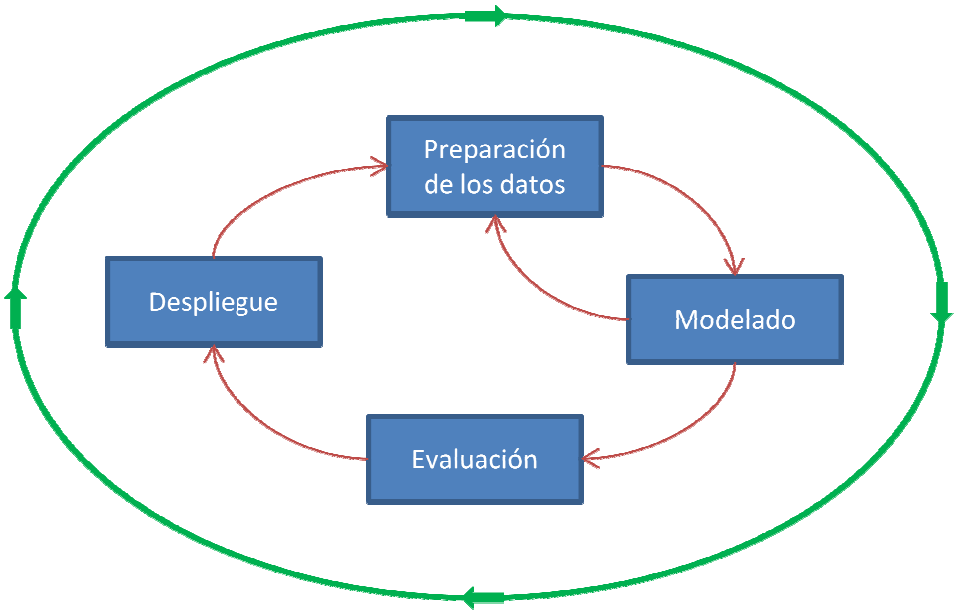


Figura 3.1. ***Diagrama de la metodología CRISP seguida en este proyecto***

La fase de preparación de datos incluye la limpieza, selección y transformación de los datos, con la construcción de un almacén de datos, en su caso, para utilizarse como repositorio de datos en las siguientes fases. El resultado final de esta fase es una o más vistas minables, que incluyen todos los datos relevantes para cada modelo a aprender.

La fase de modelado es la que realmente aplica las técnicas de modelado de minería de datos para obtener modelos y patrones a partir de los datos. Para ello, se pueden utilizar una o más herramientas de modelado y, en cada una de ellas, uno o más tipos de técnicas (árboles de decisión, regresión lineal, regresión no lineal, series ARIMA, redes neuronales, etc.).

La fase de evaluación analiza la fiabilidad de los modelos extraídos y selecciona los mejores. La fase de evaluación se suele basar en técnicas como la validación por partición (entrenamiento y comprobación) o en técnicas más sofisticadas (validación cruzada o bootstrapping).

Finalmente, la fase de despliegue aplica los modelos extraídos para la mejora de procesos y toma de decisiones. Incluye también un plan de monitorización y revisión de los modelos.

Como se ha comentado anteriormente, este proyecto se divide en tres etapas o fases que se describen a continuación y se identifican con su fase correspondiente en la metodología CRISP:

- Fase A: esta primera fase corresponde con la preparación de los datos. En ella se realiza un tratamiento de los datos LiDAR para que la información pueda ser incluida en la imagen para realizar la segmentación de la misma. La preparación de los datos finaliza generando las vistas minables que serán utilizadas en la siguiente fase.
- Fase B: la siguiente etapa es el modelado en el que mediante el software Weka se realizan diferentes pruebas con las diferentes vistas minables generadas previamente. Estas pruebas consisten en un modelado con los atributos originales, añadiendo nuevos atributos, y realizando una selección de atributos. Esta fase finaliza aplicando un test estadístico para concluir qué clasificador y prueba son los que obtienen un mejor resultado y poder trabajar con el modelo generado en la siguiente fase.
- Fase C: la última fase corresponde con las fases de evaluación y despliegue en la metodología CRISP. La evaluación se llevará a cabo aplicando el modelo generado en la fase anterior para clasificar una nueva zona, y verificando los resultados sobre un conjunto de test y la clasificación efectuada por un software comercial como es ENVI. Y el despliegue corresponde con el hecho de emplear el modelo generado en esa nueva zona de trabajo.

4. Algunos Conceptos Previos y Descripción del Marco de Trabajo

4.1. Descripción de herramientas software utilizadas en el trabajo

En este proyecto se utilizarán cinco softwares relacionados con la cartografía y la minería de datos, donde tres de ellos son libres y los otros dos comerciales. Estos softwares son Fusion, ENVI FX5, ArcGIS 10, Weka y Keel, que se explican a continuación más detenidamente.

Fusion [https://www.dataone.org/software-tools/fusion-lidar-software] es un software libre de visualización y con herramientas de análisis de datos LiDAR desarrollado por el Silviculture and Forest Models Team, que es un equipo de investigación del US Forest Service.
Esta herramienta permitirá el visionado de los datos LiDAR así como su tratamiento, identificando los puntos que corresponden con el terreno y la posterior conversión de los datos a un modelo para que pueda ser utilizado como una imagen y añadirse como una banda más en la imagen aérea.

ENVI FX5 [http://www.exelisvis.com/] está relacionado con el campo de la teledetección y es un software comercial de procesamiento y análisis avanzado de imágenes geoespaciales. Proporciona instrumentos avanzados para explorar, preparar, analizar y compartir la información extraída de todo tipo de imágenes.
En este proyecto muchos de los pasos que normalmente se pueden realizar con ENVI, como la selección de muestras de aprendizaje o la clasificación de imágenes, se realizarán con otros softwares como ArcGIS y Weka, respectivamente. Así pues, la utilización de esta herramienta en este proyecto se centrará en la modificación de las resoluciones espaciales de las imágenes y en la obtención de los diferentes objetos de la imagen por medio de la segmentación con sus respectivos atributos.

El tercer software que se utiliza es ArcGIS 10 [http://www.esri.es/es/productos/arcgis/] que está relacionado con el campo GIS, es comercial y es una plataforma de información que permite crear, analizar, almacenar y difundir datos, modelos, mapas y globos en 3D, poniéndolos a disposición de todos los usuarios según las necesidades de la organización.
En el trabajo se empleará para la selección de la muestra de aprendizaje a partir de los segmentos obtenidos con ENVI, para realizar cualquier operación espacial como

puede ser una intersección entre capas o recortes de áreas, o para la visualización de los usos del suelo tras la clasificación realizada con Weka.

Para las tareas propias de la minería de datos usaremos la suite Weka [http://www.cs.waikato.ac.nz/ml/weka/], consistente en un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos. Es un software que ha sido desarrollado en la Universidad de Waikato (Nueva Zelanda) bajo licencia GPL lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años.
En el presente proyecto se utilizará Weka para comparar los diferentes clasificadores, estudiar qué nuevos atributos se pueden añadir o realizar una selección de los mismos, así como clasificar con el modelo obtenido toda la imagen de la zona de estudio, que posteriormente puede ser visualizada con el software ArcGIS.

Por último, Keel (Knowledge Extraction based on Evolutionary Learning) [http://www.keel.es/] es un software de código abierto en JAVA similar a Weka y desarrollado por seis grupos de investigación españoles. Este software permite evaluar problemas de minería de datos incluyendo regresiones, clasificaciones, agrupamiento, etc. En este proyecto se utilizará para realizar el test estadístico de Wilcoxon, ya que posee unos módulos que permiten su cálculo.

4.2. Descripción de los datos

En todo trabajo relacionado con la minería de datos la calidad de los datos es muy importante, ya que con ellos aprenderá el modelo para poder posteriormente hacer las predicciones. Dependiendo de qué atributos se empleen en la descripción de la información, su relación entre ellos, así como su correlación, el resultado puede variar en cuanto a precisión.

En este proyecto los atributos utilizados para cada una de las instancias (segmentos de la imagen) son los que establece el software ENVI FX5 [27], los cuales se dividen en: atributos espectrales, de textura y espaciales.
Más adelante se observará cómo estos atributos pueden combinarse entre sí para incrementar la precisión, creando así nuevos atributos.
A continuación se menciona y se explica cada uno de los atributos utilizados por ENVI una vez efectuada la segmentación de la imagen.

- Atributos Espectrales:

Los atributos espectrales se obtienen para cada una de las bandas introducidas en la imagen. El valor del atributo se obtiene a partir de los valores que tienen los píxeles que forman parte de dicho objeto. Entre los atributos espectrales tenemos:

- Media espectral (AVG): corresponde con la media de los valores de los píxeles comprendidos dentro del objeto.
- Máximo espectral (MAX): es el valor máximo de los píxeles que se encuentran dentro del objeto.
- Mínimo espectral (MIN): es el valor mínimo de los píxeles comprendidos en el objeto.
- Desviación típica espectral (STD): corresponde con la desviación típica de los valores de los píxeles del objeto.

- Atributos de Textura:

Los atributos de textura también se calculan para cada una de las bandas de la imagen. En una textura se busca una región de la imagen que presenta propiedades locales constantes, lentamente variables o aproximadamente periódicas.
El cálculo se divide en dos procesos: en el primero se aplica a la imagen una ventana o *kernel* con un tamaño predefinido, donde los atributos que se explican a continuación son calculados para todos los píxeles en un vecindario *kernel* y el resultado se le asigna al píxel central. Una vecindario *kernel* es una matriz de números que es utilizada para calcular el valor del nuevo píxel en función de los valores de sus vecinos y el filtro empleado en la matriz. Esta matriz se desplaza por toda la imagen de entrada, calculando únicamente el valor del píxel central y asignando el nuevo valor a la imagen de salida. Normalmente el tamaño de las matrices es de 3x3, 5x5 o 7x7 píxeles, aumentando el efecto del filtrado cuanto mayor sea el número [21].
El siguiente proceso es parecido al apartado anterior, ya que los resultados obtenidos se promedian según los píxeles contenidos dentro del objeto. Los atributos de textura que ofrece ENVI son:

- Rango textura (TXRAN): corresponde con el intervalo de valores de texturas comprendidos dentro del objeto.
- Media textura (TXAVG): es el valor medio de las texturas de los píxeles que se encuentran dentro del objeto.
- Varianza textura (TXVAR): corresponde con la varianza extraída a partir de los valores de textura de píxeles comprendidos en el objeto.
- Entropía textura (TXENT): media de los valores de entropía de los píxeles que se encuentran dentro del objeto, la cual es un buen indicador de la distribución aleatoria de los valores de grises.

- Atributos Espaciales:

Por último, los atributos espaciales se calculan a partir del polígono definido por el propio objeto, por lo que no se necesita información de las diferentes bandas.

- Área (FX_AREA): corresponde con la superficie del objeto en las unidades en las que el mapa se encuentra georreferenciado, normalmente en metros.

- Longitud (FX_LENGTH): indica el perímetro del objeto en las unidades en las que se ha georreferenciado el mapa, normalmente en metros.
- Compacidad (FX_COMPACT): es una medida de forma del objeto que indica la compacidad del mismo, cuya fórmula es: $\sqrt{4 * \frac{Area}{\pi}} \Big/ Longitud$. Un círculo sería lo más compacto con un valor de $\frac{1}{\pi}$ y la de un cuadrado sería $\frac{1}{2\sqrt{\pi}}$.
- Convexidad (FX_CONVEX): los polígonos pueden ser convexos o cóncavos. Este atributo mide la convexidad del objeto, siendo el valor 1.0 cuando el objeto es convexo y sin agujeros, y siendo menor que 1.0 cuando es cóncavo. La fórmula para obtener el valor es: $\frac{Longitud\ pare\ convexa}{Longitud}$.
- Solidez (FX_SOLID): es una medida que compara el área del polígono con el área de una figura convexa envolviendo el polígono. La solidez para un polígono convexo sin agujeros es igual a 1.0, mientras que para un objeto cóncavo es menor que 1.0. La fórmula es: $\frac{Area}{Area\ convexa}$.
- Redondez (FX_ROUND): es una comparación entre el área del polígono con el cuadrado del mayor diámetro del polígono. El mayor diámetro corresponde con la longitud del mayor eje de un rectángulo que encuadra el polígono. La redondez para un círculo es igual a 1, y para un cuadrado es igual a $\frac{4}{\pi}$. La fórmula es: $\frac{4*Area}{\pi*Máximo\ Diámetros^2}$.
- Factor de forma (FX_FORMFAC): es una medida que compara el área de un polígono con el cuadrado del perímetro total. El valor para un círculo es igual a 1, y el valor para un cuadrado es igual a $\frac{\pi}{4}$. La fórmula es igual a: $\frac{4*\pi*Area}{Perímetro^2}$.
- Elongación (FX_ELONG): medida que calcula el ratio entre el eje mayor del polígono y el eje menor. Los ejes se obtienen del rectángulo que encuadra al polígono. La elongación para un cuadrado es igual a 1.0, y el valor para un rectángulo es mayor que 1.0. La fórmula es: $\frac{Eje\ mayor}{Eje\ menor}$.
- Forma rectangular (FX_RECT_FI): muestra lo bien que se puede ajustar el polígono a un rectángulo. El atributo compara el área del polígono con el área del rectángulo que envuelve al mismo. El valor para un rectángulo es igual a 1.0, mientras que para una forma no rectangular es menor que 1.0. La fórmula es: $\frac{Area}{Eje\ Mayor*Eje\ Menor}$.
- Dirección principal (FX_MAIN_DI): corresponde con el ángulo entre el eje mayor y el eje de las X en grados sexagesimales. El valor está comprendido entre los 0 y los 180 grados. 90 grados es una dirección Norte/Sur y un valor entre 0 y 180 es una dirección Este/Oeste.
- Mayor longitud (FX_MAJAXLN): es la longitud del eje mayor del rectángulo que encuadra al polígono. Las unidades de medida son en las que se encuentra georreferenciado el mapa, siendo normalmente en metros.

- Menor longitud (FX_MINAXLN): es la longitud del eje menor del rectángulo que encuadra al polígono. Las unidades de medida son en las que se encuentra georreferenciado el mapa, siendo normalmente en metros.
- Número de agujeros (FX_NUMHOLE): corresponde con el número de agujeros que tiene el polígono.
- Área del agujero/Área Sólida (FX_HOLESOL): es el ratio entre el área del polígono y el área del contorno del polígono. Si el objeto no tiene agujeros su valor será igual a 1.0. La fórmula es: $\frac{Área}{Área\ contorno\ exterior}$.

4.3. Descripción de las técnicas de minería de datos utilizadas

Los modelos pueden ser de dos tipos: predictivos o descriptivos. Los primeros pretenden estimar valores futuros o desconocidos de variables de interés, conocidas como variables objetivo o dependientes, utilizando otros campos de la base de datos, conocidos como variables independientes o predictivas.
Sin embargo, los modelos descriptivos identifican patrones que agrupan los datos, sirviendo para explorar propiedades de los datos analizados, no para predecir nuevos datos.

Dentro de los modelos predictivos encontramos la clasificación y la regresión, mientras que el agrupamiento (*clustering*), las reglas de asociación, las reglas de asociación secuenciales y las correlaciones son tareas descriptivas.

La clasificación es posiblemente la tarea más utilizada y consiste en que cada instancia corresponde a una clase. Este atributo puede tomar diferentes valores discretos, correspondiendo cada uno de ellos a una clase, mientras que el resto de atributos de la instancia son utilizados para predecir la clase.
El objetivo es asignar la clase correcta al mayor número posible de nuevas instancias.

La regresión en lugar de asignar clases asigna a cada registro un valor real. El objetivo es minimizar el error (normalmente se utiliza el error medio cuadrático) entre el valor predicho y el real.

Dentro de las tareas descriptivas el agrupamiento es la tarea por excelencia y consiste en obtener grupos “naturales” a partir de los datos. A diferencia de la clasificación, en lugar de analizar datos ya etiquetados, lo que hace es analizar los datos para etiquetarlos. En los grupos generados se busca que los objetos pertenecientes a un mismo grupo sean muy similares entre sí y muy diferentes de los objetos de los otros grupos.

Al ser la minería de datos un campo muy interdisciplinar existen diferentes paradigmas detrás de las técnicas utilizadas: técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéricos, aprendizaje bayesiano, programación lógica inductiva y varios tipos de métodos basados en núcleos, entre otros.

Muchos conceptos estadísticos son la base de muchas técnicas de minería de datos, entre estas técnicas podemos encontrar la regresión lineal o no lineal.
Las técnicas estadísticas no son sólo útiles para la regresión, sino que se utilizan también para discriminación (clasificación o agrupamiento).

Entre los métodos basados en núcleo las máquinas de vectores soporte son su ejemplo más representativo, en el que se busca un discriminante lineal que maximice la distancia a los ejemplos fronterizos de las distintas clases. Esta técnica se explicará más detalladamente más adelante, ya que en este proyecto se utiliza el método SMO que es una máquina de vector soporte.

La siguiente técnica corresponde con los árboles de decisión que son una serie de decisiones o condiciones organizadas en forma jerárquica, con forma de árbol. Son muy utilizados para encontrar estructuras en espacios de alta dimensionalidad y en problemas donde se mezclan datos numéricos y categóricos. Pueden emplearse tanto para la clasificación, agrupamiento y regresión. Esta técnica también es empleada en el TFM, por lo que se desarrollará más detenidamente más adelante.

El aprendizaje basado en instancias o casos almacena las instancias en memoria, por lo que cuando llega un nuevo registro para clasificar, se intenta relacionar éste con el resto de registros almacenados (de los cuales su clase o valor se conocen) buscando los que más se parecen, con el objetivo de utilizar estas instancias similares para estimar los valores a obtener de la nueva instancia.
Todo el proceso de aprendizaje basado en instancias se realiza cuando llega una instancia para clasificar y no al procesar el conjunto de entrenamiento. Se conoce pues como un método retardado o perezoso, al retrasar el trabajo real tanto como sea posible.
Una variante conocida es el método de los “k vecinos más próximos” que es utilizado en este trabajo y se explicará más adelante.

Como en este TFM se utilizará el método Random Forest, también es necesario hablar de la combinación de modelos. Con el objetivo de mejorar la precisión de las predicciones, surge un interés en la definición de métodos que combinan hipótesis. Estos métodos construyen un conjunto de hipótesis (*ensemble*), y combinan las predicciones del conjunto de alguna manera (normalmente por votación) para clasificar ejemplos o para hacer regresión sobre ejemplos. A estas técnicas de combinación de hipótesis se les conoce como métodos de ensamblaje de modelos (entre otros nombres). La combinación de modelos se ha desarrollado principalmente

para modelos predictivos, aunque ciertas ideas podrían extenderse a modelos descriptivos. Este punto se desarrollará más adelante centrándonos en el método Random Forest [1].

En este proyecto se utilizarán 4 técnicas predictivas presentes en Weka para analizar la precisión con la que clasifican las imágenes aéreas - y que corresponden con máquinas de vectores soporte, basados en instancias, árboles de decisión, y ensambles. Dentro de máquinas de vectores soporte se utilizará SMO (Sequential Minimal Optimization); como algoritmo basado en instancias se utilizará el vecino más próximo que en Weka se conoce como IBk (Instance-Based); dentro de los árboles de decisión se utilizarán el J48; y el Random Forest como ensamble.

Las máquinas de vectores soporte (SVM) se basan en un clasificador lineal muy sencillo, precedido de una transformación de espacio para darle potencia expresiva. El clasificador lineal empleado obtiene la línea (para 2 dimensiones o el hiperplano para un mayor número de dimensiones) que separe limpiamente las dos clases maximizando la distancia a la frontera de los ejemplos más próximos a la misma. La Figura 4.1 muestra 2 posibles particiones del espacio, de los cuales el de la derecha es preferible ya que la distancia a la frontera de los tres puntos próximos a la misma es mayor que el clasificador de la izquierda.

