PROYECTO FIN DE BOOTCAMP

Reto Minsait, Cajamar,2020

ENERO DE 2022

MARÍA MELLADO PALACIOS
The Bridge School

ÍNDICE

1. Objetivos…… 2
2. Antecedentes……... 2
3. Formato y estructura…… 2
4. Introducción…… 3
5. Análisis Exploratorio de datos…… 8
6. Análisis de Componentes Principales……11
7. Preprocesado…… 15
8. Clasificación supervisada…… 18
9. Modelización de problema multiclase…… 19
10. Modelización de variables añadidas…… 21
11. Ajuste de hiperparámetros y validación cruzada…… 21
12. Implementación…… 23
13. Conclusiones y perspectivas futuras…… 24

1 OBJETIVOS

Construcción de un modelo automatizado que pueda ser usado para la clasificación de terrenos a través de imágenes satelitales. La métrica objetivo a maximizar es la exactitud, siendo esta:

Exactitud = N.º registros correctamente clasificados/N.º total de registros proporcionados por la organización.

2 ANTECEDENTES

Los datos utilizados corresponden al reto Minsait Land Classification (University Hack 2020). Las imágenes proporcionadas corresponden al satélite Sentinel II del servicio Copernicus de la Agencia Espacial Europea. A la lista de atributos extraídos de la imagen se dispone de un conjunto de fincas catastrales.

El dataset contiene un listado de superficies sobre las que se han recortado la imagen de satélite y se han extraído una serie de características de sus geometrías. Finalmente se ha etiquetado el conjunto de los datos según una clasificación de suelo.

Los ficheros contienen un total 55 variables: las 3 primeras de ellas relativas a la identificación de los registros y las 8 últimas variables son distintas referencias geométricas y relativas al entorno (geometría del edificio, métricas geométricas generadas automáticamente -GEOM-, metros cuadrados, año construcción y nº de plantas de los edificios del entorno). El ámbito geográfico de las imágenes es una zona concreta del municipio Madrid. La referencia (ID) es distinta y representativa de un elemento diferenciador.

Las imágenes satelitales se han tratado y se ha extraído información de 4 canales (R, G, B y NIR), correspondientes a las bandas de color rojo, verde y azul, y el infrarrojo cercano. El valor mostrado corresponde a la intensidad por deciles en cada imagen. Estas variables empiezan con la letra “Q”.

3 FORMATO Y ESTRUCTURA

Los datasets proporcionados (Modelar_UH2020 y Estimar_UH2020) están en formato txt y tienen la siguiente estructura:

- **Nombres de campo**: Incluidos en la cabecera.
- **Separador**: "|".
- **Codificación**: UTF-8.

4 INTRODUCCIÓN

Sentinel (compuesto por el 2A, lanzado en 2015 y 2B, lanzado en 2017) es un satélite desarrollado por la Agencia Europea del Espacio (ESA) que cuenta con 13 bandas que proporcionan imágenes de alta resolución y calidad radiométrica. Este satélite, que forma parte del programa Copérnico, tiene la misión de desarrollar observaciones de la Tierra, cambios en la corteza terrestre y gestión de los desastres naturales.

Las partes del espectro electromagnéticas más utilizadas en teledetección (en función de la longitud de onda) son:

- El espectro visible (VIS: región comprendida entre 0,4 y 0,7 micrómetros. Se divide en tres franjas: el azul (de 0,4 a 0,5 μm), el verde (de 0,5 a 0,6 μm) y el rojo (de 0,6 a 0,7 μm).

- El infrarrojo cercano (NIR): región del espectro comprendida entre 0,7 y 1,3 micrómetros. Resulta de especial importancia para discriminar masas vegetales y concentraciones de humedad.

- El infrarrojo medio (SWIR): región del espectro comprendida entre 1,3 y 8 micrómetros. Resulta idóneo para estimar contenidos de humedad en la vegetación y para detección de focos de temperatura.

El infrarrojo térmico (TIR): región comprendida entre 8 y 100 micrómetros. Detecta el calor proveniente de la mayor parte de las cubiertas terrestres.

- Las microondas (MW): comprendido entre 1 mm y 1 metro. Se utilizan en teledetección porque es un tipo de radiación bastante transparente a las cubiertas nubosas, es decir, que atraviesa bastante las nubes. Se suele llamar dominio óptico del espectro a la parte visible e infrarroja, hasta el SWIR, puesto que es la parte que comúnmente capturan los sensores ópticos

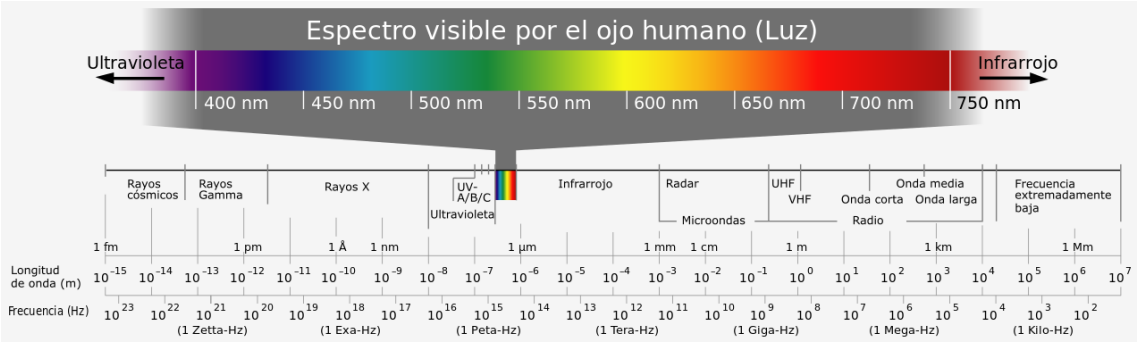

Figura 1: Espectro electromagnético. Radiación de longitudes de onda más corta (rayos gamma) hasta longitudes de onda más largas (ondas de radio).

Este satélite consta de 13 bandas espectrales, que van desde el espectro visible, el infrarrojo cercano (NIR) hasta el infrarrojo de onda corta (SWIR). Entre ellas, hay cuatro bandas de 10 m (tres del visible y una del NIR), seis bandas de 20 m, y tres bandas de 60 m de resolución.