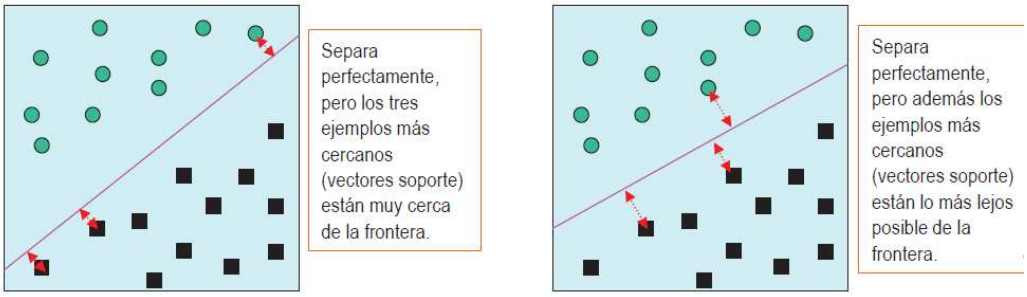


Figura 4.1. ***Ejemplos de un discriminante (clasificador) basado en vectores soporte, donde el de la derecha es mejor que el de la izquierda al encontrarse los ejemplos más cercanos lo más lejos posible de la frontera.***

El algoritmo es muy eficiente incluso para cientos de dimensiones, ya que el separador lineal puede tener únicamente en cuenta los puntos más próximos y descartar los más lejanos a la frontera.
Cuando los datos no son separables linealmente se aplica una función núcleo *kernel*. El aprendizaje de separadores no lineales con SVM se consigue mediante una transformación no lineal del espacio de atributos de entrada en un espacio de características de dimensionalidad mucho mayor y donde sí es posible separar linealmente los ejemplos. Las funciones núcleo *kernel* calculan el producto escalar de dos vectores en el espacio de características sin necesidad de calcular explícitamente las transformaciones de los ejemplos de aprendizaje [1].

El método SMO de Weka implementa el algoritmo secuencial de optimización mínima de John C. Platt para entrenar un clasificador de vector soporte usando *kernels* polinomiales escalados. Transforma la salida de SVM en probabilidades, aplicando una función sigmoide. Reemplaza todos los valores vacíos, transforma los atributos nominales en binarios, y normaliza todos los atributos numéricos [28].

El siguiente algoritmo que se utiliza es el IBk de Weka, que corresponde con el vecino más próximo (kNN). Este algoritmo no genera ningún modelo, por eso se conoce como un método perezoso (lazy en inglés). Lo que hace el algoritmo es buscar los k casos más cercanos. El valor de k se suele determinar heurísticamente, aunque $k = \sqrt{n}$, donde n es el número de ejemplos, es una opción con base teórica. Para asignar una clase a un dato puede que ocurran diferentes situaciones: si todos los k datos más cercanos son de la misma clase, entonces el nuevo caso se clasifica en esa clase; si no son todos de la misma clase, entonces se calcula la distancia media por clase o se asigna a la clase con más elementos [1]. En la Figura 4.2 se muestra cómo se buscan los k elementos más cercanos al que se quiere clasificar y cómo se genera una partición que marca las regiones pertenecientes a cada clase.

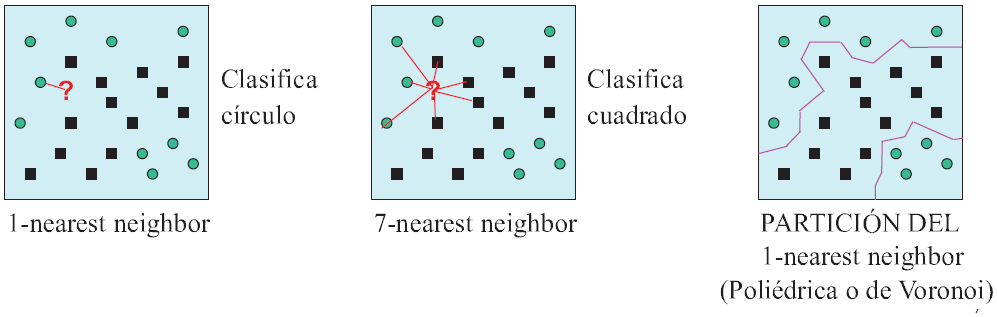


Figura 4.2. ***Ejemplo del método de los vecinos más próximos, siendo k=1 en el de la izquierda y k=7 en el de en medio, y en la figura de la izquierda se observa la partición realizada para k=1***

Los árboles de decisión son un modelo de predicción utilizado en el ámbito de la inteligencia artificial, donde dada una base de datos se construyen diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de problemas [26].
Tienen la ventaja de que son muy fáciles de entender y visualizar el resultado, y robustos al ruido, existiendo algoritmos para podar hojas poco significativas. Sin embargo sus desventajas son que suele ser muy voraz, y si el criterio de partición no está bien elegido, entonces las particiones suelen ser muy ad-hoc y generalizan poco [1].

Dentro de los árboles de decisión nos fijamos en el J48 de Weka que corresponde con el algoritmo C4.5. Dicho algoritmo fue desarrollado por Ross Quinlan y es una extensión del ID3.
El algoritmo crea un nodo raíz con todos los ejemplos; si todos los elementos son de la misma clase, el subárbol se cierra y la solución ya se da como encontrada; si esto no es así, entonces se elige una condición de partición siguiendo un criterio de partición; el problema queda subdividido en dos subárboles (los que cumplen una condición y los que no) y se vuelve a 2 para cada uno de los dos subárboles [1]. En la Figura 4.3 se muestra la partición cuadricular realizada para un ejemplo con dos variables y cómo ésta se traduce a una representación con forma de árbol.

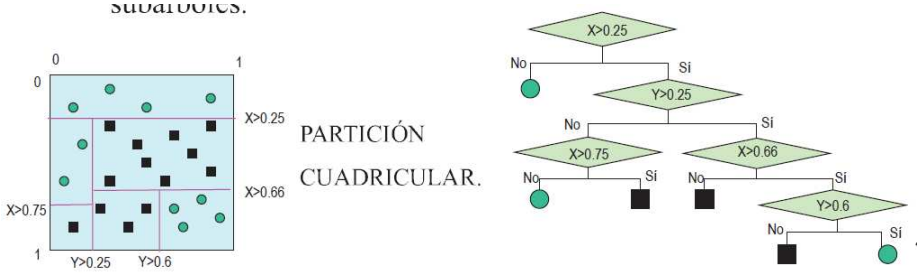


Figura 4.3. ***Ejemplo de un árbol de decisión con una partición cuadricular y su representación en forma de árbol***

Otro algoritmo utilizado en el proyecto dentro de los métodos de ensamblaje de modelos es Random Forest. Éste fue desarrollado por Leo Breiman, se basa en el desarrollo de muchos árboles de clasificación. Para clasificar un nuevo objeto desde un vector de entrada, se pone dicho vector bajo cada uno de los árboles del bosque. Cada árbol genera una clasificación, en términos coloquiales diríamos que cada árbol vota por una clase. El bosque escoge la clasificación teniendo en cuenta el árbol más votado sobre todos los del bosque [29].

4.4. Medidas de evaluación de los modelos

Para poder entrenar y probar un modelo se deben partir los datos en dos conjuntos: el conjunto de entrenamiento o aprendizaje y el conjunto de test o evaluación. Esta separación es necesaria para garantizar que la validación de la precisión del modelo es una medida independiente. Si no se utilizan conjuntos diferentes de entrenamiento y prueba, la precisión del modelo será sobreestimada, es decir, tendremos estimaciones muy optimistas.
El método empleado en este trabajo es la validación cruzada con n pliegues (en este caso n=10), que es el método que se usa normalmente. En este método los datos se dividen aleatoriamente en n grupos, reservándose un grupo para el conjunto de prueba y los otros n-1 restantes para construir un modelo y predecir el resultado de los datos del grupo reservado. Este proceso se repite tantas veces como pliegues se hayan

indicado, utilizando cada vez un grupo diferente para la prueba y entrenar el nuevo modelo.
Finalmente se construye un modelo con todos los datos y se obtienen sus ratios de error y precisión promediando los n ratios de error disponibles [1].

Como medida de evaluación de modelos se puede utilizar la precisión o fiabilidad global, o el índice Kappa. En este TFM se presentan los resultados para el índice Kappa, que será el utilizado para analizar qué clasificador lleva a cabo una mejor clasificación, ya que este índice también tiene en cuenta los valores que no se encuentran sobre la diagonal de la matriz de confusión, es decir, los valores que han sido clasificados erróneamente, mientras que la precisión únicamente se fija en la diagonal, que son las instancias bien clasificadas.
En la bibliografía consultada existen dos fórmulas diferentes para calcular el índice Kappa, en este proyecto se usará la siguiente fórmula, ya que es la empleada por Weka:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

donde Pr(a) es el acuerdo observado entre los evaluadores (precisión = $\frac{\sum diagonal}{total}$), y Pr(e) es la probabilidad hipotética de la posibilidad de acuerdo ($\sum_1^{nº\ clases}(\frac{\sum fila}{total} * \frac{\sum columna}{total})$). Cuanto más próximo sea el valor de κ a 1, mejor será el resultado.

5. Preprocesado de los Datos

Tal y como puede observarse en la Figura 3.1 una de las primeras etapas en el desarrollo de una aplicación de minería de datos es el adecuado procesamiento de los datos con el objeto de generar una vista minable de calidad a partir de la cual entrenar los modelos. Este preprocesado permite eliminar datos no válidos (erróneos), transformar el formato, rango, etc. de algunos datos para que puedan ser utilizados por los algoritmos de aprendizaje, etc.

Cuando se trabaja con datos espaciales el primer paso es delimitar el área de trabajo. En la Figura 5.1 se muestra la localización de la zona de estudio y en la Figura 5.2 se ha delimitado la misma. En este trabajo se pretende obtener un modelo de clasificación para una zona urbana y periurbana de tipo mediterráneo, y que posteriormente se pueda utilizar para clasificar otras zonas con unas características similares. La zona escogida es la del Puerto de Sagunto junto con una parte del término de Canet d'en Berenguer, ya que es una ciudad de un tamaño considerable, con diferentes tipos de edificaciones, presencia de cultivos a sus alrededores, industria, zonas boscosas, y es una zona costera; por lo que también posee playa y mar. Al ser una ciudad con unos usos del suelo tan variados, se pretende que se pueda utilizar el modelo de clasificación en otra zona, sin importar si ésta es costera o de interior, ya que el modelo obtenido con el aprendizaje ha sido creado en una zona con los diferentes usos posibles.

Puerto de Sagunto forma parte del término municipal de Sagunto, situado en la comarca del Camp de Morvedre, a unos 25 km al norte de la ciudad de Valencia. Canet d'en Berenguer también es una localidad costera que forma parte de la misma comarca y su término municipal se encuentra rodeado por el de Sagunto.

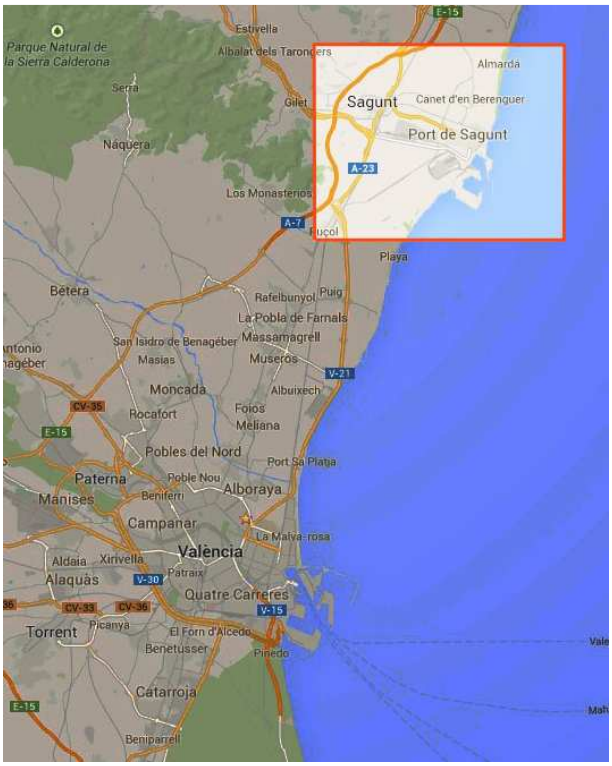


Figura 5.1. ***Localización de la zona de estudio A***

Figura 5.2. ***Delimitación de la zona de trabajo A***

Los datos de los que se dispone y con los que se trabaja son datos LiDAR y ortofotografías de alta resolución con una resolución espacial de 0,5 metros.
Los datos LiDAR corresponden con el vuelo que realizó el Instituto Geográfico Nacional (IGN) en la sección dedicada al Plan Nacional de Ortofotografía Aérea (PNOA) en el año 2009 y con una resolución de 0,5 ptos/m^2, y han sido facilitados por el Departamento de Ingeniería Cartográfica, Geodesia y Fotogrametría de la Universidad Politécnica de Valencia.
Es muy importante que los datos LiDAR y las ortofotografías fueran tomados con una diferencia temporal pequeña, ya que si no podría haber algunas diferencias como que un edificio aparezca en el modelo de alturas y no en la imagen, o viceversa.

Una vez se conoce la zona de estudio se deben recortar tanto los datos LiDAR como las ortofotografías según los límites, ya que trabajar con más datos de los necesarios supondría un mayor uso de memoria y de tiempo. Para recortar las imágenes se ha utilizado el software ArcGIS, mientras que la selección de datos LiDAR se ha ejecutado con el software FUSION.

- **Tratamiento de datos LiDAR:**

En la Figura 5.3 se presenta un diagrama de flujo describiendo todos los procesos que se explican a continuación y que tienen que ver con el preprocesado de los datos.
La finalidad de trabajar con datos LiDAR es obtener dos modelos: uno que corresponde con el terreno y con cualquier elemento presente sobre él, como pueden ser árboles o edificaciones, y es conocido como Modelo Digital de Superficie (MDS); y un segundo que corresponde únicamente con el terreno, conocido como Modelo Digital del Terreno (MDT). Realizando la diferencia entre estos dos modelos lo que se obtiene es un modelo normalizado, que contiene la altura de todos los objetos presentes, pudiendo así añadir esta información como una banda más de la ortofotografía; de este modo se puede diferenciar entre objetos con los mismo valores radiométricos pero con una altura diferente, como podría ser un árbol de un arbusto.

Antes de haber sido tratados, los datos LiDAR son un conjunto de puntos láser tomados desde un avión y que han podido intersectar con cualquier tipo de objeto presente, por lo que es muy probable que se hayan capturado puntos erróneos, que no hacen más que añadir ruido. Estos puntos pueden ser por ejemplo, pájaros o cualquier otro elemento que no permita capturar la superficie. Por eso el primer paso será aplicar un filtro para eliminar estos puntos, basándose en los datos obtenidos en su entorno y detectando si estas diferencias son muy grandes.

Habiendo eliminado los puntos que no corresponden con la superficie, el siguiente paso será separar con una serie de algoritmos facilitados en FUSION los puntos que corresponden con el terreno, para posteriormente obtener con estos datos el MDT. Para filtrar los puntos del terreno estos algoritmos se basan en los rebotes de cada uno de los pulsos láser enviados desde el avión. El último rebote suele corresponder con el

terreno, y si existen otros rebotes suelen pertenecer a objetos como pueden ser las ramas de los árboles.

Con los puntos correspondientes con el terreno ya identificados, se genera el MDT mediante una interpolación de estos puntos, dando una imagen como salida, en la que cada uno de sus píxeles tiene como valor radiométrico la altura del terreno en ese punto.

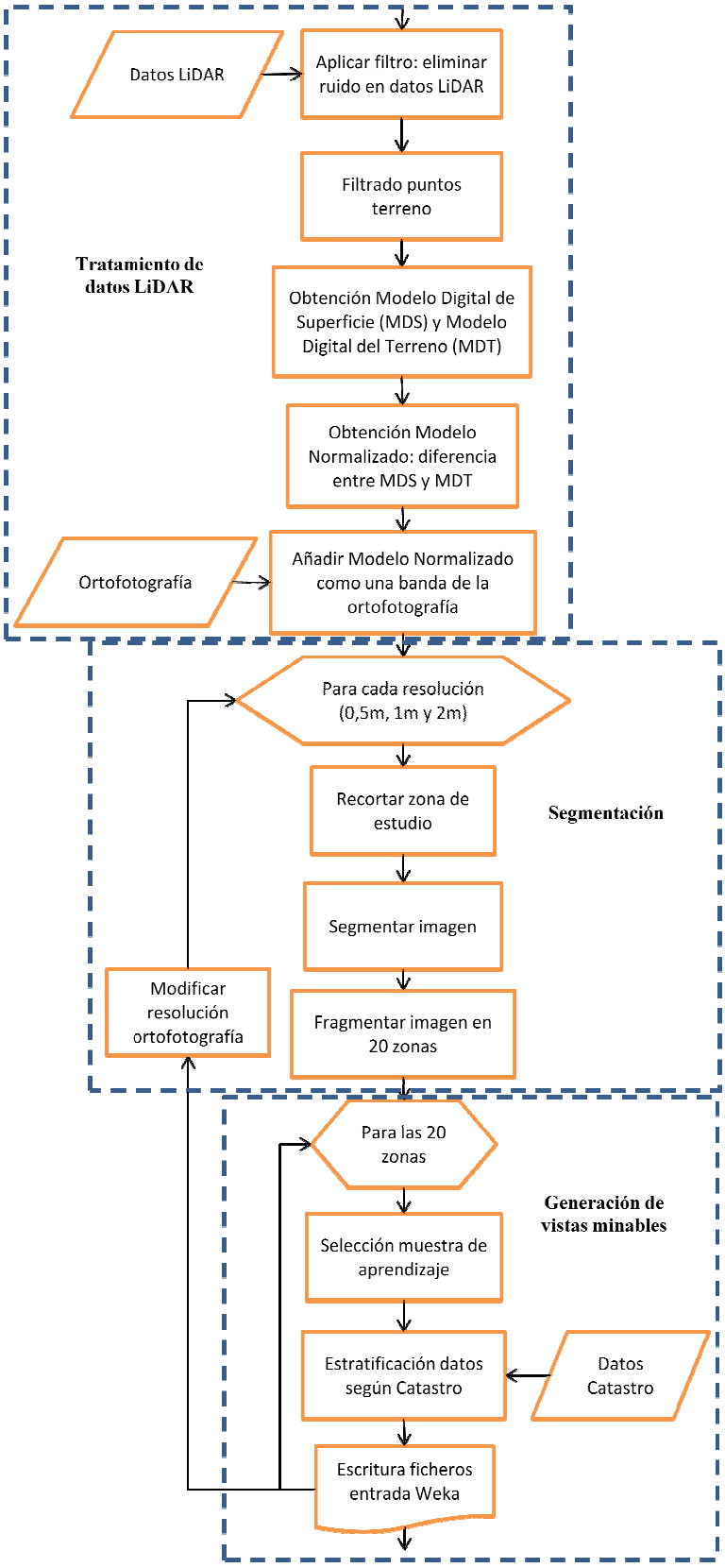


Figura 5.3. ***Detalle de las diferentes etapas realizadas durante la fase de preprocesado de los datos***

Un punto importante a la hora de obtener tanto el MDT como el MDS es saber qué resolución espacial es la más idónea para los datos LiDAR de los que se dispone, ya que puede pasar que si se escoge una resolución muy baja, se le esté exigiendo más resolución de la que realmente tienen los datos; y si la resolución es muy alta, se esté perdiendo información. Esta resolución varía según la densidad de los datos LiDAR.

En este proyecto se han probado para la generación del MDT y del MDS las resoluciones de 5, 4, 3, 2 y 1 metros, escogiendo finalmente una resolución de 2 metros. En la Figura 5.4 se muestra el MDS de la zona del Castillo de Sagunto para la resolución escogida de 2 metros. Para poder escoger esta resolución se han obtenido diferentes modelos a las diferentes resoluciones mencionadas y se ha analizado si los modelos tenían muchas variaciones altimétricas, por lo que la resolución estaba siendo más alta de lo que debería; o si el modelo estaba muy suavizado, por lo que la resolución estaba siendo más baja de lo que debería.