10 metre spatial resolution:

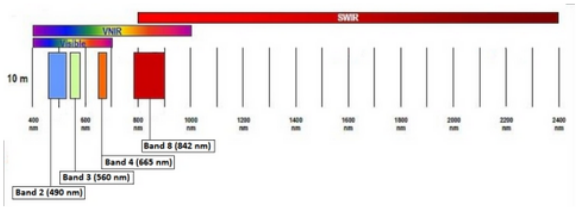

Figure 1: SENTINEL-2 10 m spatial resolution bands: B2 (490 nm), B3 (560 nm), B4 (665 nm) and B8 (842 nm)

20 metre spatial resolution:

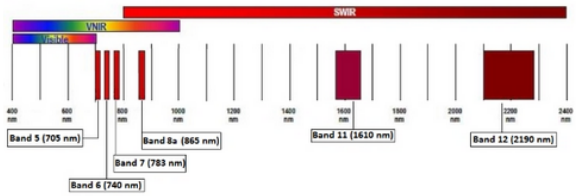

Figure 2: SENTINEL-2 20 m spatial resolution bands: B5 (705 nm), B6 (740 nm), B7 (783 nm), B8a (865 nm), B11 (1610 nm) and B12 (2190 nm)

60 metre spatial resolution:

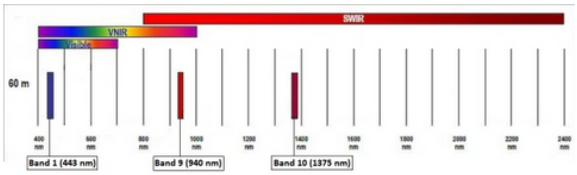

Figure 3: SENTINEL-2 60 m spatial resolution bands: B1 (443 nm), B9 (940 nm) and B10 (1375 nm)

Figura 2: Resolución espacial de Sentinel-2 (tomado de la página web del programa Copernicus(sentinel.copernicus.eu).

	Nombre de la banda	Longitud de onda (nm)	Ancho de onda(nm)	Resolución(m)
1	Coastal aerosol	443,9	27	60
2	Azul	496,6	98	10
3	Verde	560	45	10
4	Rojo	664,5	38	10
5	Borde Rojo vegetación	703,9	19	20
6	Borde Rojo vegetación	740,2	18	20
7	Borde Rojo vegetación	782,5	28	20
8	NIR	835,1	145	10
8a	NIR estrecho	864,8	33	20
9	Vapor de agua	945	26	60
10	SWIR - Cirrus	1373,5	75	60
11	SWIR	1613,7	143	20
12	SWIR	2202,4	242	20

Tabla 1: Longitudes de onda y anchos de banda para las tres resoluciones espaciales del satélite Sentinel 2A, Sensor MSI (tomado de la página web del programa Copernicus (sentinel.copernicus.eu).

	Nombre de la banda	Longitud de onda (nm)	Ancho de onda(nm)	Resolución(m)
1	Coastal aerosol	442,3	45	60
2	Azul	492,1	98	10
3	Verde	559	46	10
4	Rojo	665	39	10
5	Borde Rojo vegetación	703,8	20	20
6	Borde Rojo vegetación	739,1	18	20
7	Borde Rojo vegetación	779,7	28	20
8	NIR	833	45	10
8a	NIR estrecho	864	32	20
9	Vapor de agua	943,2	27	60
10	SWIR - Cirrus	1376,9	76	60
11	SWIR	1610,4	141	20
12	SWIR	2185,7	238	20

Tabla 2: Longitudes de onda y anchos de banda para las tres resoluciones espaciales del satélite Sentinel 2B, Sensor MSI (tomado de la página web del programa Copernicus (sentinel.copernicus.eu).

Las imágenes de las distintas bandas se pueden combinar entre ellas para producir una imagen en color real *o* falso color en función de las bandas escogidas. Con las bandas 4, 3 y 2 podemos componer una imagen como la veríamos nosotros; bandas rojo (Red), Verde (Green) y Azul (Blue); la conocida como RGB. Mediante la combinación de otras bandas se pueden conseguir otras composiciones que potencien algún elemento.

Bandas 8-4-3 (Falso color): combinación de bandas muy útil para estudio de vegetación, patrones de suelo o etapas de crecimiento de cultivos. En general, los tonos rojos intensos indican hojas anchas y/o vegetación más sana, mientras que los rojos más claros significan pastizales o áreas escasamente vegetadas. Las áreas urbanas densamente pobladas se muestran en azul claro.

Bandas 12 (11), 8A, 4: banda idónea para el estudio de la salud de la vegetación, suelos perturbados y detección de camuflaje. Así, la vegetación vigorosa e irrigada, y las áreas ribereñas se exhiben en verde claro, mientras que las tierras secas y las áreas naturales son verde opaco. Los suelos aparecen como bronceado, marrón y malva.

Bandas 4,11 y 12: falso color para la detección de zonas urbanas. Las áreas urbanas aparecen en tonos magentas mientras que los pastos aparecen en tonos verdes claros. Las áreas forestales aparecerán en color verde oliva.

Bandas 12,11,8A: análisis sobre penetración atmosférica.

Bandas 4,3,2: a esta combinación se le denomina color natural puesto que involucra las tres bandas visibles.

A partir de estas bandas también se pueden calcular otras (neo-canales) que resalten alguna característica en cuestión. Así, la luminancia (o luminosidad), del sistema de color HSL (Matiz, Saturación y Luminosidad):

$$L = \frac{\max(R, G, B) + \min(R, G, B)}{2}$$

El canal c3 del sistema de color c1c2c3 se define como:

$$c = \arctan\left(\frac{B}{\max(R, G)}\right)$$

Índice Diferencial Normalizado de Vegetación, NDVI:

$$NDVI = \frac{NIR - R}{NIR + R}$$

Índice de vegetación ajustado al suelo modificado, MSAVI:

$$MSAVI = \frac{(NIR - R) \times (1 + L)}{NIR + R + L}$$

Índice de exceso Verde:

$$ExG = \frac{2 \times G - R - B}{R + G + B}$$

Índice Diferencial Modificado de Agua Normalizado, MNDWI:

$$MNDWI = \frac{G - NIR}{G + NIR}$$

Una imagen hiperespectral (HSI, por sus siglas en inglés, Hyperspectral Satellite Images) es aquella que contiene varias bandas espectrales a lo largo de todo el espectro electromagnético. Un píxel de dicha imagen no vendrá definido por un único valor de intensidad, sino que tendrá tantos como bandas se hayan analizado. Este tipo de imágenes se diferencian de las imágenes múltiples fundamentalmente en el número de bandas espectrales. Las imágenes multiespectrales poseen generalmente de 3 a 15 bandas mientras que las hiperespectrales poseen cientos de bandas mucho más estrechas, del orden de los nanómetros. La posición e intensidad de estas bandas, así como la forma de la curva espectral, se puede utilizar para identificar y clasificar diferentes materiales.

La detección hiperespectral (detección, clasificación y cuantificación) incluye una gran cantidad de necesidades a nivel comercial, civil, de defensa y de inteligencia. Las aplicaciones son numerosas y variadas, entre las que cabe citar:

- La identificación de recursos naturales y la evaluación de impactos ambientales
- La exploración en industrias petroleras y mineras
- El monitoreo de emisiones de gases de efecto invernadero
- La medición de la calidad del agua
- La medición de la distribución de temperatura y humedad en la atmósfera para el modelamiento climatológico
- La detección, rastreo y monitoreo de objetivos de interés y actividades ilegales
- La evaluación del progreso estacional de los cultivos y el estrés en los cultivos

5 ANÁLISIS EXPLORATORIO DE LOS DATOS (EDA)

Para la creación de un modelo primero hemos realizado el análisis exploratorio de datos sobre el fichero denominado “Modelar_UH2020”. Este fichero consta de 103230 filas y 56 columnas. Todas las variables son de tipo int64 o float64 excepto tres variables que son de tipo objecto y que son el ID, CADASTRALQUALITYID y la variable objetivo CLASE. La suma total de valores vacíos (Nan, por sus siglas en inglés, Not a Number) en todo el dataset fue de 20, por lo que se decidió eliminar por completo las filas que los contenían (del archivo Modelar_UH2020). Se ha podido comprobar que estos valores vacíosestán en la misma fila, por lo que solo es eliminada una fila.

La variable CLASE está compuesta de las siguientes categorías:

- Industrial
- Residential
- Public
- Office
- Retail
- Agriculture
- Other

Para poder estudiar la variable dependiente con más detalle, representamos la gráfica de conteo, donde se puede ver claramente que hay una clase mayoritaria (La clase “Residential”, que corresponde a la clase 5).

Figura 3: Gráfica de conteo de la variable CLASE.

	Conteo de la clase
RESIDENTIAL	90173
INDUSTRIAL	4490
PUBLIC	2976
RETAIL	2093
OFFICE	1828
OTHER	1332
AGRICULTURE	338

A continuación, se representa también mediante gráfica de conteo (countplot) las variables MAXBUILDINGFLOOR y CADASTRALQUALITYID.

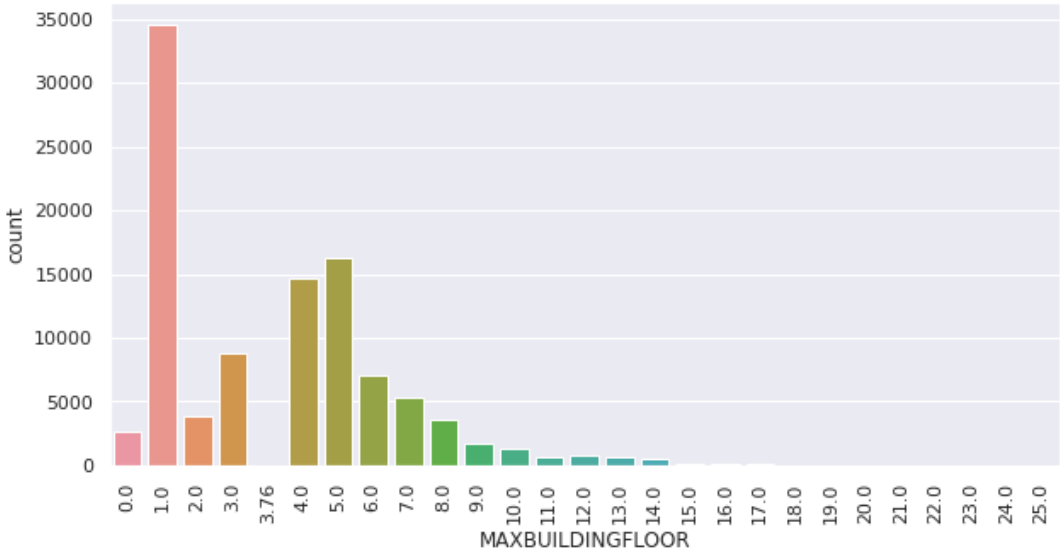

Figura 4: Grafica de conteo de la variable MAXBUILDINGFLOOR.

Se puede observar mediante la gráfica que el número de pisos más frecuente es de 1 y que a partir de 15 pisos no se obtienen resultados.

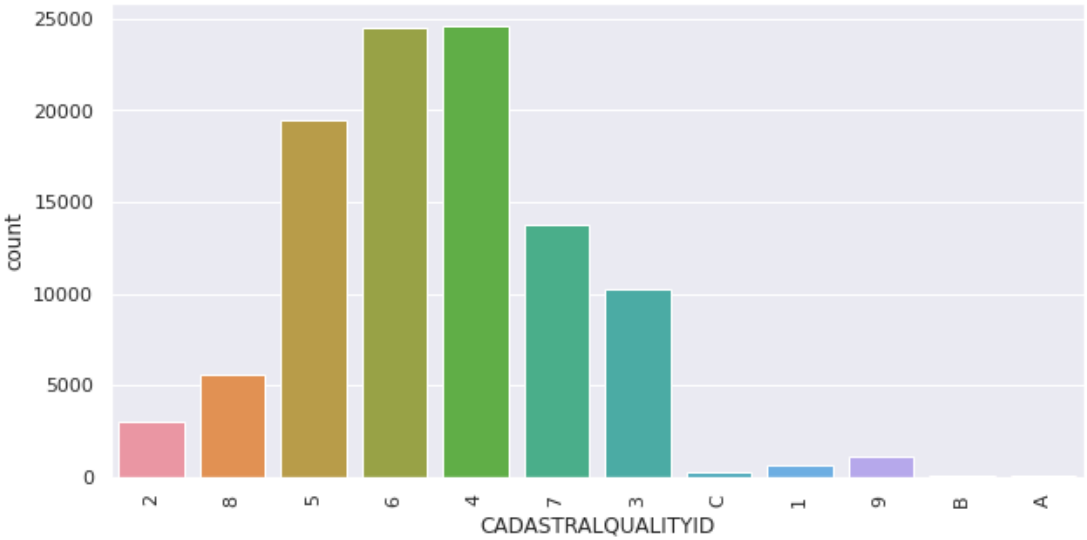

Figura 5: Grafica de conteo de la variable CADASTRALQUALITYID.

La variable CADASTRALQUALITYID se ha codificado mediante LabelEncoder para codificar las etiquetas de las características categóricas en valores numéricos entre 0 y 1.

A continuación, se representa el histograma del año de construcción. Se puede observar una distribución asimétrica hacia la derecha. Se puede concluir que la gran mayoría de edificios fueron construidos a partir de 1950.