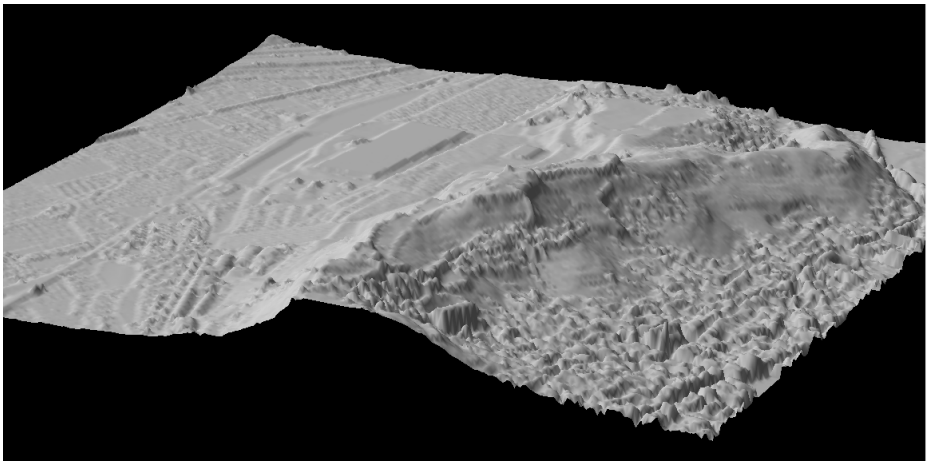


Figura 5.4. ***Modelo Digital de Superficie de la zona del Castillo de Sagunto generado a partir de los datos LiDAR y con una resolución de 2 metros***

Una vez determinada la resolución más adecuada al problema ya se puede aplicar el mismo proceso al resto de zonas que forman parte de la zona de estudio.
Como los datos LiDAR del área de trabajo vienen en diferentes ficheros, es necesario unificarlos todos antes de efectuar el filtrado y la obtención de los modelos.

En estos momentos ya se dispone del MDT y MDS con una resolución de 2 metros y en formato imagen. Para poder utilizar la altura de los objetos, extraída de los dos modelos, como una banda más de la ortofografía, es necesario que tanto la ortofotografía como los modelos tengan los mismos límites y la misma resolución.
Con ENVI se puede modificar la resolución de los modelos para transformarla de 2 metros a 0,5 metros, por lo que un píxel original de los modelos corresponderá con 16 nuevos.

Al ya disponer de la ortofotografía, del MDS y del MDT con los mismo límites y la misma resolución espacial, se puede obtener el modelo de alturas que será la diferencia entre el MDS y el MDT.
En la Figura 5.5 se observa la altura de los elementos presentes en el terreno, representándose en verde lo elementos más bajos o cuando el punto coincide con el terreno y en rojo los objetos más elevados.

Figura 5.5. ***Modelo normalizado con las alturas de los objetos de la zona de Puerto de Sagunto y Canet d'en Berenguer***

El siguiente paso será añadir este modelo de alturas como una banda más de las ortofotografías, siendo el valor radiométrico de cada píxel, la altura del elemento a partir del terreno.

- **Segmentación de la imagen:**

Cuando ya se dispone de la imagen con la que se quiere trabajar es el momento de la segmentación de la misma, que consiste en ir agrupando píxeles que presentan unos valores radiométricos similares. De esta manera se pasa de trabajar con píxeles a trabajar con segmentos u objetos, los cuales ya no presentarán valores radiométricos puntuales sino datos estadísticos como la media, valor máximo y mínimo, etc.
Para realizar la segmentación de la imagen se utilizará el software ENVI FX5, que posee las herramientas con los algoritmos introducidos para poder efectuar este proceso.

Para que el software conozca a qué nivel debe realizar la segmentación, es necesario introducirle un factor de escala. Cuanto mayor sea este factor, menor número de segmentos generará y los píxeles estarán más agrupados. Por lo tanto, el objetivo es

encontrar un factor de escala que realice una mejor segmentación pero con un menor número de segmentos, ya que así se utilizará menos memoria.

El principal problema en este proyecto apareció a la hora de segmentar la línea de costa, ya que con una buena segmentación del resto de elementos, la línea de costa agrupaba zonas de playa con zonas de mar; esto supuso introducir un factor de escala más bajo, generando así más objetos y aumentando el tiempo de procesamiento. Finalmente los valores introducidos tanto en el factor de escala como en la agrupación (merge) para las diferentes resoluciones espaciales fueron los siguientes:

- **0,5m:** Factor de escala 20, Merge 55.
- **1m:** Factor de escala 20, Merge 55.
- **2m:** Factor de escala 20, Merge 55.

Tras realizar la segmentación de la imagen el siguiente paso es seleccionar qué atributos tendrá cada uno de estos segmentos. En este caso se introdujeron todos los atributos que ofrecía ENVI FX5, ya que una de los objetivos de este proyecto es combinar atributos entre sí para generar otros nuevos y realizar una selección de atributos, por lo que de entrada se necesita trabajar con todos.

Después de efectuar estos pasos ya se dispone de un fichero en formato shapefile de tipo polígono en el que se muestran los diferentes segmentos generados y cada uno de ellos lleva asociado sus propios valores correspondientes con cada uno de los atributos seleccionados.

El siguiente paso es dividir la zona de trabajo en 20 particiones iguales, cada una de las cuales se tomará como un *dataset*. Todo ello nos va a permitir evaluar los modelos mediante validación cruzada y efectuar el correspondiente test estadístico para analizar la significancia de los resultados obtenidos.

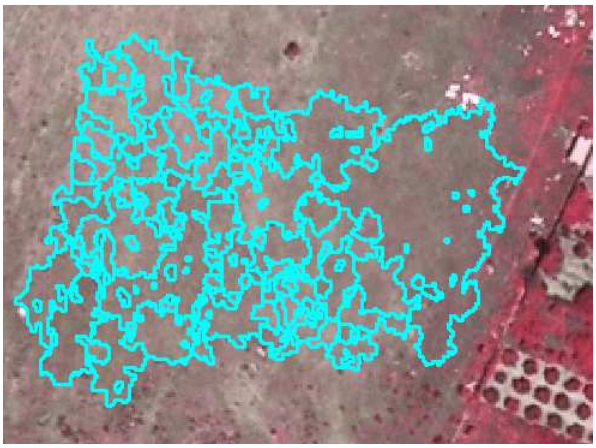
- **Generación de vistas minables:**

Al ya disponer de los diferentes ficheros shapefile con cada uno de los segmentos, ahora se puede trabajar desde ArcMap con la ortofotografía de fondo e ir etiquetando los polígonos almacenando en una tabla con la clase de cada uno. En la Figura 5.6 se muestra cómo se realiza la selección de los objetos con ArcMap y en la Figura 5.7 se observa la tabla con las características de los objetos seleccionado y donde el valor del atributo CLASE debe ser modificado por el de la clase correspondiente. Este proceso se repetirá para las 20 particiones realizadas en el apartado anterior.
Las clases que se van a considerar son: Edificación, Vía, Suelo Desnudo, Vegetación, Playa y Agua.
La clase Edificación corresponde con cualquier tipo de construcción ya sea un edificio o un chalet.
Como Vía se etiquetan las carreteras, vías férreas, calles, parkings asfaltados y aceras.

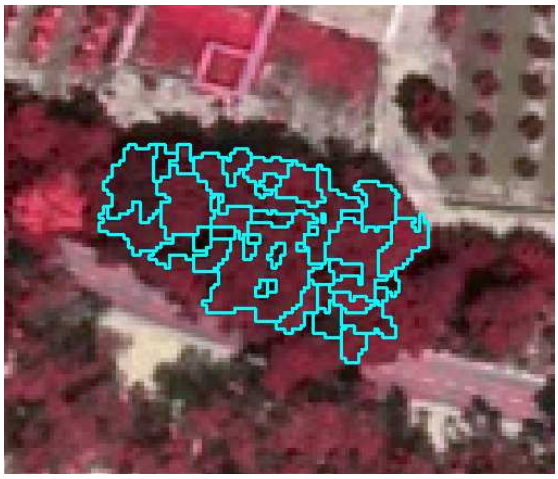
Dentro de la clase Suelo Desnudo se encuentran aquellas superficies que no contienen vegetación, no están asfaltadas y no son playa, como pueden ser descampados, zonas fluviales sin vegetación ni agua o zonas de cultivo sin vegetación.
La clase Vegetación corresponde con cualquier tipo de vegetación como pueden ser palmeras, pinos, arbustos, cultivos, etc.
Como Playa se etiqueta todo segmento que corresponda con una playa de arenas.
Por último, la clase Agua contiene tanto los objetos de agua de mar, como piscinas o balsas.

(a)

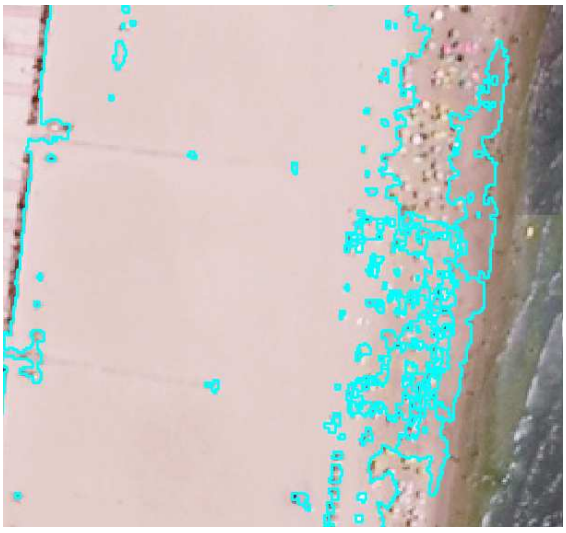
(b)



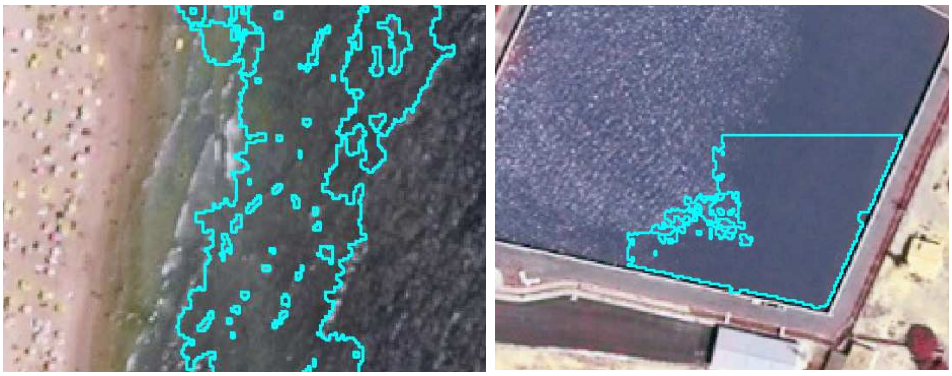
(c)



(d)

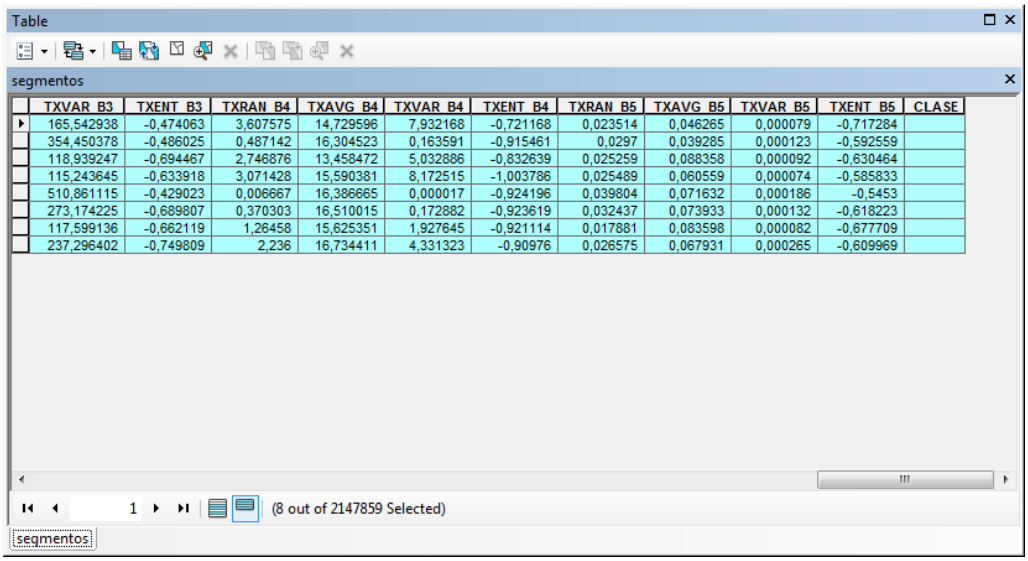


(e)



(f)

Figura 5.6. ***Ejemplo de selección de segmentos de las diferentes clases: (a) Edificación, (b) Vía, (c) Suelo Desnudo, (d) Vegetación, (e) Playa y (f) Agua***



	TXVAR_B3	TXENT_B3	TXRAN_B4	TXAVG_B4	TXVAR_B4	TXENT_B4	TXRAN_B5	TXAVG_B5	TXVAR_B5	TXENT_B5	CLASE
▸	165,542938	-0,474063	3,607575	14,729596	7,932168	-0,721168	0,023514	0,046265	0,000079	-0,717284	
	354,450378	-0,486025	0,487142	16,304523	0,163591	-0,915461	0,0297	0,039285	0,000123	-0,592559	
	118,939247	-0,694467	2,746876	13,458472	5,032886	-0,832639	0,025259	0,088358	0,000092	-0,630464	
	115,243645	-0,633918	3,071428	15,590381	8,172515	-1,003786	0,025489	0,060559	0,000074	-0,585833	
	510,861115	-0,429023	0,006667	16,386665	0,000017	-0,924196	0,039804	0,071632	0,000186	-0,5453	
	273,174225	-0,689807	0,370303	16,510015	0,172882	-0,923619	0,032437	0,073933	0,000132	-0,618223	
	117,599136	-0,662119	1,26458	15,625351	1,927645	-0,921114	0,017881	0,083598	0,000082	-0,677709	
	237,296402	-0,749809	2,236	16,734411	4,331323	-0,90976	0,026575	0,067931	0,000265	-0,609969	

Figura 5.7. ***Muestra de la tabla con los segmentos seleccionados y donde el valor dela clase debe ser introducido***

Es necesario realizar una selección de segmentos que pertenezca a una misma clase pero con características diferentes, para que así se pueda obtener un mejor modelo (como por ejemplo seleccionar edificaciones altas pero también bajas). La selección debe ser representativa de cada una de las particiones, esto quiere decir que si hay gran presencia de edificios, entonces tendrá que haber un número elevado de instancias de entrenamiento etiquetadas como edificación con respecto a las otras clases.

Con la finalidad de mejorar el modelo aportando al entrenamiento la máxima información posible se añade más información por medio de un atributo relacionado con los estratos. Este tipo de información separa la información de la imagen en

diferentes usos como podría ser: uso residencial, forestal, agrícola, etc. dando una información muy útil para la clasificación. Como esa información a veces no existe o es difícil conseguirla, es necesario que el propio usuario identifique esas zonas manualmente.
Ya que este proyecto está relacionado con la minería de datos y lo que se busca es una automatización de los procesos, se utilizará la cartografía catastral, de acceso público, que indicará si un segmento es de tipo urbano, rústico u otro (zonas de playa y mar). Al estar esta cartografía disponible para todo el que lo desee, cruzando esta información con la de los segmentos el proceso se automatiza, sin necesidad de realizar ninguna operación de edición manual.

En lugar de dividir la imagen original en diferentes imágenes según el estrato y obtener diferentes modelos, lo que se hará será añadir un nuevo atributo en el que se indique a qué estrato corresponde cada segmento, siendo éste un atributo de tipo nominal con los valores U, R y O (Urbano, Rústico y Otro, respectivamente).

Para poder realizar este proceso es necesario disponer de los datos del Catastro, los cuales se pueden descargar gratuitamente de su página web (www.sedecatastro.gob.es). El área de estudio comprende partes del término municipal de Sagunto y del de Canet d'en Berenguer, por lo que es necesario descargarse los datos de ambos municipios así como la cartografía rústica y urbana, que vienen por separado.

La capa que interesa es la llamada MASA, que contiene las manzanas de urbano y los polígonos de rústico, ya que para este caso no interesa información más detallada como puede ser la de las parcelas.
Como la cartografía catastral viene separada por municipios y por urbano o rústico, lo primero es juntar, como se ve en la Figura 5.8, toda la información en una misma capa en la que sólo habrá un atributo en el que se indique a qué tipo de Catastro corresponde.

Figura 5.8. ***Cartografía con los tipos de Catastro***

Es posible observar que únicamente las manzanas de las casas están marcadas como uso urbano, el resto corresponde con el polígono rústico (incluyendo las vías urbanas), por lo que este nuevo atributo ofrece información extra que puede ser de utilidad.

El siguiente paso será cruzar esta información con la de los segmentos generados. El problema es que los límites de los segmentos y los del Catastro serán muy similares pero no tienen por qué coincidir del todo, por lo que esta operación espacial para asignar un tipo a cada segmento se complica un poco, al no ser una intersección directa.
La solución adoptada es analizar cada segmento y ver qué tipo de Catastro es mayoritario en ese objeto, y tras esa comprobación se le asigna la clase mayoritaria (según el área que ocupa).

6. Generación de Modelos de Minería de Datos

En la Figura 6.1 se puede observar un diagrama de flujo en el que se indican los diferentes procesos seguidos en la generación de modelos de minería de datos.
Llegados a este punto ya se dispone de 20 ficheros en formato shapefile con los segmentos, sus características y la clase a la que pertenecen (si éstos han sido etiquetados en la muestra de aprendizaje).
El siguiente paso será introducir esta información en el software Weka, para poder llevar a cabo el aprendizaje de los modelos. Para ello es necesario escribir unas líneas de código en lenguaje Python que transformen los shapefiles en ficheros de entrada de Weka de tipo .xrff.

En este proyecto se ha utilizado el formato .xrff para Weka, el cual tiene la estructura de un documento XML, ya que permite definir qué atributo muestra la etiqueta en cada instancia y dar un peso según el área a cada instancia. El formato más conocido, el .arff, también permite ponderar cada instancia, sin embargo, el atributo que indica la etiqueta de cada instancia es el que se encuentra en último lugar, por lo que si algún nuevo atributo es añadido al final, se deben reordenar los atributos, mientras que en un fichero .xrff el atributo clase ya se encuentra etiquetado como tal y puede ocupar cualquier posición.
Hay que tener en cuenta que para este proyecto es muy importante la ponderación de las instancias, ya que al trabajar con segmentos de diferentes dimensiones, cada uno tendrá un área y no se le puede dar el mismo peso a un objeto pequeño que a uno grande, por lo que se realiza una ponderación a partir del área de cada objeto, buscando así que se clasifique correctamente la mayor superficie posible.

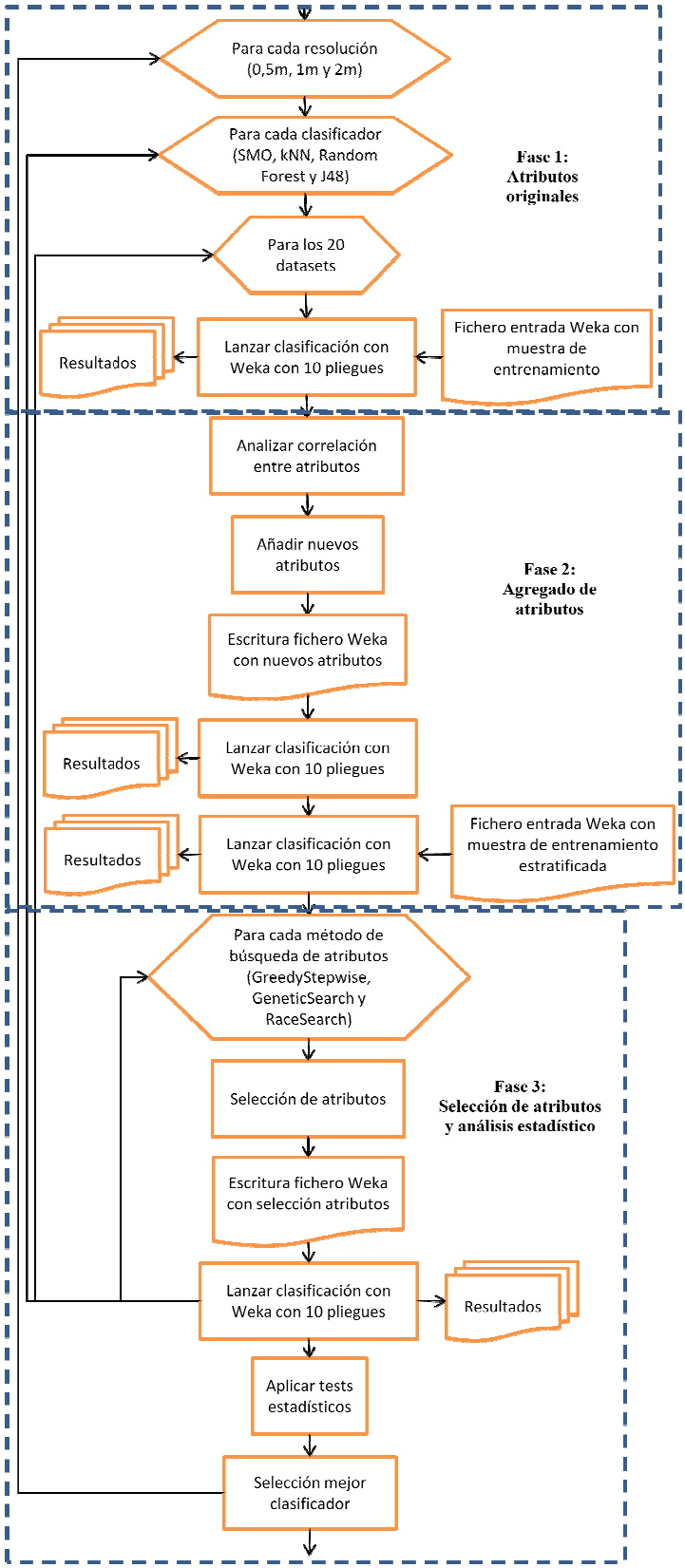


Figura 6.1. ***Detalle de las diferentes etapas realizadas durante la fase de generación de modelos de minería de datos***

Uno de los problemas que aparece cuando se trabaja con un fichero muy grande es que se producen algunos errores por falta de memoria, ya que las matrices no pueden almacenar tanta información. Para solucionar este problema cuando el número de instancias es muy elevado se divide el shapefile en las partes necesarias, escribiendo en el fichero .xrff una parte y a continuación las otras, quedando nuevamente en el fichero de entrada de Weka toda la zona de estudio unificada.

- **Fase 1: Utilización de Weka con los atributos originales generados por ENVI**

Tras haber obtenido el fichero de entrada con todos los atributos generados por ENVI es el momento de comenzar a trabajar con Weka.
Como se observará en éste y en los siguientes apartados, lo que se pretende es realizar diferentes pruebas añadiendo y eliminando atributos y probando con diferentes clasificadores, pero siempre con el mismo conjunto de datos de entrenamiento; de esta manera lo que se busca es hacer un estudio acerca de qué clasificadores y atributos pueden aumentar la precisión a la hora de clasificar una imagen con estas características.

Para la evaluación de los modelos se aplicará validación cruzada de 10 pliegues con cada clasificador (SMO, kNN con k=5, Random Forest y J48) para cada uno de los 20 *datasets* de los que se compone la zona con los atributos originales generados por ENVI. Estos resultados serán almacenados en una tabla Excel para poder ser comparados posteriormente con el resto de pruebas. Este mismo proceso se repetirá con cada una de las transformaciones que se le haga al fichero, las cuales se analizan en los siguientes apartados.

- **Fase 2: Agregación de nuevos atributos**

La siguiente prueba ya conlleva un poco más de trabajo, ya que se debe analizar la correlación entre los diferentes atributos y ver si el ratio entre dos atributos originales de ENVI es capaz de discriminar mejor una clase, de ese modo ese ratio sería añadido como un nuevo atributo.

Para poder obtener un primer listado sobre qué ratios entre atributos podrían ayudar a mejorar la clasificación se visualiza con Weka los datos como en la Figura 6.2, donde cada uno de los ejes corresponde con un atributo original generado por ENVI y los puntos están coloreados según la clase a la que pertenecen. De esta manera si unos puntos de un mismo color se encuentran separados o bien delimitados de otros, entonces el ratio entre esos dos atributos podría ser de ayuda para mejor la precisión.
En la Figura 6.2 se observa cómo alguna clase como la verde o la azul se compactan entre sí y sus valores quedan localizados.

Para comprobar qué ratios de los introducidos mejoran el modelo y cuáles no, se puede hacer de dos maneras: una primera sería ir añadiendo estos nuevos ratios uno a uno como atributos e ir analizando la precisión obtenida, o una segunda opción que sería añadir todos los anotados y posteriormente la selección de atributos se encargará de eliminar aquellos que no hagan mejorar el modelo.

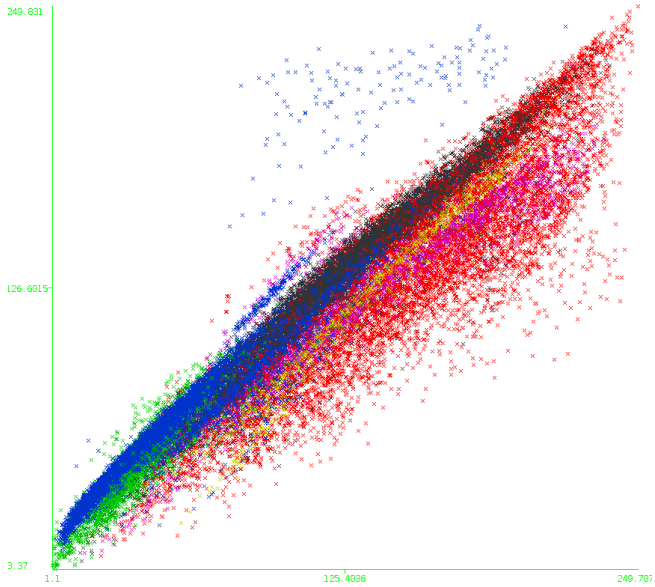


Figura 6.2. ***Distribución de las instancias con respecto a las variables TXAVG_B2 y TXAVG_B3, y representadas según la clase a la que pertenecen***

Tras haber añadido los nuevos atributos relacionados con los originales, hay un nuevo atributo que puede ser añadido y con el que conseguir buenos resultados. Este nuevo atributo está relacionado con el punto "Generación de vistas minables" del Capítulo 5, en el que cruzando los segmentos con los datos del Catastro se obtenía de qué tipo era cada objeto, es decir, si era de tipo rústico o urbano. Esta nueva información puede por ejemplo facilitar la diferenciación de edificaciones o solares con otro tipo de clases.

- **Fase 3: Selección de atributos**

Es importante mencionar los diferentes métodos que se van a utilizar para hacer una selección de atributos, ya que gracias a ellos se conseguirá disminuir el número de atributos, eliminando así ruido y disminuyendo el tiempo de cálculo.
Para poder hablar de la selección de atributos es necesario hablar en un primer lugar del método de evaluación de atributos, donde encontramos dos grupos principales de métodos: *filter* y *wrapper*, que buscan describir la naturaleza de la métrica utilizada para evaluar el valor de los atributos [22], dotándoles de un peso según su importancia. Los métodos de filtrado realizan una selección analizando si son relativamente independientes de la clasificación. Muchos métodos utilizan para ello coeficientes simples de correlación (como el criterio de discriminación de Fisher), mientras que

otros efectúan tests estadísticos (t-test, F-test) [23]. Un número elevado de métodos de filtrado producen un ranking de atributos en lugar de una lista con los mejores atributos, y el punto de corte en el ranking es escogido ejecutando diferentes test mediante validación cruzada.
En los *wrappers*, que es el método escogido en este TFM, se emplea un modelo predictivo para puntuar los subconjuntos de atributos. Cada subconjunto es utilizado para entrenar un modelo, que es evaluado sobre un conjunto de test, y a partir del resultado de la evaluación se le atribuye una puntuación al subconjunto. Como los métodos *wrapper* entrenan un modelo para cada subconjunto son más costosos, pero obtienen una mejor selección de atributos para cada conjunto de datos [26].

Una vez se han evaluado los atributos es necesario elegir un método de búsqueda, que será el encargado de generar el espacio de pruebas. Dentro de los *wrapper* se emplean los algoritmos de búsqueda para buscar a través del espacio de posibles atributos y evaluar cada subconjunto ejecutando un modelo con ese subconjunto.
Los métodos de filtrado son similares en la búsqueda pero en lugar de evaluar el conjunto contra un modelo, se evalúa un simple filtro [26].
A continuación se explica más detenidamente los tres métodos de búsqueda empleados: GreedyStepwise, GeneticSearch y RaceSearch.

GreedyStepwise realiza una búsqueda hacia adelante y hacia atrás a través del espacio del subconjunto de atributos. Puede empezar sin o con todos los atributos o desde un punto arbitrario en el espacio. Éste se detiene cuando la agregación o eliminación de los atributos restantes, disminuyen los resultados de la evaluación. También puede generar un ranking de atributos atravesando el espacio de un lado a otro y registrando el orden con el que los atributos son seleccionados.

GeneticSearch efectúa una búsqueda empleando un simple algoritmo genético descrito por Goldberg en 1989 [30]. Este tipo de algoritmos se denominan con este nombre porque se inspiran en la evolución biológica y su base genético-molecular. Estos algoritmos hacen evolucionar una población de individuos sometiéndola a acciones aleatorias semejantes a las que actúan en la evolución biológica (mutaciones), así como también a una selección de acuerdo con algún criterio, en función del cual se decide cuáles son los individuos más adaptados, que sobreviven, y cuáles los menos aptos, que son descartados [26].

Por último, RaceSearch tiene cuatro modos de funcionar pero la utilizada para este trabajo será la *Forward selection race* en la cual se comienza con una lista vacía de atributos y se van añadiendo, seleccionando el ganador y formando a partir de éste la nueva lista de atributos base. Este proceso se repite con la nueva lista de atributos base hasta que ya no se mejore la precisión [30].

Llegados a este punto ya se han añadido todos los nuevos atributos, por lo que es el momento de realizar una selección de atributos para empezar a eliminar todos aquellos

que sólo hacen que añadir ruido o están muy correlados con otros, por lo que la utilización de memoria es mayor y las clasificaciones más pobres.

Como ya se comentó en párrafos anteriores se utilizan tres métodos de búsqueda de atributos como son GreedyStepwise, GeneticSearch y RaceSearch.
Cada uno de estos métodos dará un listado diferente de los atributos que se deben de mantener, por lo que cada resultado corresponderá con una prueba. De esta manera el fichero que se tenía de la fase anterior será modificado eliminando los atributos correspondientes y lanzando el proceso de clasificación para cada clasificador.

También hay que tener en cuenta que la zona de estudio está dividida en 20 *datasets*, por lo que para cada *dataset* y cada método de búsqueda se mostrará un listado diferente de atributos. La manera con la que se procede para obtener unos atributos en común para los 20 *datasets* es viendo para todos ellos cuáles son los atributos que más aparecen en los resultados para ser mantenidos, y esos serán los seleccionados para toda la zona de trabajo.

- **Análisis de los resultados:**

El test estadístico que se explica a continuación se aplicará sobre 20 *datasets* para comparar diferentes clasificadores y métodos. Este test dirá si la diferencia de resultados obtenidos entre diferentes clasificadores es estadísticamente significativa y se puede concluir que un clasificador obtiene mejores resultados que otro.

El test de los rangos con signo de Wilcoxon (Wilcoxon, 1945) es un test no paramétrico que permite comparar un clasificador con otro con el objetivo de afirmar si uno mejora los resultados del otro y esta diferencia es significativa, o si por el contrario los resultados no se pueden considerar diferentes ya que la diferencia existente puede deberse a la aleatoriedad [25].
Para ello se obtienen las diferencias de valores entre los dos clasificadores para cada *dataset*. Seguidamente se ordena el valor absoluto de estas diferencias de menor a mayor, asignándoles su orden correspondiente y un orden promedio si hay diferencias que se repiten. Conociendo el orden de cada diferencia y el signo de la misma se obtienen la suma de los órdenes (rankings) positivos R^+ y la de los negativos R^-.
Para poder rechazar la hipótesis nula, el menor de los valores entre R^+ y R^- debe ser menor o igual que el valor correspondiente de la tabla de valores críticos de Wilcoxon, o el p-value menor que $1-\alpha$ (en este caso se toma $\alpha=0.9$, que es el valor más utilizado [25]), o como se verá más tarde, observar la salida de Keel donde estas comprobaciones se presentan en forma de tabla comparando todos los clasificadores entre sí.

Este test se encuentra disponible en un módulo del software Keel. Los valores a analizar corresponden al índice Kappa ponderado por la superficie de la muestra de entrenamiento de cada *dataset*.

Como se ha comentado, la herramienta Keel mostrará como salida una matriz en la que se muestra si el resultado entre dos clasificadores es significativamente diferente y quién gana a quién.

Asimismo, en la matriz también se indica para cada clasificador las veces que gana y las veces que empata o gana, así como el ranking. El clasificador que mejor ranking obtenga será porque tiene menos derrotas, es decir, que empata o gana más veces. Si en esa columna hay un empate, entonces obtiene mejor ranking el que posea un mayor número de victorias. Si aún así existe algún empate, entonces se asigna un ranking promedio entre los clasificadores empatados, de esta manera si hay tres clasificadores que quedan empatados para la posición 7, entonces a los tres se les atribuye la posición 8, y el siguiente clasificador estará situado en la posición 10.

Con estos rankings ya se puede ver cuál ha sido el clasificador que ha realizado una mejor clasificación, y si la diferencia con el siguiente es pequeña, o éste posee muchos menos atributos que el primero, entonces se pueden tener los dos en cuenta.

7. Resultados del Modelado

Una vez se han explicado en el Capítulo 6 todos los pasos necesarios para generar los modelos de minería de datos, en este capítulo se presentan los resultados obtenidos en cada uno de los puntos indicados. El mismo proceso se repite para las tres granularidades estudiadas: 0.5, 1 y 2 metros.

7.1. Resolución 0,5 metros

7.1.1. Atributos agregados

En el Capítulo 6 se explicaba cómo se debían analizar las gráficas en Weka para poder determinar si el ratio entre dos atributos originales aportaba una información extra con tal de ayudar en la discriminación de clases.
A continuación, en la Tabla 7.1 se enumeran los nuevos atributos añadidos tras haber analizado las gráficas:

Nuevo Atributo	Descripción
AVG_B2/AVG_B3	Ratio entre la media espectral de la banda del rojo y la del verde
FX_COMPACT/TXRAN_B5	Ratio entre la compacidad y el rango de textura del NDVI
AVG_B2/TXAVG_B3	Ratio entre la media espectral de la banda del rojo y la media de textura de la banda del verde
FX_CONVEX/TXAVG_B5	Ratio entre la convexidad y la media de textural del NDVI
FX_ELONG/MAX_B5	Ratio entre la elongación y el máximo espectral del NDVI
AVG_B2/MIN_B4	Ratio entre la media espectral de la banda del rojo y el mínimo espectral de las alturas
FX_ROUND/MIN_B5	Ratio entre la redondez y el mínimo espectral del NDVI
FX_COMPACT/MIN_B4	Ratio entre la compacidad y el mínimo espectral de las alturas
TXAVG_B2/TXAVG_B3	Ratio entre la media de textura de la banda del rojo y la del verde
AVG_B1/TXAVG_B3	Ratio entre la media espectral del infrarrojo cercano y la media de textura del verde
TXAVG_B1/TXAVG_B3	Ratio entre la media de textura del infrarrojo cercano y la del verde
MIN_B2/TXAVG_B4	Ratio entre el mínimo espectral del rojo y la media de textura de las alturas
AVG_B1/AVG_B3	Ratio entre la media espectral del infrarrojo cercano y la del verde
AVG_B3/TXAVG_B1	Ratio entre la media espectral del verde y la media de textura del infrarrojo cercano
MAX_B1/TXAVG_B5	Ratio entre el máximo espectral del infrarrojo cercano y media de textura del NDVI

Tabla 7.1. ***Enumeración y descripción de los nuevos atributos agregados***

Los nuevos atributos creados son tal como aparecen en el listado, es decir, el cociente entre ambos atributos originales generados por ENVI. La explicación de cada atributo se encuentra en el apartado 4.2. La banda 1 (B1) corresponde con el infrarrojo cercano, la segunda (B2) con el rojo, la tercera (B3) con el verde, la cuarta (B4) con la altura de los objetos y la quinta (B5) con el índice NDVI (*Normalized Difference Vegetation Index*). Este índice sirve para medir el crecimiento de las plantas, determinar coberturas vegetales y controlar la biomasa [31]. La fórmula para calcular este índice es:

$$NDVI = \frac{infrarrojo\ cercano - rojo}{infrarrojo\ cercano + rojo}$$

Cuando la vegetación es muy vigorosa refleja mucha radiación solar en el infrarrojo cercano y poca en el rojo, por lo que se obtiene un valor de NDVI elevado. Sin embargo, cuando está enferma el valor es bajo. Por lo tanto, este índice no es sólo interesante para conocer el estado de la vegetación, sino también para diferenciar la vegetación de otros objetos.

Analizando el listado de los nuevos atributos se puede observar que las combinaciones de las bandas 1 y 2 con la banda 3 aparecen en varias ocasiones. Esto puede ser debido a que tal vez esas bandas no aportan toda la información de forma individual pero sí combinándolas entre sí. También se observa que la banda del índice NDVI suele obtener buenos resultados combinada con atributos espaciales, ya que a lo mejor discrimina objetos alargados con NDVI bajo como una vía u objetos con un NDVI más elevado y forma más redonda como un árbol.

7.1.2. Selección de atributos

Tras haber añadido variables al listado original, ahora es el momento de hacer una selección de las mismas. A continuación se nombrarán los atributos seleccionados para cada uno de los métodos de búsqueda empleados.

- **Greedy Stepwise:**

Los atributos seleccionados por el método Greedy Stepwise para una resolución espacial de 0,5 metros son los siguientes (Tabla 7.2):

Atributo	Descripción
AVG_B4	Media espectral de las alturas
TXAVG_B4	Media de textura de las alturas
AVG_B2/TXAVG_B3	Ratio entre la media espectral del rojo y la media de textura del verde
FX_CONVEX/TXAVG_B5	Ratio entre la convexidad y la media de textura del NDVI
TXAVG_B2/TXAVG_B3	Ratio entre la media de textura del rojo y del verde
TXAVG_B1/TXAVG_B3	Ratio entre la media de textura del infrarrojo cercano y del verde
MIN_B2/TXAVG_B4	Ratio entre el mínimo espectral del rojo y la media de textura de las alturas
TIPO	Tipo de Catastro

Tabla 7.2. ***Enumeración y descripción de la selección de atributos con Greedy Stepwise para una resolución de 0,5 metros***

Como se observa la reducción de atributos es significativa, ya que se pasa de 70 (sin contar el de la clase) a 8. De los atributos originales se mantienen los que se refieren a la banda 4, que es la altura de los objetos, y hay muchos de los introducidos en el apartado anterior que se mantienen, lo que quiere decir que la agregación de los nuevos atributos ha sido buena. Entre estas nuevas variables sigue siendo significativa la combinación entre las bandas 1, 2 y 3, y la banda 5 con los atributos espaciales. Una variable muy importante y que también ha sido seleccionada es TIPO, que corresponde con el tipo de Catastro al que pertenece el objeto.

- **Genetic Search:**

Con el método Genetic la selección de atributos es muchísimo mayor, por lo que no aporta tanto como Greedy. Los atributos seleccionados son 68, que corresponden con los originales y los añadidos menos STD_B1 y TXRAN_B1 (banda del infrarrojo cercano), que han sido eliminados.

- **Race Search:**

Este método también reduce mucho el número de variables, quedando tras las selección 6 atributos (Tabla 7.3):

Atributo Descripción	Descripción
TXAVG_B4	Media de textura de las alturas
TXAVG_B5	Media de textura del NDVI
TXAVG_B2/TXAVG_B3	Ratio entre la media de textura del rojo y del verde
TXAVG_B1/TXAVG_B3	Ratio entre la media de textura del infrarrojo cercano y del verde
MIN_B2/TXAVG_B4	Ratio entre mínimo espectral del rojo y media de textura de las alturas
TIPO	Tipo de Catastro

Tabla 7.3. ***Enumeración y descripción de la selección de atributos con Race Search para una resolución de 0,5 metros***

Más o menos realiza la misma selección que Greedy, pero en este caso también selecciona la media de las texturas del índice NDVI.