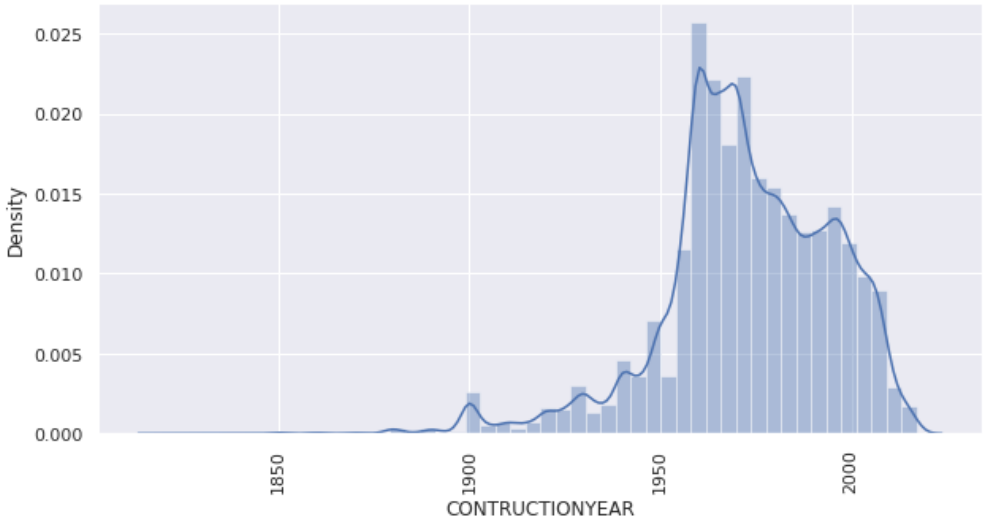

Figura 6: Histograma de la variable CONSTRUCTIONYEAR.

También se han representado las variables de longitud y latitud:

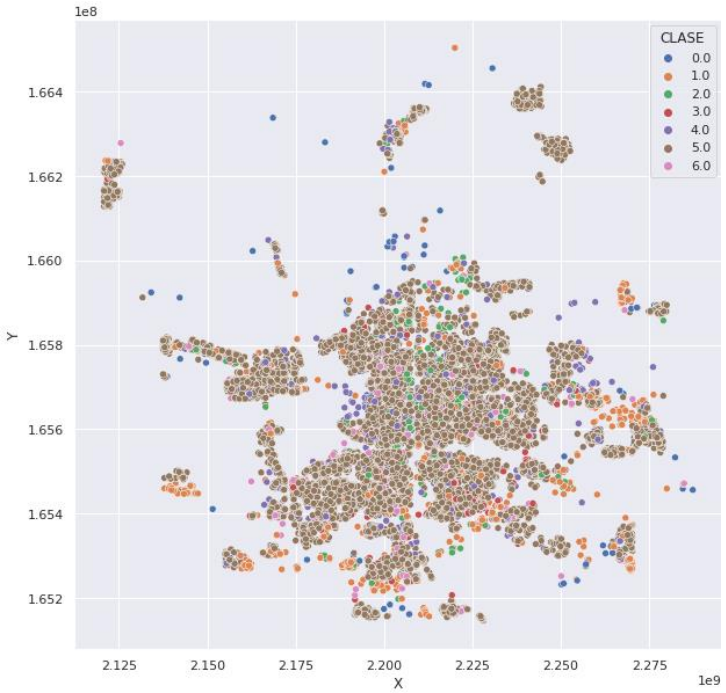

Figura 7: Representación gráfica de las variables X e Y.

En cuanto a las variables geométricas (GEOMR1, GEOMR2, GEOMR3 y GEOMR4), no se han incluido dentro de las variables utilizadas para el modelado puesto que son variables utilizadas el cálculo de la variable AREA, que si se ha incluido en el modelado. Estas cuatro variables geométricas aumentarían el coste computacional mientras que no aportarían gran información, por lo que se eliminan de las variables incluidas en el modelo.

6 ANÁLISIS DE COMPONENTES PRINCIPALES

El principal objeto del análisis de componentes principales (PCA, por sus siglas en inglés) es la reducción de dimensionalidad. Es un método muy efectivo cuando la correlación entre variables es alta. Mediante este análisis se identifican aquellas direcciones en las que la varianza es mayor. Para que las variables tengan media igual a cero y desviación estándar igual a 1 es aconsejable estandarizar los datos antes de realizar PCA. De esta manera, se evita que aquellas variables cuya escala sea mayor sean predominantes. En este caso, se ha construido un "pipeline" en el que se realiza un escalado de los datos seguido de un análisis de componentes principales. A continuación, se ha creado un dataframe donde se recogen todas las componentes principales. Al conjunto de datos original de 55 variables se le ha eliminado las variables ID, GEOMR1, GEOMR2, GEOMR3 y GEOMR4 por lo que se queda un conjunto de datos con 50 variables. A partir de este conjunto de datos se crean 49 (n -1) componentes principales. En el siguiente mapa de calor se detallan las variables y las componentes principales.