7.1.3. Resultados de la generación de modelos

En las Tablas 7.5 y 7.6 se muestran los resultados de cada *dataset* para cada clasificador. Estas tablas son una única pero han sido divididas para poder ser presentadas. Como ya se ha comentado la medida que aparece en cada celda es el índice Kappa ponderado por la superficie del conjunto de entrenamiento de cada

dataset. En rojo se marca el índice más alto de cada *dataset*. La última fila corresponde con el índice medio de cada clasificador, y en esta fila también se encuentra en rojo la media más elevada.
En el Capítulo 6 se comentó que los 4 clasificadores que se iban a utilizar en cada una de las pruebas eran SMO, kNN (k=5), Random Forest y J48. En las Tablas 7.5 y 7.6 las columnas indicadas como "Original" corresponden con las pruebas realizadas con los atributos originales generados por ENVI; "Extendido" son aquellas en las que se han añadido nuevos atributos correspondientes con ratios calculados a partir de dos atributos originales; las de "Catastro" contienen los atributos de "Extendido" y se les añade uno más en el que se muestra si el tipo de Catastro es rústico, urbano u otro; y por último, el resto de columnas corresponden con la selección de atributos, donde se diferencian dos de los tres métodos de búsqueda: Greedy Stepwise y Race Search. Como se puede observar no se tienen en cuenta los resultados obtenidos con el método de búsqueda Genetic Search, ya que no hace una gran reducción en el número de los atributos, por lo que éstos son muy similares a los de "Catastro".

Como se puede observar los métodos que obtienen mejores resultados en los *datasets* son el SMO con los atributos de "Catastro". Sin embargo, la media más elevada teniendo en cuenta los 20 *datasets* la consigue el método SMO con Race, a pesar de no conseguir el mejor resultado en ningún *dataset*, pero casi siempre obteniendo buenos resultados.

Con el objeto de ahorrar espacio, en las tablas de resultados se usarán una serie de abreviaturas para los siguientes clasificadores y métodos que se enumeran en la Tabla 7.4:

Clasificador/Método	Abreviatura
Random Forest	RF
Original	O
Extendido	E
Catastro	C
Greedy Stepwise	G
Race Search	R

Tabla 7.4. ***Abreviaturas de los clasificadores y métodos utilizadas en las tablas de resultados***

Datasets	SMO-O	kNN-O	RF-O	J48-O	SMO-E	kNN-E	RF-E	J48-E	SMO-C	kNN-C	RF-C
Dataset 1	70477,7	80908,5	78444,5	65344,4	**81623,0**	81072,7	69607,1	57131,2	73097,7	79717,5	69442,8
Dataset 2	46792,8	29584,3	46929,1	45050,5	45629,7	30718,3	46568,9	45366,9	34202,8	31039,5	33516,6
Dataset 3	29625,0	27348,3	29161,7	28517,3	**29717,0**	27854,5	29088,0	28544,9	29579,0	26660,9	29324,3
Dataset 4	51687,4	50040,3	50704,3	49794,5	51901,7	50981,5	50306,9	51227,2	**51964,5**	51342,3	51839,0
Dataset 5	121040,1	121887,5	120520,8	264467,8	124047,0	122106,2	121860,2	111308,8	124539,0	123035,6	120821,5
Dataset 6	19834,2	34304,1	38206,5	36736,6	20063,8	37207,8	38142,1	38266,9	21054,5	36543,3	38931,4
Dataset 7	19768,7	15788,4	20583,5	15916,9	16956,1	15943,1	17862,4	17230,7	17749,2	17167,5	20437,6
Dataset 8	17869,3	17516,9	18003,6	20325,6	17905,0	17466,6	18009,9	20445,2	18200,7	18184,0	18290,9
Dataset 9	10478,4	7462,4	9454,3	8935,0	10777,3	8072,2	**11703,1**	10315,8	9028,1	8649,1	10479,7
Dataset 10	28089,5	51747,2	125832,8	110974,7	65849,6	121567,5	126357,9	63902,7	57011,6	122912,4	126703,8
Dataset 11	27130,3	23003,2	26198,9	24610,2	27118,1	23837,1	26631,1	26189,8	27160,7	26049,8	26972,0
Dataset 12	20000,9	21277,3	23407,4	24080,0	24816,4	21623,9	24087,7	21955,1	**24953,9**	23942,4	24910,6
Dataset 13	19253,7	18560,3	18630,8	11275,2	19332,0	18732,6	18856,0	11623,9	**19343,8**	18932,4	19059,8
Dataset 14	22117,1	21877,5	23639,9	21111,1	22298,6	21880,1	24250,1	23912,2	22480,1	22230,5	23428,2
Dataset 15	26407,4	24957,3	34048,4	19135,9	27998,2	25221,3	27846,9	25362,1	26889,5	25355,0	**34551,7**
Dataset 16	8686,9	10771,2	11428,8	10282,1	9791,8	11621,4	11504,8	12099,1	9794,3	11572,0	11869,8
Dataset 17	**27642,1**	26587,0	25498,9	21831,7	26370,6	26563,0	24843,6	19556,3	26857,6	23629,2	26472,8
Dataset 18	14788,9	13938,5	14508,4	14138,5	14902,5	14253,8	14713,1	14519,0	**14990,5**	14441,7	14820,7
Dataset 19	13315,1	12241,2	14535,5	13709,4	13545,8	12317,6	14038,3	14569,8	13654,9	13207,6	14817,6
Dataset 20	26806,0	29565,8	26787,7	26879,0	27521,5	**29989,3**	27200,3	29186,2	27335,3	24531,7	27346,3
Media	31090,6	31968,4	37826,3	41655,8	33908,3	35951,5	37173,9	32135,7	32494,4	35957,2	37201,8

Tabla 7.5. ***Índice Kappa ponderado obtenido por cada método en cada dataset, y media de cada método para una resolución de 0,5 metros***

Datasets	J48-C	SMO-G	kNN-G	RF-G	J48-G	SMO-R	kNN-R	RF-R	J48-R
Dataset 1	59554,1	80908,5	24311,1	59808,7	59669,1	80974,2	24344,0	60391,8	59332,3
Dataset 2	28542,9	31618,6	30820,5	**47605,5**	33297,6	47255,1	31745,1	47235,7	30956,7
Dataset 3	28931,5	29563,6	29422,5	28772,0	29177,0	29597,4	28026,4	29066,5	29210,7
Dataset 4	48356,6	50599,8	42474,1	38677,9	42735,5	50484,7	42510,7	38092,3	38128,9
Dataset 5	111800,9	**265971,2**	265287,9	264686,5	265096,5	265834,6	265779,9	264959,8	265451,9
Dataset 6	37574,2	37191,7	39076,4	39189,2	39245,6	37654,8	**39527,5**	39205,3	37900,4
Dataset 7	20226,2	20152,2	20699,0	20773,1	20590,1	20119,5	20391,8	**21202,3**	18080,3
Dataset 8	18200,7	20577,3	**20667,5**	20527,0	20516,5	20560,5	20527,0	20516,5	20388,5
Dataset 9	8304,3	8280,7	11531,3	8925,8	9387,4	8431,5	9799,1	9614,2	9112,0
Dataset 10	63287,8	125576,6	126562,9	46265,1	45547,8	125461,3	126691,0	**126806,2**	126153,0
Dataset 11	26685,9	26348,1	27270,3	27011,6	24963,2	26381,5	**27644,6**	26755,9	25818,5
Dataset 12	22077,4	24021,4	24928,5	23430,3	23455,8	23960,3	24908,1	24258,4	22398,4
Dataset 13	11694,4	18728,7	13447,6	19134,2	18965,7	18407,5	13010,8	18844,3	18799,2
Dataset 14	24005,5	24320,7	**24603,1**	23970,2	23478,6	23771,1	22913,8	23246,6	23655,1
Dataset 15	25383,2	33196,6	34055,4	32306,2	31795,8	33256,5	34027,2	31074,3	31732,5
Dataset 16	12100,4	11923,0	**12495,7**	12090,2	12092,8	11873,6	12483,0	12018,0	12015,5
Dataset 17	19634,4	22874,8	25709,3	25480,9	25553,0	26124,1	25036,0	25664,2	25060,0
Dataset 18	14549,3	14752,5	14808,6	14587,2	14513,0	14514,5	14751,0	14713,1	14629,7
Dataset 19	14498,1	8966,6	14318,9	13511,5	13400,8	9936,1	**14881,5**	14795,8	14432,6
Dataset 20	29313,9	26714,7	27218,5	27032,3	26893,6	26316,8	29730,1	27255,0	24907,7
Media	31236,1	44114,4	41485,4	39689,3	39018,8	**45045,8**	41436,4	43785,8	42408,2

Tabla 7.6. ***Índice Kappa ponderado obtenido por cada método en cada dataset, y media de cada método para una resolución de 0,5 metros***

7.1.4. Test de Wilcoxon

Tras introducir las tablas anteriores (salvo la fila de media) en Keel, la herramienta facilita una tabla en la que se indica qué método gana a quién y si esa diferencia es significativa o hay un empate. Esos resultados han sido transformados en la Tabla 7.7. En ella se muestra el resultado tras aplicar el test de Wilcoxon de comparar la fila con la columna, de este modo si aparece una G, significa que el método de la fila gana al de la columna, si aparece una P es que pierde, y si hay una E es que empatan.

Como ya se explicó en el Capítulo 6, en la Tabla 7.8 se muestran 3 columnas a modo resumen. En ellas se indica el número de victorias, no derrotas (empate o victoria) y ranking de cada método. Como ya se dijo, el ranking viene por el mayor número de no derrotas, en caso de empate se observa el número de victorias, y si el empate continúa se calcula un ranking promedio entre los métodos afectados.

Datasets	SMO-O	kNN-O	RF-O	J48-O	SMO-E	kNN-E	RF-E	J48-E	SMO-C	kNN-C	RF-C	J48-C	SMO-G	kNN-G	RF-G	J48-G	SMO-R	kNN-R	RF-R	J48-R
SMO-O	-	E	E	E	P	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E
kNN-O	E	-	P	E	P	P	P	E	P	P	P	E	P	P	E	E	P	P	P	P
RF-O	E	G	-	G	E	G	E	E	E	G	P	G	E	E	E	E	E	E	E	E
J48-O	E	E	P	-	E	E	P	E	E	E	P	E	P	P	E	E	P	P	P	E
SMO-E	G	G	E	E	-	E	E	E	E	G	E	G	E	E	E	E	E	E	E	E
kNN-E	E	G	P	E	E	-	P	E	E	E	P	E	P	E	E	E	P	E	P	E
RF-E	E	G	E	G	E	G	-	G	E	G	E	G	E	E	E	E	E	E	E	E
J48-E	E	E	E	E	E	E	P	-	E	E	P	E	E	E	E	E	P	E	P	E
SMO-C	E	G	E	E	E	E	E	E	-	E	E	E	E	E	E	E	P	E	E	E
kNN-C	E	G	P	E	P	E	P	E	E	-	P	E	P	P	P	E	P	P	P	P
RF-C	E	G	G	G	E	G	E	G	E	G	-	G	E	E	E	G	E	E	E	G
J48-C	E	E	P	E	P	E	P	E	E	E	P	-	P	P	E	E	P	P	P	E
SMO-G	E	G	E	G	E	G	E	E	E	G	E	G	-	E	E	E	E	E	E	E
kNN-G	E	G	E	G	E	E	E	E	E	G	E	G	E	-	G	E	E	E	E	G
RF-G	E	E	E	E	E	E	E	E	E	G	E	E	E	P	-	E	E	E	E	G
J48-G	E	E	E	E	E	E	E	E	E	E	P	E	E	E	E	-	E	E	E	E
SMO-R	E	G	E	G	E	G	E	G	G	G	E	G	E	E	E	E	-	E	E	G
kNN-R	E	G	E	G	E	E	E	E	E	G	E	G	E	E	E	E	E	-	E	G
RF-R	E	G	E	G	E	G	E	G	E	G	E	G	E	E	E	E	E	E	-	G
J48-R	E	G	E	E	E	E	E	E	E	G	P	E	E	P	P	E	P	P	P	-

Tabla 7.7. ***Resultados de comparación entre los diferentes métodos tras aplicar el test de Wilcoxon para un resolución de 0,5 metros***

Datasets	Victorias	Victoria/Empate	Ranking
SMO-O	0	21	13
kNN-O	0	8	20
RF-O	7	22	9
J48-O	0	13	17,5
SMO-E	6	23	8
kNN-E	1	16	16
RF-E	8	23	4,5
J48-E	0	18	14
SMO-C	1	22	11
kNN-C	1	10	19
RF-C	**11**	**23**	**1**
J48-C	0	13	17,5
SMO-G	7	23	6,5
kNN-G	8	23	4,5
RF-G	3	22	10
J48-G	0	22	12
SMO-R	**10**	**23**	**2**
kNN-R	7	23	6,5
RF-R	**9**	**23**	**3**
J48-R	3	17	15

Tabla 7.8. ***Resumen con los resultados de comparación de cada método y su posición en el ranking para una resolución de 0,5 metros***

En la Tabla 7.8 se observa que el método mejor clasificado es el Random Forest con la agregación de todos los nuevos atributos, incluido el tipo de Catastro. También se observa que tras él se encuentran el SMO Race y el Random Forest Race. Estos métodos también se van a tener en cuenta ya que quedan muy próximos al primer clasificado y porque estos últimos reducen mucho el número de variables, mientras que el Random Forest tiene todos los atributos posibles.

7.2. Resolución 1 metro

7.2.1. Atributos agregados

Las variables añadidas para una resolución de 1 metro son las mismas que las mencionadas para 0,5 metros.

7.2.2. Selección de atributos

- **Greedy Stepwise**

Los atributos seleccionados por el método de búsqueda Greedy para una resolución de 1 metro son (Tabla 7.9):

Atributo	Descripción
AVG_B4	Media espectral de las alturas
AVG_B5	Media espectral del NDVI
TXAVG_B4	Media de textura de las alturas
TXAVG_B5	Media de textura del NDVI
AVG_B2/AVG_B3	Ratio entre la media espectral del rojo y del verde
TIPO	Tipo de Catastro

Tabla 7.9. ***Enumeración y descripción de la selección de atributos con Greedy Stepwise para una resolución de 1 metro***

Como se puede observar las variables son diferentes a las seleccionadas para la resolución de 0,5 metros, habiéndose seleccionado en este caso un menor número de nuevas variables agregadas. Esta diferencia de atributos entre resoluciones puede ser debida a que los píxeles contienen valores más mezclados, y por lo tanto los segmentos generados también, haciendo que los valores y los atributos para identificar una clase en una granularidad u otra no sean los mismos.
Como se puede observar, tanto la banda de la altura de los objetos como la del índice NDVI tienen gran importancia, así como la variable referente al tipo de Catastro, que hace mejorar todos los modelos.

- **Genetic Search**

Para esta granularidad los atributos seleccionados por Genetic han sido 66, por lo que sigue siendo un número muy elevado con respecto al número original. De todos los atributos introducidos los no seleccionados han sido: MAX_B2 (banda del rojo), MIN_B3, TXENT_B3 (banda del verde) y FX_CONVEX/TXAVG_B5 (banda del NDVI).

- **Race Search**

La reducción efectuada por Race en este caso también es considerable, pasando de 70 a 5 variables (Tabla 7.10):

Atributo	Descripción
AVG_B2/AVG_B3	Ratio entre la media espectral del rojo y la del verde
TXAVG_B2/TXAVG_B3	Ratio entre la media de textura del rojo y la del verde
TXAVG_B1/TXAVG_B3	Ratio entre la media de textura del infrarrojo cercano y la del verde
MIN_B2/TXAVG_B4	Ratio entre mínimo espectral del rojo y la media de textura de las alturas
TIPO	Tipo de Catastro

Tabla 7.10. ***Enumeración y descripción de la selección de atributos con Race Search para una resolución de 1 metro***

Como se puede observar, en este caso todos los atributos seleccionados han sido añadidos en el proceso de agregación, lo que habla de la importancia de los mismos. Se siguen manteniendo las relaciones entre las bandas 1, 2 y 3, así como la información sobre los estratos.

7.2.3. Resultados de la generación de modelos

En las Tablas 7.11 y 7.12, que originariamente eran una sola tabla pero han sido divididas para poder ser expuestas en este documento, se muestran los resultados de cada uno de los 20 *datasets* con cada clasificador. Al igual que en la granularidad anterior se marcan en rojo los mejores índices para cada *dataset*, y en este caso además se subraya el valor si existe un empate entre dos clasificadores.
Para esta resolución de 1 metro tampoco se tienen en cuenta los resultados obtenidos con el método de búsqueda Genetic Search, ya que no hace una gran reducción en el número de los atributos, por lo que éstos son muy similares a los de "Catastro".

Como se puede observar el clasificador que obtiene un índice Kappa ponderado más elevado en un mayor número de *datasets* es el SMO con todos los nuevos atributos añadidos, incluido el del Catastro. Sin embargo, y como pasaba para una resolución de 0,5 metros, el que obtiene una mejor media es el SMO pero para una selección de atributos con el método Greedy.

Datasets	SMO-O	kNN-O	RF-O	J48-O	SMO-E	kNN-E	RF-E	J48-E	SMO-C	kNN-C	RF-C
Dataset 1	89206,3	89186,1	76139,1	94936,4	89792,4	89418,5	98594,8	85295,2	88104,7	88185,6	87720,7
Dataset 2	78568,7	54792,3	58168,0	71760,3	79178,7	55898,5	57997,2	72215,8	**80358,2**	60608,3	77714,6
Dataset 3	50264,0	41710,9	47643,8	48796,3	50243,3	42646,3	47886,7	49881,5	**50827,3**	42739,4	47974,5
Dataset 4	58934,2	57920,7	58308,0	53602,3	58243,5	58417,8	61774,4	55680,8	58966,4	59837,9	62387,6
Dataset 5	-55869,6	-58323,7	114898,1	115028,7	-55765,2	-58114,8	-60281,7	8014,9	9503,0	117404,4	-59289,6
Dataset 6	53317,5	50674,5	51483,0	48109,5	54254,3	52085,2	53624,2	49024,0	**55469,8**	53010,8	53049,8
Dataset 7	26089,7	24505,9	25750,0	24266,7	26152,1	24817,8	31891,2	25569,8	26661,5	25895,6	25677,3
Dataset 8	36077,3	34142,5	38668,0	20340,4	36142,9	34113,8	**39619,0**	20406,0	28813,5	28264,2	36577,4
Dataset 9	18164,9	11616,3	14957,4	16136,1	20018,4	11802,5	15636,6	16775,9	**20844,4**	11679,8	18953,6
Dataset 10	66674,7	61786,6	179601,5	176787,1	67045,0	61601,5	180582,8	177305,5	83523,9	180860,5	182175,1
Dataset 11	45659,7	38478,7	44328,2	42144,0	45809,3	40817,5	45011,4	41780,0	**46278,1**	43166,3	44991,5
Dataset 12	44551,3	38620,1	41624,9	37558,4	43249,7	40304,8	42114,2	39935,6	43429,7	41061,8	43831,3
Dataset 13	**47165,3**	37853,4	43603,0	43742,4	43184,8	39156,2	39872,5	45102,9	46838,4	44641,4	44482,8
Dataset 14	48778,3	46995,7	50172,4	33528,1	51353,7	48985,8	50661,9	50124,5	47948,2	47783,2	47926,9
Dataset 15	36245,9	32543,3	40993,7	25688,2	41906,6	33029,6	41471,5	41480,0	**41949,2**	33183,1	41066,2
Dataset 16	29486,6	26881,0	27882,0	21403,8	29528,5	27726,6	28509,5	22581,1	29528,5	28375,0	25918,8
Dataset 17	30095,3	37846,8	39425,4	39192,1	40725,0	39370,5	38839,7	35124,1	41534,9	36185,7	37888,0
Dataset 18	17426,2	24989,8	26303,1	26492,4	27817,0	25469,9	26537,5	26506,5	**28076,8**	26201,4	27447,0
Dataset 19	18169,9	39749,4	38869,6	42370,5	18375,0	39954,5	39950,0	39927,2	18443,4	38746,5	43291,3
Dataset 20	33875,9	42074,5	39041,5	44242,4	34519,7	41772,7	39393,6	43845,0	35822,4	36028,7	44368,1
Media	38644,1	36702,2	52893,0	51306,3	40088,7	37463,8	45984,4	47328,8	44146,1	52193,0	46707,6

Tabla 7.11. ***Índice Kappa ponderado obtenido por cada método en cada dataset, y media de cada método para una resolución de 1 metro***

Datasets	J48-C	SMO-G	kNN-G	RF-G	J48-G	SMO-R	kNN-R	RF-R	J48-R
Dataset 1	86538,3	**99736,8**	27508,7	86912,2	97634,7	97372,0	27397,6	87619,6	96108,7
Dataset 2	77706,4	75559,0	72272,8	77730,8	79195,0	76689,7	75591,5	60681,5	61982,9
Dataset 3	49917,7	50362,2	50527,5	50279,5	50486,2	50682,6	47302,7	50734,3	50269,1
Dataset 4	58133,7	61948,6	**62929,8**	62290,8	52001,5	41976,9	51833,7	62813,6	62697,4
Dataset 5	-58193,1	**253762,8**	253527,8	11252,2	252535,7	252118,0	252091,9	248932,9	252561,8
Dataset 6	49682,0	53674,4	53936,4	51516,5	52358,4	50144,8	52079,6	50490,5	49096,5
Dataset 7	31146,1	32438,8	31901,6	26748,1	32404,1	32106,1	**33332,9**	32882,4	31849,6
Dataset 8	21275,0	31334,5	31555,9	20541,2	13502,9	32277,3	32170,7	18717,1	13334,8
Dataset 9	18081,7	18581,2	17284,2	15660,7	16285,1	17021,3	14760,2	14744,9	19267,0
Dataset 10	178398,0	181119,8	**182526,9**	182064,1	180638,3	179916,2	180527,3	181675,2	179564,4
Dataset 11	42124,0	42926,9	37057,4	34364,5	35546,4	42488,1	44831,9	37840,3	35840,6
Dataset 12	37526,1	44237,4	40304,8	40738,7	39063,2	44399,0	40941,8	**45631,4**	36741,4
Dataset 13	45146,2	46838,4	45583,6	44646,2	41867,5	46564,4	45001,9	45636,5	41184,9
Dataset 14	50326,7	49938,3	50544,9	48262,1	48485,6	49161,4	48836,8	49613,7	**51422,8**
Dataset 15	41497,1	34880,9	40328,2	40899,8	34944,9	40622,6	41202,7	38749,9	41032,1
Dataset 16	22581,1	28802,3	29211,7	27457,7	25814,2	29316,3	29202,8	25479,6	24699,7
Dataset 17	35723,6	38702,5	36675,4	40642,6	37778,1	38043,5	**42019,9**	40381,8	36606,7
Dataset 18	26523,4	27557,1	27624,9	27370,7	26856,7	23360,2	26952,7	27480,9	26461,3
Dataset 19	40278,2	**43327,8**	33686,7	38422,9	42329,5	34999,5	40301,0	43177,3	42051,4
Dataset 20	44171,9	48115,3	**48251,1**	39509,3	41692,2	41858,2	38588,8	45429,4	47280,4
Media	44929,2	**63192,2**	58662,0	48365,5	60071,0	61055,9	58248,4	60435,6	60002,7

Tabla 7.12. ***Índice Kappa ponderado obtenido por cada método en cada dataset, y media de cada método para una resolución de 1 metro***

7.2.4. Test de Wilcoxon

Tras haber aplicado el test de Wilcoxon con Keel, se toma la tabla de salida de esta herramienta y se transforma en las Tabla 7.13, donde se indica el resultado de comparar el método de cada fila con el de la columna.