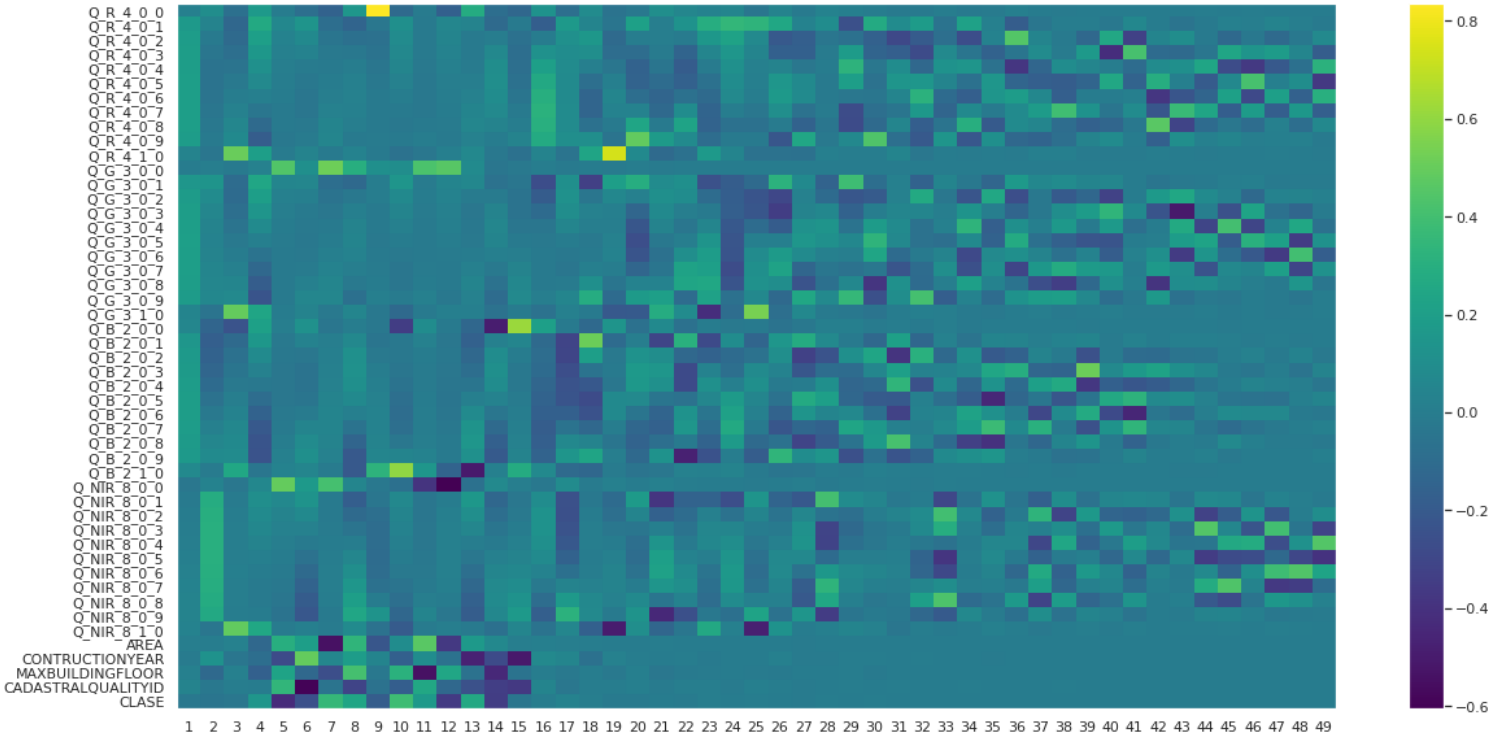

Figura 8: mapa de calor de las variables y las componentes principales.

Teniendo en cuenta que las variables están normalizadas para tener media cero, la varianza total presente en el set de datos se define como:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

y la varianza explicada por la componente *m* es:

$$\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2$$

Por lo tanto, la proporción de varianza explicada por la componente *m* viene dada por el siguiente ratio:

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}$$

Tanto la proporción de varianza explicada como la proporción de varianza explicada acumulada son dos valores de gran utilidad a la hora de decidir el número de componentes principales a utilizar en los análisis posteriores. Si se calculan todas las componentes principales de un set de datos, entonces, aunque transformada, se está almacenando toda la información presente en los datos originales. El sumatorio de la proporción de varianza explicada acumulada de todas las componentes es siempre 1.

En las siguientes figuras (9 y 10) se representan el porcentaje de varianza y varianza acumulada por cada componente obtenida.

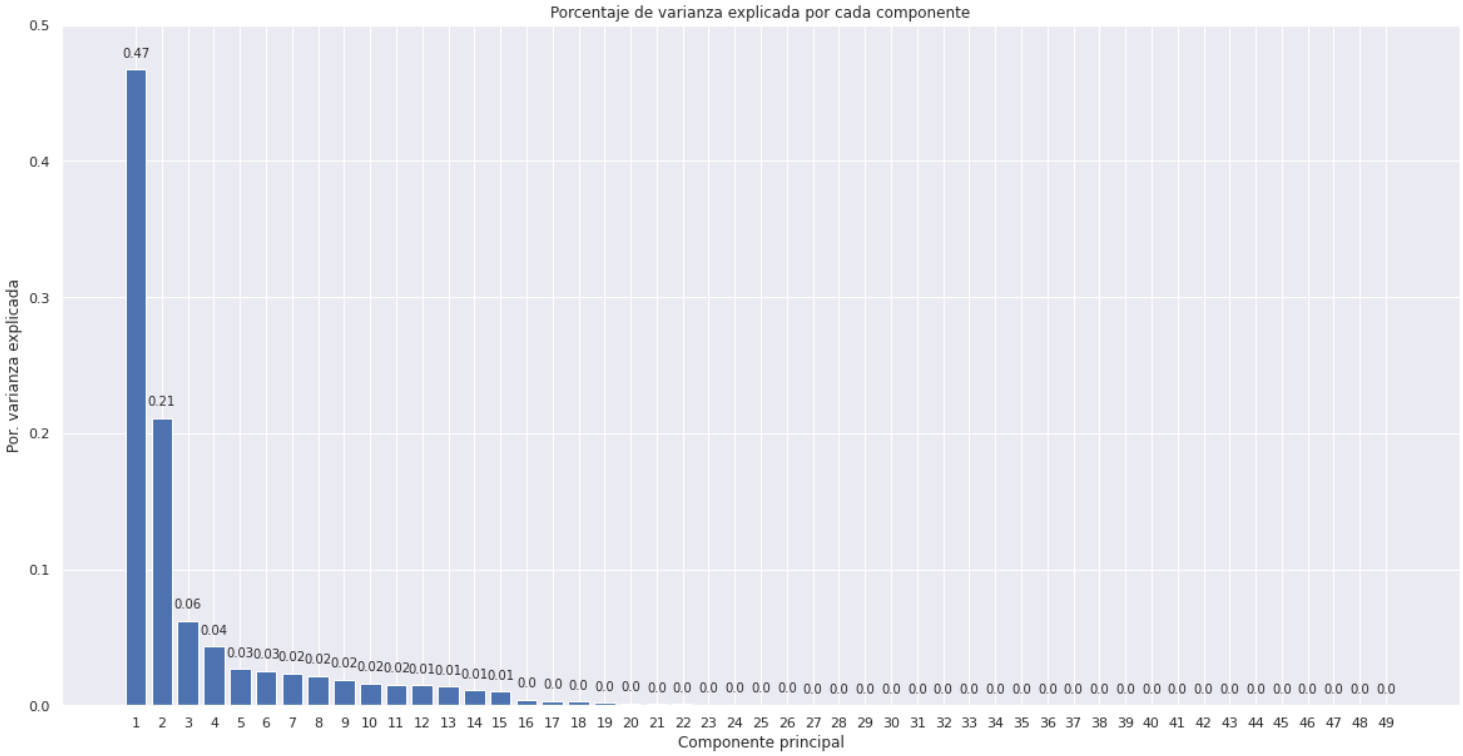

Figura 9: Porcentaje de varianza explicada por cada componente.