Seguidamente, en la Tabla 7.14 se muestran 3 columnas resumen en las que se indica el número de victorias, de no derrotas y el ranking de cada método.

Datasets	SMO-O	kNN-O	RF-O	J48-O	SMO-E	kNN-E	RF-E	J48-E	SMO-C	kNN-C	RF-C	J48-C	SMO-G	kNN-G	RF-G	J48-G	SMO-R	kNN-R	RF-R	J48-R
SMO-O	-	E	E	E	P	E	E	E	P	E	E	E	P	E	E	E	E	E	E	E
kNN-O	E	-	P	E	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
RF-O	E	G	-	E	E	G	P	E	E	E	P	E	P	E	E	E	E	E	P	E
J48-O	E	E	E	-	E	E	E	E	E	E	P	E	P	P	E	P	P	P	P	E
SMO-E	G	G	E	E	-	G	E	E	P	E	E	E	E	E	E	E	E	E	E	E
kNN-E	E	G	P	E	P	-	P	E	P	E	P	E	P	P	E	E	P	P	P	P
RF-E	E	G	G	E	E	G	-	E	E	G	E	E	P	E	E	E	E	E	E	E
J48-E	E	G	E	E	E	E	E	-	E	E	P	P	P	P	E	E	E	E	P	E
SMO-C	G	G	E	E	G	G	E	E	-	E	E	E	E	E	E	E	E	E	E	E
kNN-C	E	G	E	E	E	E	P	E	E	-	P	E	P	P	E	E	P	P	P	E
RF-C	E	G	G	G	E	G	E	G	E	G	-	G	E	E	E	E	E	E	E	E
J48-C	E	G	E	E	E	E	E	G	E	E	P	-	P	E	E	E	E	E	E	E
SMO-G	G	G	G	G	E	G	G	G	E	G	E	G	-	E	G	G	G	G	G	G
kNN-G	E	G	E	G	E	G	E	G	E	G	E	E	E	-	E	E	E	E	E	G
RF-G	E	G	E	E	E	E	E	E	E	E	E	E	P	E	-	E	E	E	E	E
J48-G	E	G	E	G	E	E	E	E	E	E	E	E	P	E	E	-	E	E	E	E
SMO-R	E	G	E	G	E	G	E	E	E	G	E	E	P	E	E	E	-	E	E	E
kNN-R	E	G	E	G	E	G	E	E	E	G	E	E	P	E	E	E	E	-	E	E
RF-R	E	G	G	G	E	G	E	G	E	G	E	E	P	E	E	E	E	E	-	E
J48-R	E	G	E	E	E	G	E	E	E	E	E	E	P	P	E	E	E	E	E	-

Tabla 7.13. ***Resultados de comparación entre los diferentes métodos tras aplicar el test de Wilcoxon para una resolución de 1 metro***

Datasets	Victorias	Victoria/Empate	Ranking
SMO-O	0	18	15
kNN-O	0	3	20
RF-O	2	18	14
J48-O	0	15	18
SMO-E	3	22	8
kNN-E	1	10	19
RF-E	5	21	10
J48-E	1	17	16
SMO-C	4	23	4
kNN-C	1	15	17
RF-C	9	23	2
J48-C	2	20	13
SMO-G	**17**	**23**	**1**
kNN-G	7	23	3
RF-G	1	21	12
J48-G	2	22	9
SMO-R	5	22	6,5
kNN-R	5	22	6,5
RF-R	8	22	5
J48-R	2	21	11

Tabla 7.14. ***Resumen con los resultados de comparación de cada método y su posición en el ranking para una resolución de 1 metro***

En la Tabla 7.14 se observa que el método mejor clasificado es el SMO con la selección de atributos Greedy. Como en la resolución anterior se observan los métodos que se encuentran tras éste en el ranking, pero en este caso se observa cómo los siguientes poseen todos los atributos, por lo que no aportan una disminución del número de atributos, quedándose el tercer método (kNN-G) muy lejos del primero. Por lo tanto, para la resolución de 1 metro únicamente se tendrá en cuenta el clasificador SMO con una selección de atributos Greedy.

7.3. Resolución 2 metros

7.3.1. Atributos agregados

Los nuevos atributos añadidos para una resolución de 2 metros son los mismos que los agregados para las resoluciones de 0.5 y 1 metro.

7.3.2. Selección de atributos

- **Greedy Stepwise**

Las variables que han sido seleccionadas por el método de búsqueda Greedy han sido (Tabla 7.15):

Atributo	Descripción
AVG_B4	Media espectral de las alturas
TXAVG_B2	Media de textura del rojo
TXAVG_B3	Media de textura del verde
TXAVG_B4	Media de textura de las alturas
TXAVG_B5	Media de textura del NDVI
AVG_B2/AVG_B3	Ratio entre la media espectral del rojo y del verde
TXAVG_B2/TXAVG_B3	Ratio entre la media de textura del rojo y del verde
MIN_B2/TXAVG_B4	Ratio entre mínimo espectral del rojo y media de textura de las alturas
AVG_B3/TXAVG_B1	Ratio entre media espectral del verde y media de textura del infrarrojo cercano
TIPO	Tipo de Catastro

Tabla 7.15. ***Enumeración y descripción de la selección de atributos con Greedy Stepwise para una resolución de 2 metros***

Se puede observar que el número de atributos ha aumentado ligeramente con respecto a las resoluciones anteriores y que éstos han variado. En este caso se mantiene la media de casi todas las texturas salvo la de la banda 1. Con respecto a los atributos agregados siguen predominando las relaciones entre las bandas 1 y 2 con la 3 y la variable del Catastro.

- **Genetic Search**

En cuanto al método Genetic el número de variables es menor que en las granularidades anteriores, siendo en este caso 63, pero aún así el número sigue siendo muy elevado. Para la resolución de 2 metros los atributos que han sido eliminados son: FX_RECT_FI, FX_MAIN_DI, STD_B1 (infrarrojo cercano), MIN_B3 (banda del verde), TXENT_B5 (banda del NDVI), FX_COMPACT/TXRAN_B5 (banda del NDVI) y FX_ROUND/MIN_B5 (banda del NDVI).

- **Race Search**

Para el método de búsqueda Race ocurre lo mismo que con Greedy para esta resolución: que el número de variables seleccionadas aumenta, pero siendo un número aún muy reducido. Los 7 atributos seleccionados son (Tabla 7.16):

Atributo	Descripción
AVG_B5	Media espectral del NDVI
TXAVG_B4	Media de textura de las alturas
TXAVG_B5	Media de textura del NDVI
TXAVG_B2/TXAVG_B3	Ratio entre la media de textura del rojo y del verde
TXAVG_B1/TXAVG_B3	Ratio entre la media de textura del infrarrojo cercano y del verde
MIN_B2/TXAVG_B4	Ratio entre mínimo espectral del rojo y media de textura de las alturas
TIPO	Tipo de Catastro

Tabla 7.16. ***Enumeración y descripción de la selección de atributos con Race Search para una resolución de 2 metros***

Como se puede ver las variables son muy similares a las ya seleccionadas para otras granularidades.

7.3.3. Resultados de la generación de modelos

En las Tablas 7.17 y 7.18 se indican los resultados para cada uno de los *datasets* con cada clasificador. En rojo se indican los mejores resultados de cada *dataset*, y en rojo y subrayado en caso de haber un empate entre varios clasificadores.
Para esta resolución de 2 metros tampoco se tienen en cuenta los resultados obtenidos con el método de búsqueda Genetic Search, ya que no hace una gran reducción en el número de los atributos, por lo que éstos son muy similares a los de “Catastro”.

Como se puede ver el clasificador que obtiene un índice Kappa ponderado más elevado en un mayor número de *datasets* y en la media es el SMO con una selección de atributos con el método Greedy.

Datasets	SMO-O	kNN-O	RF-O	J48-O	SMO-E	kNN-E	RF-E	J48-E	SMO-C	kNN-C	RF-C
Dataset 1	96656,5	79824,7	94440,0	76054,7	97260,1	84930,7	94667,6	83337,5	97260,1	85425,4	95627,4
Dataset 2	93603,9	71336,3	90051,1	72867,5	92953,4	70896,0	91101,9	76640,5	98387,7	69124,6	95805,6
Dataset 3	60597,3	51719,6	57390,7	59063,2	60886,5	53285,1	54995,3	58648,2	57466,2	52725,5	59767,4
Dataset 4	62820,0	43503,1	67347,2	51531,2	67834,9	51596,7	67864,0	61837,5	57594,1	50541,3	67238,0
Dataset 5	68004,7	164429,8	29152,2	34189,6	68208,9	164821,3	160924,1	33321,6	**167680,3**	165127,6	64856,4
Dataset 6	67606,1	59741,4	61771,4	62883,1	65569,1	63041,9	63048,8	63476,9	**67792,5**	57511,1	62952,2
Dataset 7	42154,8	27115,1	42688,3	38554,9	43073,3	32115,9	43648,5	41375,4	43857,3	36351,4	38211,6
Dataset 8	46421,1	17071,1	46105,6	16735,6	46260,9	18868,9	46356,0	45644,9	46235,8	14912,8	18929,0
Dataset 9	22490,4	22681,2	20847,8	19131,0	22389,7	22869,3	21867,8	19705,9	22652,0	22392,4	20535,2
Dataset 10	187607,5	176977,7	183186,4	120041,9	188607,0	178515,5	184166,7	123328,9	188741,6	184186,0	187146,2
Dataset 11	**70860,0**	61038,3	64182,4	60847,3	69052,9	62382,7	62617,7	55139,5	68935,3	66239,3	66187,9
Dataset 12	59323,9	54806,4	57723,9	58539,6	58715,3	56971,0	57905,9	57887,1	60327,8	57334,9	58709,0
Dataset 13	67817,0	59078,8	63517,0	50158,8	70341,7	58758,7	64608,4	60068,3	70225,2	66296,4	68100,7
Dataset 14	66788,4	63755,6	58692,0	65231,4	66788,4	62652,2	64493,5	65793,2	66889,9	61569,0	63857,1
Dataset 15	44993,1	36082,1	43387,1	33567,1	44823,6	36510,7	44503,3	43815,7	44771,8	36171,6	44249,0
Dataset 16	49988,5	48447,0	46507,3	32908,4	50495,4	48519,5	49424,7	34418,8	50495,4	48535,0	48560,8
Dataset 17	50968,5	48205,4	49329,0	39732,7	53915,1	49747,5	50446,9	49759,0	54006,8	46468,5	52653,9
Dataset 18	35396,4	32604,3	37996,0	30966,7	33107,9	37885,5	38352,2	38692,0	35883,6	38823,0	37635,8
Dataset 19	52798,9	42783,1	55180,8	53130,8	59476,0	49688,1	51790,1	59189,6	**62020,6**	52688,2	54015,8
Dataset 20	52407,6	49162,6	53194,8	47089,3	54745,3	49883,7	49048,4	53417,2	55622,6	51464,1	48940,2
Media	64965,2	60518,2	61134,6	51161,2	65725,3	62697,0	68091,6	56274,9	70842,3	63194,4	62699,0

Tabla 7.17. ***Índice Kappa ponderado obtenido por cada método en cada dataset, y media de cada método para una resolución de 2 metros***

Datasets	J48-C	SMO-G	kNN-G	RF-G	J48-G	SMO-R	kNN-R	RF-R	J48-R
Dataset 1	84168,7	92520,3	96903,9	93757,2	83743,2	96102,4	92480,7	95093,1	87523,2
Dataset 2	84226,5	**99398,5**	94734,8	97917,3	93714,0	98097,4	93694,0	95865,7	94674,7
Dataset 3	59471,9	57705,1	53995,6	**60999,7**	59515,9	57799,4	54398,0	60490,4	59616,5
Dataset 4	63096,6	67783,9	68024,1	67281,7	68031,4	68249,7	**68504,5**	67427,3	68395,3
Dataset 5	19553,9	166506,1	166982,6	164651,1	97122,9	114107,0	166999,6	99199,1	20115,5
Dataset 6	63200,7	65327,4	65244,6	60466,4	64623,1	65016,7	60846,2	61654,0	64864,8
Dataset 7	43124,3	43486,2	44070,7	**44826,9**	42419,2	42957,3	43694,9	43991,8	40888,3
Dataset 8	46631,4	48739,7	17121,2	20241,0	45710,0	**48889,9**	17221,4	20261,0	20170,9
Dataset 9	19912,6	**22935,5**	22008,2	21237,3	21504,9	21239,9	22604,3	19785,4	20201,4
Dataset 10	124847,4	**189491,3**	188837,7	189318,3	188587,8	187146,2	188856,9	188011,2	188107,3
Dataset 11	54353,4	66312,8	67451,4	59062,2	57629,8	68744,3	69816,9	56682,1	59716,0
Dataset 12	59524,7	**61984,2**	55766,3	56318,5	52246,4	59844,7	58414,1	50370,4	47691,3
Dataset 13	59486,3	69788,7	69766,9	64099,1	64062,7	**70931,0**	68566,4	67860,6	59275,3
Dataset 14	66233,2	**67072,7**	64500,2	65705,2	65976,0	65542,8	63098,9	65136,6	64987,7
Dataset 15	44310,2	**45280,4**	44559,8	44362,0	42511,0	44889,5	44423,2	44103,0	43646,1
Dataset 16	34418,8	50754,0	50686,8	37263,8	38081,0	50516,1	37424,1	37739,7	**50904,0**
Dataset 17	47563,4	50613,1	50091,5	50595,9	52046,3	**54190,3**	52969,2	49799,1	49208,6
Dataset 18	39232,4	40112,6	39093,2	39195,6	32186,7	40112,6	39854,7	38708,4	38180,3
Dataset 19	59593,1	61675,7	55851,1	48100,1	52863,9	60640,9	54777,3	46082,7	52746,8
Dataset 20	54240,5	57389,4	56181,5	49817,6	57677,8	57437,4	51536,3	52545,8	**57707,9**
Media	56359,5	**71243,9**	68593,6	66760,8	64012,7	68622,8	67509,1	63040,4	59431,1

Tabla 7.18. ***Índice Kappa ponderado obtenido por cada método en cada dataset, y media de cada método para una resolución de 2 metros***

7.3.4. Test de Wilcoxon

Tras haber aplicado el test de Wilcoxon con Keel, se toma la tabla de salida de esta herramienta y se transforma en las Tabla 7.19, donde se indica el resultado de comparar el método de cada fila con el de la columna.

Seguidamente, en la Tabla 7.20 se muestran 3 columnas resumen en las que se indica el número de victorias, de no derrotas y el ranking de cada método.

Datasets	SMO-O	kNN-O	RF-O	J48-O	SMO-E	kNN-E	RF-E	J48-E	SMO-C	kNN-C	RF-C	J48-C	SMO-G	kNN-G	RF-G	J48-G	SMO-R	kNN-R	RF-R	J48-R
SMO-O	-	G	G	G	E	G	G	G	P	G	G	G	P	E	E	E	P	E	E	G
kNN-O	P	-	P	E	P	P	P	E	P	P	P	E	P	P	P	P	P	P	P	P
RF-O	P	G	-	G	P	G	E	E	P	E	P	E	P	P	E	E	P	E	E	E
J48-O	P	E	P	-	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
SMO-E	E	G	G	G	-	G	G	G	P	G	G	G	E	E	G	G	E	E	G	G
kNN-E	P	G	P	G	P	-	P	E	P	E	P	E	P	P	P	E	P	P	E	P
RF-E	P	G	E	G	P	G	-	G	P	G	E	G	P	P	E	E	P	E	E	E
J48-E	P	E	E	G	P	E	P	-	P	E	E	P	P	P	E	P	P	P	E	E
SMO-C	G	G	G	G	G	G	G	G	-	G	G	G	E	G	G	G	E	G	G	G
kNN-C	P	G	E	G	P	E	P	E	P	-	P	E	P	P	E	E	P	P	E	E
RF-C	P	G	G	G	P	G	E	E	P	G	-	E	P	E	E	E	P	E	E	E
J48-C	P	E	E	G	P	E	P	G	P	E	E	-	P	P	E	E	P	E	E	E
SMO-G	G	G	G	G	E	G	G	G	E	G	G	G	-	G	G	G	E	G	G	G
kNN-G	E	G	G	G	E	G	G	G	P	G	E	G	P	-	E	G	P	E	G	G
RF-G	E	G	E	G	P	G	E	E	P	E	E	E	P	E	-	E	P	E	E	E
J48-G	E	G	E	G	P	E	E	G	P	E	E	E	P	P	E	-	P	E	E	E
SMO-R	G	G	G	G	E	G	G	G	E	G	G	G	E	G	G	G	-	G	G	G
kNN-R	E	G	E	G	E	G	E	G	P	G	E	E	P	E	E	E	P	-	E	E
RF-R	E	G	E	G	P	E	E	E	P	E	E	E	P	P	E	E	P	E	-	E
J48-R	P	G	E	G	P	G	E	E	P	E	E	E	P	P	E	E	P	E	E	-

Tabla 7.19. ***Resultados de comparación entre los diferentes métodos tras aplicar el test de Wilcoxon para una resolución de 2 metros***

Datasets	Victorias	Victoria/Empate	Ranking
SMO-O	12	19	6
kNN-O	0	4	19
RF-O	3	14	14
J48-O	0	1	20
SMO-E	15	22	4
kNN-E	2	9	18
RF-E	8	15	12
J48-E	1	11	17
SMO-C	**20**	**23**	**1**
kNN-C	2	12	16
RF-C	6	17	9
J48-C	2	14	15
SMO-G	**19**	**23**	**2**
kNN-G	13	19	5
RF-G	3	18	8
J48-G	3	16	11
SMO-R	**18**	**23**	**3**
kNN-R	7	19	7
RF-R	2	17	10
J48-R	3	15	13

Tabla 7.20. ***Resumen con los resultados de comparación de cada método y su posición en el ranking para una resolución de 2 metros***

En la Tabla 7.20 se observa que el método que aparece en el número uno del ranking es el SMO con todos los atributos, incluido el del tipo de Catastro. Para esta resolución espacial de 2 metros ocurre como en la de 0,5 metros, en la cual hay unos métodos como el SMO con selección Greedy y Race que se encuentran cerca del primer clasificado y ofrecen una reducción de variables muy importante, por lo que también se tomarán para el siguiente capítulo.

8. Estudio de Adaptabilidad del Modelo

Esta nueva zona sobre la que probar los modelos obtenidos en el apartado anterior comprende parte de los términos municipales de Valencia, Tavernes Blanques, Alboraya, Bonrepós i Mirambell, y Almàssera; los cuales forman parte de la comarca de l'Horta Nord.

En la Figura 8.1 se delimita la zona de estudio que es mayoritariamente zona de huertas y urbana, incluyendo alguna zona industrial y costera.

Figura 8.1. ***Delimitación de la zona de estudio B***

Como datos se dispone de ortofotografías de alta resolución con una resolución espacial de 0,5 metros y de un modelo normalizado. Este modelo corresponde con la altura de los objetos presentes sobre el terreno, por lo que ya no será necesario efectuar el proceso para su obtención como se hizo en el Capítulo 5.

- **Tratamiento de los datos:**

En la Figura 8.2 se muestra un diagrama de flujo que indica los procesos efectuados para el desarrollo de este capítulo, en el que se explica la aplicación del modelo.
El tratamiento que precisan los datos para esta nueva zona de trabajo es muy similar que el realizado en el Capítulo 5, incluso conlleva menos trabajo ya que ya se facilita el modelo normalizado y no es necesario obtenerlo a partir de los datos LiDAR.

El primer paso es efectuar un recorte de la zona de estudio, de manera que no existan datos sobrantes y se necesite más memoria y los procesos sean más costosos.

El siguiente paso es efectuar la segmentación como en el Capítulo 5. Como se van a comparar los modelos generados con los modelos obtenidos por los clasificadores kNN y SVM de ENVI, es necesario usar parte de los segmentos para entrenar los modelos de ENVI y otros como test para comparar todos los modelos entre sí.
En este caso los valores introducidos tanto en el factor de escala como en la agrupación (merge) para las diferentes resoluciones espaciales fueron los siguientes:

- **0,5m:** Factor de escala 20, Merge 55.
- **1m:** Factor de escala 20, Merge 55.
- **2m:** Factor de escala 30, Merge 55.

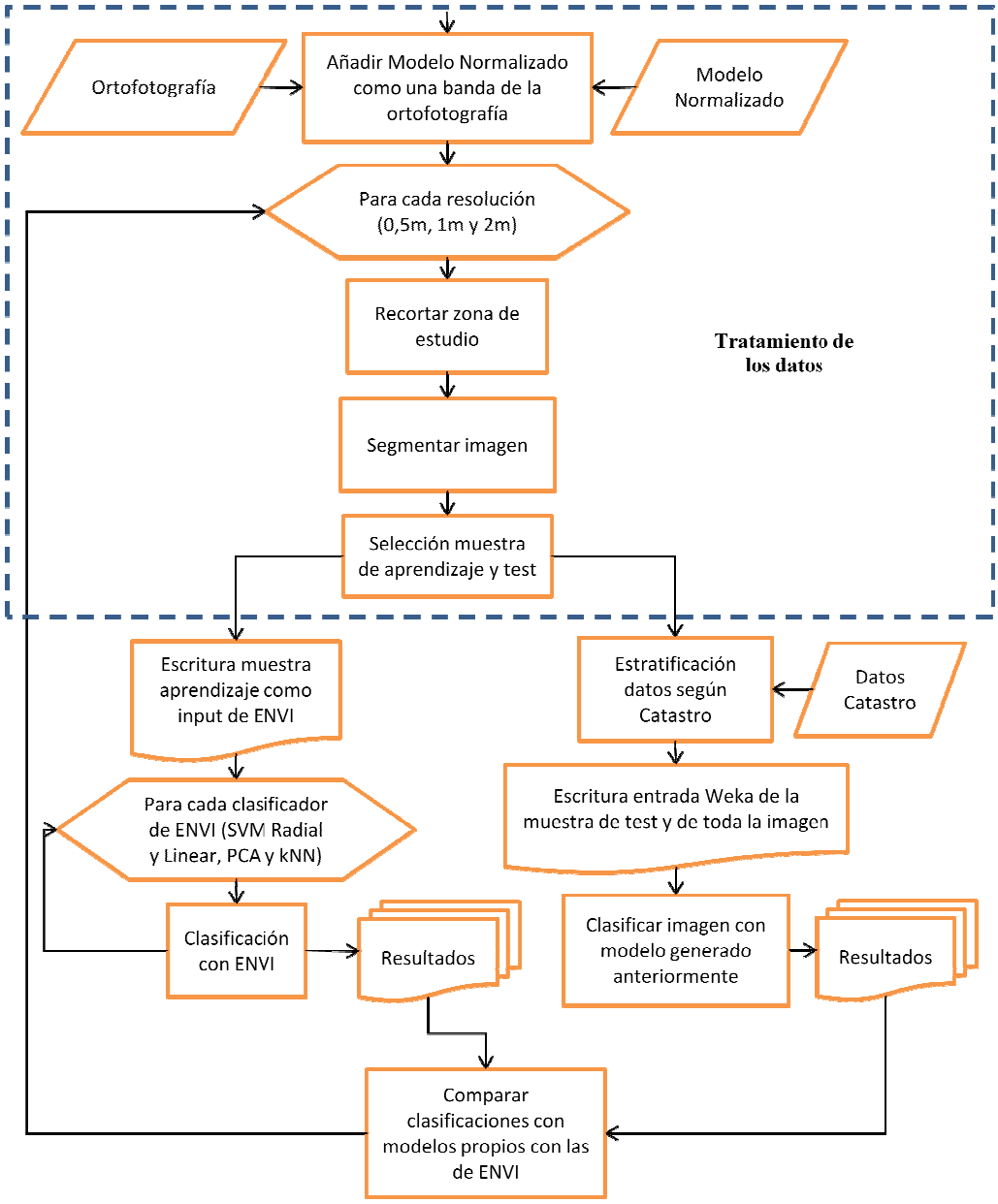


Figura 8.2. ***Detalle de las diferentes etapas realizadas durante la fase de aplicación de los modelos***

Tras obtener los segmentos ya se pueden tomar las muestras. Cabe recordar que la muestra de aprendizaje únicamente será utilizada para realizar las clasificaciones con el software ENVI FX5, ya que nuestro modelo ya ha sido generado con los datos de aprendizaje del Capítulo 5. Sin embargo, la muestra de test sí que será utilizada para comparar las clasificaciones realizadas con ENVI con la efectuada por nuestro modelo. Como se explicó en el Capítulo 5, las muestras de aprendizaje y test se toman con ArcGIS y el valor de cada objeto se modifica en la tabla. Para poder introducir estos datos como de entrada para ENVI es necesario transformarlos a una capa de tipo punto en lugar de tipo polígono y realizar unas modificaciones en el orden y el nombre de

los atributos de la tabla. Todo este proceso se lleva a cabo por medio de unas líneas de código escritas en Python. De esta manera la muestra de aprendizaje ya puede ser introducida como dato de entrada en ENVI y el software puede efectuar la clasificación, sin tener que realizar la selección de los segmentos manualmente para cada uno de los clasificadores.

Para la estratificación de los segmentos según los datos del Catastro y la escritura del fichero de entrada de Weka con la muestra de test y el fichero con todos los segmentos (necesarios para la clasificación) se siguen los mismos pasos que en 5.5 y en 5.6.

- **Muestra de aprendizaje y test entre diferentes resoluciones espaciales:**

Sin lugar a duda uno de los trabajos más costosos es la selección de los datos de entrenamiento y de test, y éste se incrementa cuando se desea trabajar con diferentes resoluciones espaciales sobre una misma zona, que es el caso de este proyecto. El problema es que la segmentación de la imagen cambia según la variación de su granularidad, por lo que los segmentos no son iguales. Por lo tanto, puede suceder que a una resolución espacial de 2 metros, un objeto etiquetado como edificación, pueda corresponder con varios objetos etiquetados como edificación y vía en una resolución de 0,5 metros.

Para ver si este proceso puede reducirse, se va a analizar en este punto si las muestras utilizadas con una resolución de 1 y 2 metros pueden ser obtenidas a partir de las introducidas en 0,5 metros. De esta manera únicamente sería necesario realizar la selección en la resolución más alta y obtener automáticamente las muestras para las resoluciones más bajas.
Para llevar a cabo este estudio se verificará por qué segmentos de mayor resolución está compuesto el de una menor resolución, y ver a partir del área cuál es la clase mayoritaria, siendo ésta la nueva etiqueta del segmento de menor resolución.
Por ejemplo, en la Figura 8.3 se observa con línea de color naranja un segmento extraído de la ortofotografía de resolución espacial 2 metros, en azul se representan segmentos de la imagen de 0,5 metros de resolución etiquetados como agua y en rojo segmentos etiquetados como suelo.
Si cruzamos esa información etiquetada con el segmento extraído de la resolución de 2 metros se observa cómo está compuesto por tres clases: en azul agua, en rojo suelo y en blanco sin clasificar. Para asignar una clase al segmento marcado con línea naranja hay que buscar cuál de esas tres clases por las que está compuesto es mayoritaria según el área, y se observa claramente como la clase mayoritaria es suelo (rojo), por lo que este segmento sería etiquetado como suelo.

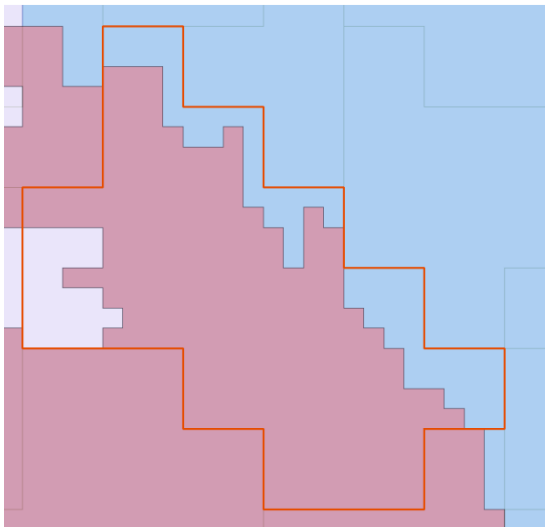


Figura 8.3. ***Comparación entre los segmentos de resolución 0,5 y 2 metros***

Este estudio se puede realizar porque se han tomado las muestras de aprendizaje para una resolución de 0,5, 1 y 2 metros, así que los resultados obtenidos con este método se pueden comparar con los etiquetados realizados manualmente.

Como resultado se obtiene que al convertir la muestra de aprendizaje seleccionada en una resolución de 0,5 metros a una de 1 metro hay 9132 (99,97%) instancias que mantienen la misma etiqueta y 3 (0,03%) que tienen un etiquetado diferente.

Si observamos la conversión de 0,5 metros a 2 metros se ve que 3041 (99,97%) instancias tienen la misma clase y 1 (0,03%) tiene una clase diferente. Lo que supone el mismo porcentaje que en el párrafo anterior.

Por lo tanto se puede decir que si las muestras seleccionadas en una alta resolución son transformadas a una resolución menor teniendo en cuenta la clase mayoritaria, el error cometido es mínimo, por lo que este método se puede utilizar cuando trabajamos con una misma zona pero con diferentes granularidades.
Así pues, en esta Zona B de trabajo se utilizará este método para a partir de las muestras de aprendizaje y de test seleccionadas en 0,5 metros, transformarlas a las muestras de las resoluciones de 1 y 2 metros, ahorrándose así tiempo de trabajo.

- **Clasificación con ENVI:**

El objetivo de esta nueva zona de trabajo es clasificar la ortofotografía a partir de los modelos propios generados en la zona anterior y poder analizar esos resultados comparándolos con la clasificación realizada por un software comercial como es ENVI.

Hay que tener en cuenta que la clasificación realizada por ENVI sí que cuenta con una muestra de entrenamiento tomada en esta Zona B, mientras que la clasificación realizada por los modelos propios no.
La finalidad es ver si disponiendo ya de un modelo obtenido en una zona de características similares, éste se puede emplear para clasificar una nueva zona y si los resultados son buenos y se asemejan a los obtenidos por ENVI, o si por el contrario los resultados no son aceptables y se deduce que es necesario actualizar el modelo con datos de esta zona.

Como ya se explicó en el Capítulo 5, el fichero de entrada de ENVI con el conjunto de entrenamiento es generado a partir de un fichero shapefile de ArcGIS.
Este proceso consiste únicamente en generar la segmentación como ya se hizo: cargar el fichero con los datos de entrenamiento, seleccionar un clasificador y realizar la clasificación.
ENVI dispone de tres clasificadores (SVM y kNN). En este caso se utiliza el SVM con una función *kernel* radial y otra lineal, y kNN con k=5. Así pues el proceso se repetirá para cada resolución y cada uno de los cuatro clasificadores utilizados.
Tras este proceso ya se dispondrá de la imagen clasificada al completo, y ésta se puede comparar con la muestra de test seleccionada para conocer la precisión obtenida en la clasificación, y poder comparar los resultados con los otros clasificadores.

- **Clasificación con los modelos propios:**

En la Zona A ya se hicieron las pruebas para poder concluir qué atributos y clasificadores obtenían una mejor clasificación.
Tras haber obtenido esos modelos es el momento de lanzar toda la imagen de la Zona B para que a través de Weka efectúe la clasificación.
El resultado de la clasificación es como el que se observa en la Figura 8.4, que es el fichero de salida de una clasificación efectuada por Weka, donde se indica qué clase se ha asignado a cada instancia y con qué precisión.

Con la clasificación de toda la imagen realizada es el momento, como se hizo con la clasificación de ENVI, de comparar los resultados con la muestra de test para poder realizar una comparación con la clasificación del software comercial.

inst#	actual	predicted	error	prediction
1	1:?	1:Bosque		0.998
2	1:?	1:Bosque		0.998
3	1:?	1:Bosque		0.998
4	1:?	1:Bosque		0.998
5	1:?	1:Bosque		0.998
6	1:?	1:Bosque		0.998
7	1:?	1:Bosque		0.998
8	1:?	1:Bosque		0.998
9	1:?	1:Bosque		0.998
10	1:?	1:Bosque		0.998
11	1:?	1:Bosque		0.998
12	1:?	1:Bosque		0.998
13	1:?	1:Bosque		0.998
14	1:?	1:Bosque		0.998
15	1:?	1:Bosque		0.998

Figura 8.4. ***Fichero de salida de Weka tras ser clasificado, donde se indica la clase predicha y la probabilidad de acierto en la clasificación***

En el capítulo anterior se han extraído los clasificadores que efectuaban una mejor clasificación para cada una de las tres resoluciones estudiadas. Estos clasificadores son: Random Forest Catastro, SMO Race y Random Forest Race para 0,5 metros; SMO Greedy para 1 metro; SMO Catastro, SMO Greedy y SMO Race para 2 metros.
Una vez generados estos modelos es el momento de utilizarlos para clasificar la nueva zona de trabajo B y comparar los resultados obtenidos con la clasificación realizada por ENVI.
Los clasificadores utilizados en ENVI serán: SVM Radial y Linear y kNN.

8.1. Resolución 0,5 metros

En la Tabla 8.1 se muestra la fiabilidad global y el índice Kappa obtenido tras clasificar la zona B con los modelos generados y con ENVI. Los resultados obtenidos por los modelos propios son muy bajos, quedándose el más alto de ellos a unos 34 puntos porcentuales de la mejor clasificación efectuada por ENVI con el SVM – Linear. Esto es debido a que las clasificaciones realizadas con ENVI sí que sus modelos hacían un aprendizaje con un conjunto de entrenamiento en esta zona de estudio B. Esto demuestra que un modelo generado en una zona A no se puede utilizar en una zona similar B, por lo que se debe realizar un aprendizaje para cada zona nueva o actualizar el modelo que se tiene. Fundamentalmente la diferencia en los resultados se debe a las diferencias radiométricas entre las imágenes utilizadas en una y otra zona, ya que incluso tomando una misma zona en momentos diferentes pueden existir estas diferencias radiométricas debidas, por ejemplo, a la dispersión atmosférica o las diferencias entre los sensores con los que las imágenes han sido capturadas. Sin embargo, aún realizando las correcciones radiométricas, dada su complejidad, los modelos diferirán entre una y otra zona.

	Clasificador	Fiabilidad Global	Índice Kappa
Modelos Generados	Random Forest – Catastro	62,7120%	0,5462
	SMO – Race	54,3990%	0,4562
	Random Forest – Race	49,7812%	0,4186
ENVI	SVM – Radial	94,6790%	0,9345
	SVM – Linear	96,2772%	0,954
	kNN	73,0010%	0,6726

Tabla 8.1. ***Comparación de resultados entre los modelos generados y las clasificaciones efectuadas por ENVI para una resolución de 0,5 metros***

Lo que también se observa en la Tabla 8.1 es que los modelos propios generados mantienen el mismo orden que el obtenido en el capítulo anterior tras aplicar el test estadístico.

Fiabilidad Usuario			
Clase	**Random Forest - Catastro**	**SMO - Race**	**SVM - Linear**
Vía	7,4362%	53,6470%	95,4737%
Edificación	97,6526%	97,1620%	94,5425%
Suelo	81,3915%	22,4792%	95,6416%
Vegetación	86,7116%	96,8700%	99,8898%
Agua	98,4167%	98,0477%	98,4850%
Playa	0,0000%	0,1383%	95,1168%

Tabla 8.2. ***Comparación de resultados de la fiabilidad de usuario de cada clase entre modelos generados y la clasificación con SVM – Linear realizada con ENVI para una resolución de 0,5 metros***

En la Tabla 8.2, donde se muestra la fiabilidad del usuario, que es la relación entre las instancias clasificadas como una clase X y el número de instancias etiquetadas como esa clase X, muestra claramente que los mayores errores en la clasificación de los modelos propios se producen en la clase Vía y en Playa.

A continuación se intentan extraer algunas explicaciones, aparte de las diferencias radiométricas, haciendo un análisis más concreto, a nivel de clases.

La mala clasificación de Vía es debida a una confusión entre esta clase y la de Agua, lo cual puede ser debido a una elevada presencia de sombras sobre las vías de las edificaciones, al ser éstas más elevadas para la zona B, y a las diferentes características de vías presentes en las dos zonas.

En cuanto a la pésima clasificación de Playa, ésta puede estar condicionada por la diferencia de fechas en las que se tomaron las imágenes de las dos zonas. Mientras la zona A fue tomada en una época estival, por la presencia de sombrillas en la playa; la zona B fue tomada en una época más húmeda y fría, ya que no aparecen sombrillas y

el arena tiene un aspecto más oscuro debido a la humedad y la línea de costa es más difícil diferenciarla, lo que condiciona la clasificación.
Con el SMO – Race también se observa que el resultado de la clase Suelo es muy bajo, al tener muchas similitudes con la clase Playa.
Sin embargo, otras clases más sencillas de clasificar ya que no varían tanto de una zona a otra, como Edificación, Agua o Vegetación (para SMO – Race), han igualado o mejorado los resultados.

8.2. Resolución 1 metro

En la Tabla 8.3 se observa la precisión y el índice Kappa obtenido tras la clasificación con el modelo generado y con ENVI. Los resultados son muy similares a los vistos anteriormente. La precisión obtenida por el modelo propio sigue siendo muy baja con respecto a los resultados más alto obtenido con la clasificación efectuada con ENVI, debido también a las diferencias radiométricas entre ambas imágenes.