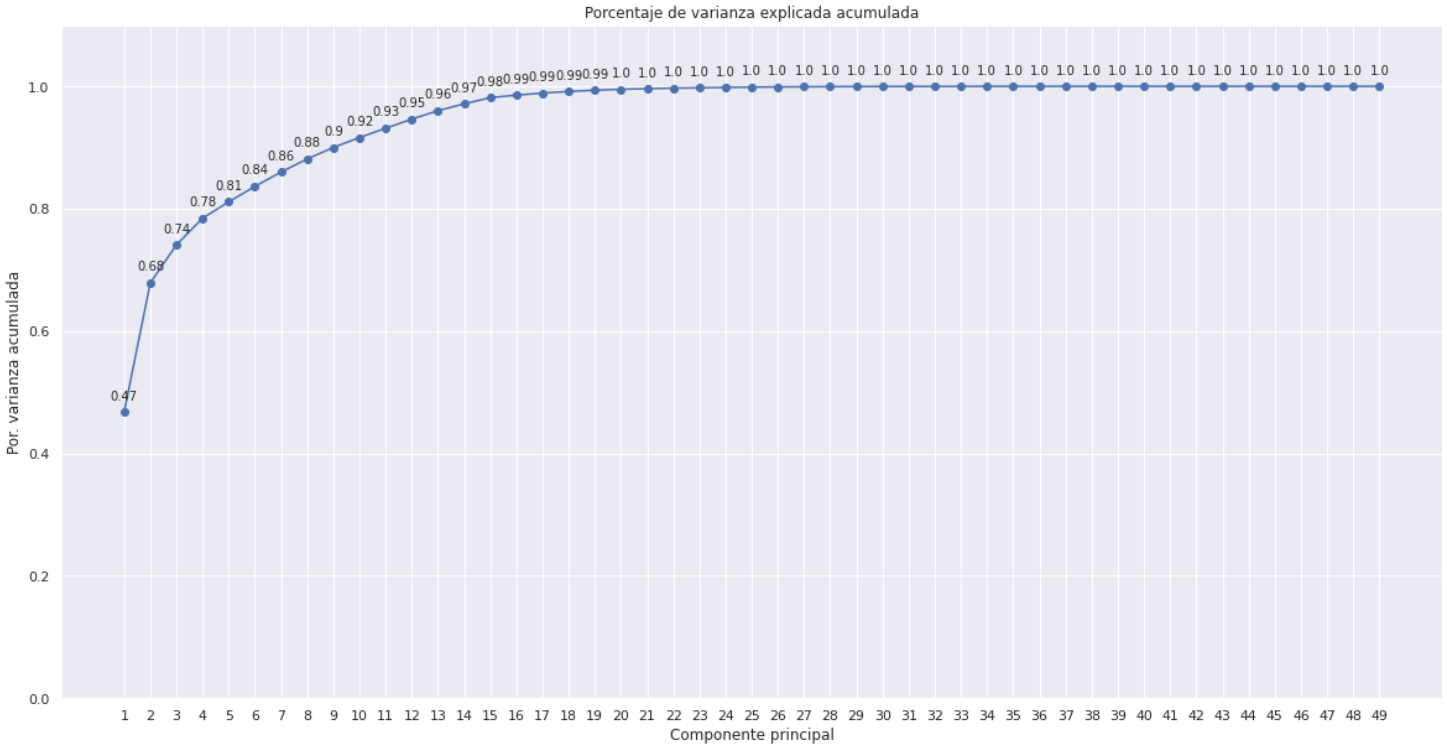

Figura 10: Porcentaje de varianza acumulada explicada por cada componente.

Se observa claramente en la figura 9 que el 47 % de la varianza puede ser explicada mediante la primera componente principal, así como el 26 % mediante la segunda componente. A su vez, a través de la figura 10, se puede observar que el 99 % de la varianza se puede explicar mediante la utilización de las primeras 15 componentes principales. Posteriormente, se realiza la proyección de las observaciones (reducción de la dimensionalidad de las nuevas observaciones proyectándolas en el espacio definido por las componentes) seguido de una reconstrucción (revirtiendo así la proyección). Seguidamente se volvieron a crear las variables X y Y atendiendo a los datos guardados en la variable reconstrucción. Se aplicó el clasificador de bosques aleatorios y se obtuvo una puntuación de 92.16.

Atendiendo a las figuras anteriores se puede deducir que con las 15 primeras componentes es suficiente para explicar la varianza casi en su totalidad. Es por ello que creamos un conjunto de datos únicamente con estas componentes y volvemos a aplicar el clasificador aumentando la puntuación a 99.92%.

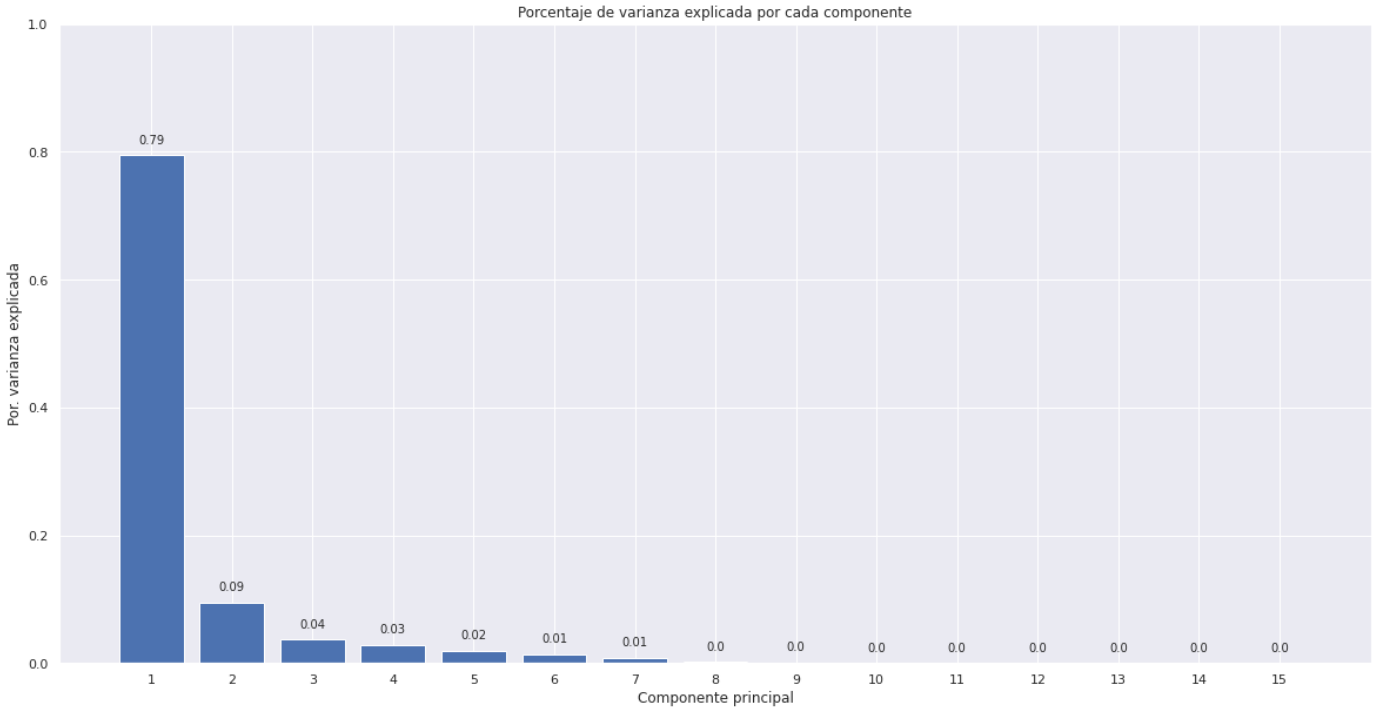

Figura 11: Porcentaje de varianza explicada por cada componente (15 componentes).