	Clasificador	Fiabilidad Global	Índice Kappa
Modelos Generados	SMO - Greedy	60,3356%	0,5166
ENVI	SVM – Radial	96,2046%	0,9531
	SVM – Linear	96,0760%	0,9516
	kNN	71,5405%	0,6546

Tabla 8.3. ***Comparación de resultados entre los modelos generados y las clasificaciones efectuadas por ENVI para una resolución de 1 metro***

En la Tabla 8.4, se muestra la fiabilidad del usuario para cada clase del modelo generado y del mejor clasificador de ENVI. Se puede apreciar cómo la clase Vía y Suelo vuelven a obtener una fiabilidad de usuario muy escasa. Ambas clases se confunden en este caso según la matriz de confusión con la clase Playa. La confusión de la clase Vía con Playa puede ser debido nuevamente a la presencia de sombras y, a como se comentó en el apartado anterior, a que la arena está húmeda y por lo tanto se ve más oscura. En cuanto a la clase Suelo, tiene nuevamente que ver con la similitud en todas las bandas entre ambas clases.
La gran diferencia con la resolución anterior es que en ésta se consigue un muy buen resultado en la clase Playa, únicamente existiendo algunas confusiones con Agua, es decir, delimitando la línea de costa.

Fiabilidad Usuario		
Clase	**SMO - Greedy**	**SVM - Radial**
Vía	1,5116%	97,0017%
Edificación	98,0414%	96,0013%
Suelo	12,8513%	93,5106%
Vegetación	86,0696%	99,8753%
Agua	84,7085%	97,7276%
Playa	95,8681%	95,2697%

Tabla 8.4. ***Comparación de resultados de la fiabilidad de usuario de cada clase entre el modelo generado y la clasificación con SVM – Radial realizada con ENVI para una resolución de 1 metro***

8.3. Resolución 2 metros

En la Tabla 8.5 se muestran la fiabilidad global y el índice Kappa obtenidos por los modelos generados y con ENVI. Se observa que el SMO – Catastro consigue un resultado más aceptable, aunque sigue siendo bajo con respecto a la clasificación de ENVI y con los resultados obtenidos a la hora de generar los modelos en el Capítulo 7, debido igualmente a las diferencias radiométricas entre las imágenes de ambas zonas.

	Clasificador	**Fiabilidad Global**	**Índice Kappa**
Modelos Generados	SMO – Catastro	80,9437%	0,7644
	SMO - Greedy	61,7521%	0,5428
	SMO – Race	53,1489%	0,4511
ENVI	SVM – Radial	93,2187%	0,9165
	SVM – Linear	97,0154%	0,9629
	kNN	75,7979%	0,7049

Tabla 8.5. ***Comparación de resultados entre los modelos generados y las clasificaciones efectuadas por ENVI para una resolución de 2 metros***

En la Tabla 8.6 se compara la fiabilidad de usuario de cada clase para dos modelos generados y el mejor clasificador de ENVI. Los resultados del método SMO – Catastro mejoran en general salvo en la clase Vía que sigue teniendo los mismos problemas que en la resoluciones anteriores, confundiéndose en este caso principalmente con la clase Agua.

El método SMO – Greedy aparte de tener malos resultados en la clase Vía, también los tiene en Suelo, pero principalmente en Playa. En estos dos casos la confusión se produce con la clase Vía, probablemente debido a las sombras de las vías y la oscuridad de la playa y de algunos suelos.

Fiabilidad Usuario			
Clase	**SMO - Catastro**	**SMO - Greedy**	**SVM - Linear**
Vía	3,7417%	11,5083%	98,2658%
Edificación	95,4705%	96,8606%	96,3376%
Suelo	92,4513%	67,3914%	96,6370%
Vegetación	81,6234%	98,5982%	99,8105%
Agua	98,7818%	98,8444%	99,7246%
Playa	89,7603%	0,0167%	93,7412%

Tabla 8.6. ***Comparación de resultados de la fiabilidad de usuario de cada clase entre modelos generados y la clasificación con SVM – Linear realizada con ENVI para una resolución de 2 metros***

9. Conclusiones y Trabajos Futuros

En este TFM se ha realizado un estudio sobre las posibilidades que nos ofrece la minería de datos para la clasificación de imágenes áreas o satélite en una zona de estudio urbana y periurbana costera con elementos geográficos típicos de un clima mediterráneo.
Para ello ha sido necesario realizar un preprocesado de los datos cartográficos, como son las ortofotografías y los datos LiDAR, eliminando datos no válidos (erróneos), transformando el formato, rango, etc. de algunos datos con el objetivo de generar una vista minable de calidad a partir de la cual entrenar los modelos.
Tras disponer de la vista minable lo que se ha hecho ha sido realizar diferentes pruebas añadiendo y eliminando atributos y probando con diferentes clasificadores, con la finalidad de efectuar un estudio acerca de qué clasificadores y atributos pueden aumentar la precisión a la hora de clasificar una imagen con estas características.
Con los resultados obtenidos para cada clasificador, lo que se ha hecho ha sido aplicar el test de Wilcoxon para poder hacer un ranking de los clasificadores para cada una de las resoluciones espaciales con las que se ha trabajado. De este modo se ha extraído el mejor clasificador y se ha analizado su proximidad con los siguientes clasificados, y si éstos estaban próximos al primero y ofrecían algunas ventajas como una reducción considerable de los atributos, entonces también se tomaban para el último paso: clasificar una nueva zona con los modelos generados en una zona diferente.
Por último, se han aplicado los modelos generados y seleccionados en el punto anterior para clasificar una nueva zona de estudio, y posteriormente se han comparado con la clasificación realizada por un software comercial como es ENVI, pero en este caso generando sus modelos con un conjunto de aprendizaje de esta segunda zona.

En el Capítulo 7, se ha extraído que para la zona A hay una serie de atributos, como pueden ser el ratio entre diferentes atributos originales y la información sobre en qué tipo de Catastro se encuentra el segmento, que ayudan a mejorar la precisión de la clasificación. También se ha conseguido una selección de atributos con el objeto de reducir el ruido, el número de atributos y por lo tanto disminuir el tiempo de cálculo.
Según la resolución se ha obtenido un conjunto de variables u otro, lo que puede parecer curioso ya que se intentan clasificar las mismas clases, pero al cambiar la granularidad y por lo tanto los segmentos, éstos ya no contienen los mismos valores, y es por eso que esta selección varía.
Los métodos que mejor resultado han obtenido para una resolución de 0,5 metros son:

- Random Forest – Catastro: que contiene los atributos originales generados por ENVI y se le ha añadido los siguientes: AVG_ROJO/AVG_VERDE, FX_COMPACT/TXRAN_NDVI, AVG_ROJO/TXAVG_VERDE, FX_CONVEX/TXAVG_NDVI, FX_ELONG/MAX_NDVI, AVG_ROJO/MIN_ALTURA, FX_ROUND/MIN_NDVI, FX_COMPACT/MIN_ALTURA, TXAVG_ROJO/TXAVG_VERDE, AVG_INFRARROJO/TXAVG_VERDE, TXAVG_INFRARROJO/TXAVG_VERDE, MIN_ROJO/TXAVG_ALTURA, AVG_INFRARROJO/AVG_VERDE, AVG_VERDE/TXAVG_INFRARROJO, MAX_INFRARROJO/TXAVG_NDVI y TIPO con la información del Catastro.
- SMO – Race y Random Forest: en los cuales la selección de atributos ha sido: TXAVG_ALTURA, TXAVG_NDVI, TXAVG_ROJO/TXAVG_VERDE, TXAVG_INFRARROJO/TXAVG_VERDE, MIN_ROJO/TXAVG_ALTURA y TIPO.

El método recomendado para una resolución de 1 metro es:

- SMO – Greedy: donde la selección de atributos es: AVG_ALTURA, AVG_NDVI, TXAVG_ALTURA, TXAVG_NDVI, AVG_ROJO/AVG_VERDE y TIPO.

Por último, para una granularidad de 2 metros los métodos recomendados son:

- SMO – Catastro: que contiene los mismos atributos que el Random Forest – Catastro seleccionado para la resolución de 0,5 metros.
- SMO – Greedy: cuya selección de atributos ha sido: AVG_ALTURA, TXAVG_ROJO, TXAVG_VERDE, TXAVG_ALTURA, TXAVG_NDVI, AVG_ROJO/AVG_VERDE, TXAVG_ROJO/TXAVG_VERDE, MIN_ROJO/TXAVG_ALTURA, AVG_VERDE/TXAVG_INFRARROJO y TIPO.
- SMO – Race: ha realizado la siguiente selección de atributos: AVG_NDVI, TXAVG_ALTURA, TXAVG_NDVI, TXAVG_ROJO/TXAVG_VERDE, TXAVG_INFRARROJO/TXAVG_VERDE, MIN_ROJO/TXAVG_ALTURA y TIPO.

Como se puede observar el clasificador que más veces se repite como mejor método, salvo en un caso en 0,5 metros, es el SMO, por lo que se puede considerar que es el que realiza una mejor clasificación para este tipo de zonas a nivel general.

De los métodos expuestos todos aportan algo nuevo, ya que se añaden nuevos atributos y en algunos casos se hace una selección de éstos. En la Tabla 9.1 se puede observar la fiabilidad global y el índice kappa obtenidos por cada método a la hora de

generar estos modelos. Como los resultados obtenidos son muy buenos y muy parecidos, o incluso mejores, entonces siempre se recomienda trabajar con el menor número posible de atributos, lo que permite cálculos más rápidos y trabajar con áreas más extensas. Por lo que se recomienda no trabajar con los métodos "Catastro" si la cantidad de datos es muy elevada.

Resolución	Método	Fiabilidad Global	Índice Kappa
0,5m	**RF – Catastro**	88,1734%	0,8475
	SMO – Race	96,5165%	0,9555
	RF – Race	95,1226%	0,9378
1m	**SMO – Greedy**	96,8299%	0,9594
2m	**SMO – Catastro**	97,0912%	0,9633
	SMO – Greedy	97,4376%	0,9677
	SMO - Race	95,2561%	0,9403

Tabla 9.1. ***Comparación de la fiabilidad global y del índice Kappa entre los modelos generados para cada resolución***

Por último, se ha observado cómo estos modelos generados en una zona A no sirven para clasificar una zona B, aunque ésta tenga unas características similares. La mala clasificación es debida a las diferencias radiométricas existentes entre dos imágenes diferentes, debida a factores como la dispersión atmosférica o la diferencia entre sensores. El problema es que aún llevando a cabo correcciones radiométricas para poder equiparar las dos zonas, los modelos diferirán entre una y otra zona. A nivel de cada clase pueden aparecer otros factores como el simple hecho de que las dos imágenes fueron tomadas en estaciones diferentes, y por lo tanto los suelos se encuentran más oscuros al estar más húmedos; o a una mayor presencia de edificaciones altas, lo que hace que existan muchas sombras sobre otras edificaciones, vías o suelos. Las clases que mejores resultados han obtenido han sido las más parecidas entre las diferentes zonas como son las edificaciones, el agua y la vegetación. Por lo que si nos interesa es únicamente clasificar estas clases, entonces estos modelos sí se podrían utilizar para clasificar una zona diferente de la de entrenamiento.
Sin embargo, si lo que se busca es una buena clasificación global, entonces el utilizar un modelo aprendido en una zona A para clasificar una zona B no consigue buenos resultados, por lo que se debe generar un nuevo modelo o actualizar el modelo que se tenía con un conjunto de entrenamiento de la zona B, lo que podría mejorar la clasificación pero también aumentar el rango de cada clase al ser un modelo más general.

Por lo tanto, en este proyecto se ha demostrado que el agregar nuevos atributos y realizar una selección de los mismos disminuye la carga de memoria y mejora los resultados; el clasificador SMO suele ser el que mejor se comporta para este tipo de imágenes; y los modelos obtenidos en una zona A no pueden utilizarse para clasificar

una zona B a no ser que estos modelos sean actualizados con un conjunto de datos de entrenamiento de la segunda zona o nos interese únicamente diferenciar unas ciertas clases como edificación, agua o vegetación..

En cuanto a los trabajos futuros, hay algunos puntos relacionados con este TFM que sería interesante tenerlos en cuenta para próximos proyectos.

Como se ha podido observar en este trabajo al tratarse de imágenes aéreas o satélite tomadas durante el día, a veces hay problemas con las sombras generadas, sobre todo por las edificaciones, que generan las sombras más grandes. Sería interesante poder tratar estas sombras con el objetivo de mejorar la clasificación, ya que una vía en sombra no se diferencia de un solar en sombra.

En lo relacionado con la agregación de atributos y la potencialidad que nos pueden ofrecer los GIS, en un trabajo futuro se podría añadir un atributo en el que se indicasen las clases colindantes de cada segmento, de esta manera el modelo sería capaz de aprender que un objeto de la clase playa puede colindar con otro de la misma clase o con uno del tipo agua.

En este proyecto se ha clasificado una imagen teniendo en cuenta imágenes generales como son Vía, Edificación, Suelo, Vegetación, Agua y Playa. Un posible trabajo sería obtener modelos a diferentes niveles para conseguir clasificar objetos más específicos. Por ejemplo, a partir de los objetos clasificados como Edificación se generaría un nuevo modelo para diferenciar los diferentes tipos de edificaciones como podrían ser de tipo industrial, adosados, edificaciones históricas, etc. De este modo a este tipo de clasificaciones se le podría dar un gran uso relacionado con el Catastro y detección de tipos de edificaciones o con temas urbanísticos.

Durante el desarrollo de este trabajo se ha observado que dependiendo de la clase había un clasificador que obtenía mejores resultados que otro. Teniendo esto en cuenta sería interesante realizar una clasificación con diferentes modelos según la clase. Por ejemplo, se obtendrían diferentes modelos con diferentes clasificadores y se estudiaría cuál se comporta mejor según qué clase. A continuación estos modelos serían utilizados como modelos binarios, y se clasificaría la clase que mejor precisión hubiese obtenido con el clasificador correspondiente y se descartaría el resto de registros. Seguidamente se descartarían las instancias ya clasificadas y se pasaría a la clase con el siguiente mejor resultado, y ésta se clasificaría con el clasificador correspondiente. Así sucesivamente hasta haber clasificado todas las clases. De este modo se clasificaría cada clase con el mejor clasificador correspondiente, pudiendo mejorar el resultado.

Por último, en este TFM se ha realizado un estudio con tres resoluciones espaciales diferentes, transcurriendo de manera paralela. Sería interesante poder desarrollar un trabajo en el que se comparasen los resultados de las diferentes granularidades,

observando si hay una resolución que consigue mejorar la precisión tanto a nivel general como para cada clase. Esto es importante ya que a veces cuando se trabaja con imágenes de alta resolución éstas nos facilitan más información de la que es necesaria. Por ejemplo, es muy probable que en una vía aparezca algún banco, coche, o que en una playa aparezca alguna sombrilla, añadiendo así ruido. Es posible que disminuyendo la resolución estos detalles se homogenicen con el resto, facilitando así la clasificación.

Referencias Bibliográficas

[1] José Hernández Orallo; Mª José Ramírez Quintana; César Ferri Ramírez. "Introducción a la Minería de Datos". *Pearson Educación, S.A.*, 2004.

[2] C.M. Di Bella et al. "La teledetección como herramienta para la prevención, seguimiento y evaluación de incendios e inundaciones". *Ecosistemas*, 17 (3): 39-52, 2008.

[3] P. Cortez et al. "A data mining approach to predict forest fires using meteorological data".

[4] M. A. Lefsky et al. "LiDAR remote sensing for ecosystem studies". *Bioscience; Jan 2002; Vol. 52, No. 1; pg.19.*

[5] Jorge García Gutiérrez. "Remote Mining: Aplicando minería de datos a teledetección sobre LiDAR". Proyecto de Tesis para el grado de Ph.D. en Ingeniería Informática, Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla, 2008.

[6] Briem, G.J; Benediktsson, J.A.; Sveinsson, J.R. "Multiple classifiers applied to multisource remote sensing data". *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40, No. 10, October 2002.

[7] Rajesh Kumar Singh. "Pattern recognition in remote-sensing imagery using data mining and statistical techniques". Thesis (Ph.D.), Purdue University, 2006.

[8] R. Arbiol; Y. Zhang; V. Palà. "Advanced classification techniques: a review". Commission VII, WG VII/4.

[9] J. Martín; F. Cánovas; F. Alonso; F.J. Gomariz; J. Moreno. "Clasificación de coberturas del suelo en la demarcación hidrográfica del segura mediante técnicas de minería de datos". XV Congreso Nacional de Tecnologías de la Información Geográfica, Madrid, AGE-CSIC, 19-21 de Septiembre de 2012.

[10] Soliman, O.S.; Mahmoud, A.S. "A classification system for remote sensing satellite images using support vector machine with non-linear kernel functions". *Informatics and Systems (INFOS)*, 2012.

[11] J.R. Otukei; T. Blaschke. "Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms". *International Journal of Applied Earth Observation and Geoinformation*, Vol. 12, Supplement 1, February 2012, Pages S27-S31.

[12] Ignacio Piqueras-Salazar; Pedro García-Sevilla. "Fusion of multi-temporal and multi-sensor hyperspectral data for land-use classification". *Pattern Recognition and Image Analysis Lecture Notes in Computer Science*, Vol. 7887, 2013, pp 724-731.

[13] González Rojas, J.C.; Pérez Cutillas, P; Palazón Ferrando, J.A. "Aplicación de técnicas de clasificación basadas en un sistema de aprendizaje para generación de un mapa de usos del suelo". *El acceso a la información espacial y las nuevas tecnologías geográficas*, pág 1575-1582.

[14] Deren LI; Kaichang DI; Deyi LI. "Land use classification of remote sensing image with GIS data based on spatial data mining techniques". *International Archives of Photogrammetry and Remote Sensing.* Vol. XXXIII, Part B3. Amsterdam 2000.

[15] Jon Atli Benediktsson; Martino Pesaresi; Kolbeinn Arnason. “Classification and feature extraction for remote sensing images from urban areas based on morphological transformations”. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 41, No. 9, September 2003.
[16] Sebastiano B. Serpico; Lorenzo Bruzzone. “A new search algorithm for feature selection in hyperspectral remote sensing images”. Technical Report #DIT-02-0027, 2001.
[17] Baofeng Guo; R.I. Damper; Steve R. Gunn; J.D.B. Nelson. “A fast separability-based feature selection method for high-dimensional remotely-sensed image classification”. *Journal Pattern Recognition*, Vol. 41, Issue 5, May 2008, Pages 1670-1679.
[18] P.H. Swain; R.C. King. “Two effective feature selection criteria for multispectral remote sensing”. *LARS Technical Reports*. Paper 39, 1973.
[19] J.C. González; M. Castellón; M.J. Castejón. “Técnicas de clasificación en el entorno de Weka para la determinación de cultivos de regadío (cítricos) en Librilla, Murcia (España)”. Teledetección: Agua y desarrollo sostenible. XIII Congreso de la Asociación Española de Teledetección. Calatayud, 23-26 de septiembre de 2009. Pp. 77-80.
[20] Janez Demsar. “Statistical comparisons of classifiers over multiple data sets”. *Journal of Machine Learning Research*, 7 (2006) 1-30.
[21] Mehl, H; Peinado, O. “Fundamentos del procesamiento digital de imágenes”.
[22] Mark Hall; Geoffrey Holmes. “Benchmarking attribute selection techniques for discrete class data mining”. Working Paper Series ISSN 1170-487X, April 2002.
[23] L. Ladha; T. Deepa. “Feature selection methods and algorithms”. *International Journal on Computer Science and Engineering (IJCSE)*, ISSN: 0975-3397, Vol. 3 No. 5, May 2011.
[24] Ron Kohavi; George H. John. “Wrappers for features subset selection”. AIJ special issue on relevance. August 1996.
[25] Julián Luengo; Salvador García; Francisco Herrera. “On the choice of the best imputation methods for missing values considering three groups of classification methods”. *Knowl Inf Syst* (2012) 32:77-108.
[26] Wikipedia
[27] “Feature extraction with example-based”. Exelis.
[28] Web del software Weka: http://www.cs.waikato.ac.nz/ml/weka/
[29] Francisco José Soltero Domingo; Diego José Bodas Sagi. “Clasificadores inductivos para posicionamiento web”. *El profesional de la información*, Vol. 14, No. 1, 2005.
[30] Descripción de los evaluadores y métodos de búsqueda de Weka: http://datamininggntua.files.wordpress.com/2008/04/weka-select-attributes.pdf
[31] Institut Cartogràfic de (ICC)
[32] S. Arenas ; J.F. Haeger; D. Jordano. “Aplicación de técnicas de teledetección y GIS sobre imágenes Quickbird para identificar y mapear individuos de peral silvestre (Pyrus Bourgeana) en bosque esclerófilo mediterráneo”. *Revista de Teledetección* (2011) 35, 55-71.