7 PREPROCESADO DE LOS DATOS

Tras analizar la variable CLASE, se observa que hay una clase mayoritaria (la RESIDENTIAL), es decir, tenemos unos datos claramente desbalanceados. Para intentar solventar este problema se toman dos estrategias diferentes. Por un lado, antes de crear un modelo, se tratan los datos mediante técnicas de sobremuestreo y submuestreo (duplicando muestras de la clase minoritaria y eliminando muestras de la clase mayoritaria, respectivamente). Por otro lado, la otra estrategia a seguir para solventar de desbalanceo, consiste en eliminar la clase mayoritaria y entrenar el modelo únicamente con el resto de las clases.

Las técnicas de sobremuestreo ("oversampling") y submuestreo ("undersampling") que se han utilizado son:

- Submuestreo: random, TomekLinks, NearMiss
- Sobremuestreo: random, SMOTE
- Combinación de ambas: SMOTE-Tomek Links

Los resultados obtenidos después de aplicar estas las técnicas de submuestreo, seguido de utilizar bosque aleatorio como clasificador para el modelado se recogen en la siguiente tabla:

	Submuestreo Random			**TomekLinks**			**Nearmiss**		
	P	**E**	**F**	**P**	**E**	**F**	**P**	**E**	**V-F**
0	0.83	1.00	0.91	1.00	1.00	1.00	1.00	1.00	1.00
1	0.50	0.67	0.57	0.83	0.95	0.89	1.00	0.67	0.80
2	0.50	0.50	0.50	0.75	0.86	0.80	0.67	0.50	0.57
3	1.00	0.60	0.75	0.23	1.00	0.37	0.25	0.20	0.22
4	0.50	0.50	0.50	0.16	0.86	0.28	0.25	0.50	0.33
5	0.80	0.67	0.67	1.00	0.90	0.95	1.00	1.00	1.00
6	0.50	1.00	1.00	1.00	1.00	1.00	0.50	1.00	0.67
puntuación	1.00			0.99			0.69		

En esta tabla, al igual que las detalladas de aquí en adelante específica para cada clase la precisión (en inglés, precisión), que nos da una idea de la calidad de la clasificación. También se detalla la exhaustividad (en inglés, recall), que nos va a dar información sobre la cantidad que el modelo es capaz de clasificar. Por último, también se detalla el valor F (en inglés, F1-score), que combina ambas medidas de precisión y exhaustividad en un solo valor. A su vez se especifica la puntuación total (en inglés, score). Para una representación en tablas más sencilla se representará la precisión con una "P", la

exhaustividad con una “E” y el valor-F con una “F”. A continuación, se define cada uno de estos términos:

$$precisión = \frac{TP}{TP + FP}$$

$$exhaustividad = \frac{TP}{TP + FN}$$

$$valor\ F = 2 * \frac{precisión * exhaustividad}{precisión - exhaustividad}$$

$$exactitud = \frac{TP + TN}{TP + TN + FP + FN}$$

Donde:

TP: Verdadero positivo (en inglés, True Positive)

TN: Verdadero Negativo (en inglés, True Negative)

FP: Falso positivo (en inglés, False Positive)

FN: Falso Negativo (en inglés, False Negative)

Estos cuatro términos provienen de la matriz de confusión.

A continuación, se detallan los resultados obtenidos después de aplicar estas las técnicas de sobremuestreo, seguido de utilizar árbol aleatorio se recogen en la siguiente tabla:

	Sobremuestreo Random			**SMOTE**		
	P	**E**	**F**	**P**	**E**	**F**
0	0.82	1.00	0.90	1.00	0.25	0.40
1	0.55	0.64	0.59	0.79	0.95	0.87
2	0.61	0.58	0.60	0.83	0.67	0.74
3	0.78	0.64	0.70	0.88	0.44	0.58
4	0.59	0.43	0.50	1.00	0.95	0.97
5	0.57	0.40	0.47	0.97	1.00	0.99
6	0.64	0.88	0.74	0.00	0.00	0.00
puntuación	0.66			0.96		

Por último, se detallan los resultados de realizar la combinación de realizar una combinación de sobremuestreo y submuestreo mediante la técnica de SMOTE-Tomek. Mediante esta técnica, primero se aplica el método SMOTE, creando ejemplos plausibles en la clase minoritaria. A continuación, se aplica Tomek Links, método mediante el cual se identifican los pares de vecinos de distinta clase que están en el límite y se eliminan. Los resultados de aplicar el algoritmo de clasificación árbol aleatorio a los datos tratados mediante esta técnica son los siguientes:

	SMOTE-Tomek		
	P	**E**	**F**
0	1.00	1.00	1.00
1	0.97	0.80	0.88
2	0.90	0.85	0.87
3	0.78	0.97	0.86
4	0.91	0.91	0.91
5	0.89	0.89	0.89
6	1.00	0.99	1.00
puntuación	0.92		

Los resultados de la aplicación de diferentes algoritmos de clasificación con un dataset en el que se ha eliminado la clase mayoritaria (se elimina la clase 5) se detallan a continuación:

	Bosques aleatorios		
	P	**E**	**F**
0	0.50	0.20	0.29
1	0.90	0.94	0.92
2	0.82	0.75	0.78
3	0.80	0.86	0.83
4	0.81	0.93	0.86
6	0.96	0.79	0.87
puntuación	0.99		

La aplicación de los clasificadores como el árbol de decisión y aumento de gradiente da como resultado una precisión de uno para todas las clases, así como del score. En el caso de SVM ocurre todo lo contrario, da como resultado una precisión de cero para todas las clases. En el caso de regresión logística solo clasifica una clase y para k vecinos solo clasifica algunas clases.

La aplicación de técnicas de submuestreo y sobremuestreo da como resultado unos valores altos de exactitud, así como de precisión, exhaustividad y valor-F en todas las clases. Mediante la utilización del clasificador de bosques aleatorios al dataset tratado mediante este tipo de técnicas consigue clasificar todas las clases con una puntuación superior al 50% en todos los casos. Asimismo, la eliminación de la clase mayoritaria

también da buen resultado, obteniendo unos valores altos de precisión del resto de clases mediante esta estrategia.

8 CLASIFICACIÓN SUPERVISADA

Uno de los objetivos fundamentales de la teledetección es la clasificación del área de estudio en base a la litología, tipo de vegetación o usos de suelo. Para ello, se suele determinar el número de clases y propiedades de estas en relación con las variables y asignar a cada uno de los individuos una de las clases utilizando una regla de decisión basada en las propiedades de los individuos y las clases en relación con las variables. En teledetección el conjunto de variables está compuesto por la reflectividad en cada una de las bandas. Sin embargo, además de esta información espectral puede utilizarse información textural e información contextual. La información textural hace referencia a las características en la vecindad de un píxel. Se definen una serie de variables (reflectividad media, varianza, autocorrelación, etc.) que tratan de cuantificar algunas de las propiedades cualitativas que se estudian en fotointerpretación.

Para el modelado se realizó clasificación supervisada, es decir, se partió de un conjunto de clases conocido a priori. Estas clases deben caracterizarse en función del conjunto de variables mediante la medición de estas en individuos cuya pertenencia a una de las clases no presente dudas (áreas de entrenamiento).

Se entrenaron los datos con diferentes algoritmos de modelización tales como bosques aleatorios, SVM, regresión logística y k vecinos. Los resultados de cada uno de los entrenamientos se muestran en la siguiente tabla:

	k-vecinos			Bosques aleatorios			SVM		
	P	**E**	**F**	**P**	**E**	**F**	**P**	**E**	**F**
0	0.44	0.31	0.37	1.00	0.29	0.44	0.20	0.33	0.25
1	0.52	0.44	0.48	0.78	0.87	0.82	0.79	0.87	0.83
2	0.15	0.08	0.10	0.47	0.45	0.46	0.55	0.39	0.46
3	0.04	0.01	0.02	0.45	0.38	0.42	0.62	0.50	0.56
4	0.2	0.04	0.07	0.81	0.83	0.82	0.84	0.86	0.85
5	0.90	0.97	0.94	1.00	1.00	1.00	0.99	1.00	1.00
6	0.29	0.06	0.10	1.00	0.84	0.91	1.00	0.63	0.77
puntuación	0.87			0.99			0.89		

	Análisis Discriminante Lineal (LDA)		
	P	**E**	**F**
0	0.21	1.00	0.35
1	0.38	0.35	0.37
2	0.50	0.04	0.07
3	0.14	0.10	0.12
4	0.00	0.00	0.00
5	0.91	0.97	0.94
6	0.00	0.00	0.00
puntuación	0.87		

Los resultados de aplicar el clasificador árbol de decisión y aumento de gradiente no se han incluido en la table puesto que se obtiene para todas las clases una precisión de 1 y un score de 1 también. Tampoco se incluyen los resultados de aplicar regresión logística, puesto que únicamente. Para SVM y bosques aleatorios se utilizó como criterio de puntuación f1-weighted.

9 MODELIZACIÓN DE PROBLEMA MULTICLASE

Ciertos algoritmos de clasificación permiten el uso de más de dos clases intencionadamente mientras que otros restringen los posibles resultados a uno de dos valores (un modelo binario, o de dos clases). Sin embargo, incluso los algoritmos de clasificación binaria se pueden adaptar a las tareas de clasificación de varias clases con diversas estrategias.

Mediante el método uno frente a todos (One vs All) se crea un modelo binario para cada una de las distintas clases de salida. Dicho método evalúa cada uno de estos modelos binarios para las clases individuales con respecto a su complemento (todas las demás clases del modelo) como si fuese un problema de clasificación binaria. Además de la eficacia de cálculo (solo se necesitan clasificadores, n clases), una ventaja de este enfoque es su interpretación. Puesto que cada clase solo está representada por un clasificador de uno contra uno, es posible obtener información sobre la clase mediante la inspección de su clasificador correspondiente. Esta es la estrategia que se usa con más frecuencia para la clasificación multiclase y es una opción predeterminada apropiada. En esencia, el componente crea un conjunto de modelos individuales, y luego combina los resultados para crear un único modelo que predice todas las clases. Cualquier clasificador binario se puede utilizar como base para un modelo uno frente a todos.

Por otro lado, en el método uno frente a uno (One vs One), se crea un modelo binario por par de clases. En el momento de la predicción, se selecciona la clase que recibió el mayor número de votos. Puesto que requiere ajustar clasificadores *n clases* (n clases - 1) / 2*, este método suele ser más lento que uno frente a todos, debido a su complejidad. Sin embargo, este método puede ser ventajoso para algunos algoritmos. Esto se debe a que cada problema de aprendizaje individual solo implica un pequeño subconjunto de los datos, mientras que, con uno frente a todos, el conjunto de datos completo se usa tantas veces como clases haya (n).

Para la implementación de estas estrategias se ha utilizado la librería sklearn (sklearn multiclass) y los módulos OneVsOneClassifier y OneVsRestClassifier. Se utilizó el clasificador SVM, obteniéndose los siguientes resultados:

	One vs Rest			One vs One		
	P	**E**	**F**	**P**	**E**	**F**
0	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00
5	0.89	1.00	0.94	0.89	1.00	0.94
6	0.00	0.00	0.00	0.00	0.00	0.00
exactitud	0.89			0.89		

Para el uso de SVM con multiclase se pueden utilizar tres vías de trabajo:

- One vs One (OVO)
- One vs All (OVA)
- Directed Acyclic Graph(DAG)

Para la realización de estas técnicas se ha utilizado la librería sklearn(multiclass). Para OneVsRestClassifier se obtiene un score alto (87%) pero la precisión de todas las clases excepto una es de cero. Esto es, solo logra clasificar correctamente la clase mayoritaria (Clase 5). Lo mismo sucede al aplicar la técnica OneVsOneClassifier. Ninguna de estas dos estrategias para abordar la clasificación multiclase parece ser efectiva para este caso.

10 MODELIZACIÓN CON VARIABLES AÑADIDAS

Para un estudio más completo de las variables del dataset asociadas a las bandas espectrales se han añadido cuatro variables: Luminance, NDVI, ExG y MNDWI. Estás variables, cuya definición se detalla en la introducción del presente documento, hacen uso de la combinación de las bandas. Se utilizan las variables R, G, B y NIR del dataset para construir estás cuatro nuevas variables y a continuación se crea un nuevo dataframe en el que están incluidas. Por último, se dividen los datos en proporción 80/20 (train/test) y se aplica los clasificadores de bosques aleatorios y SVM. El resultado es:

	Bosques aleatorios			**SVM**		
	P	**E**	**F**	**P**	**E**	**F**
0	0.79	0.24	0.37	0.79	0.24	0.37
1	0.76	0.44	0.56	0.76	0.44	0.56
2	0.62	0.01	0.03	0.62	0.01	0.03
3	0.00	0.00	0.00	0.00	0.00	0.00
4	1.00	0.03	0.07	1.00	0.03	0.07
5	0.90	1.00	0.95	0.90	1.00	0.95
6	0.00	0.00	0.00	0.00	0.00	0.00
exactitud	0.90			0.90		

En ambos casos se obtiene una puntuación del 90 %. Atendiendo a la precisión, ninguno de los dos clasificadores consigue la clasificación de las clases 3, 4 y 6, mientras que la clasificación parece adecuada en el caso del resto de clases.

11 AJUSTE DE HIPERPARÁMETROS Y VALIDACIÓN CRUZADA

Se ha realizado validación cruzada de 10 iteraciones, es decir, el proceso de validación cruzada se ha repetido 10 veces con cada uno de los subconjuntos de datos de prueba. A continuación, se calculó la media aritmética del score de cada uno de los algoritmos utilizados. Como se puede apreciar en la figura 8, los clasificadores que resultan en una puntuación más baja, y por lo tanto se podrían descartar para el caso del presente estudio, son K-vecinos, SVC, Multiple Layer Perceptrón y LDA.

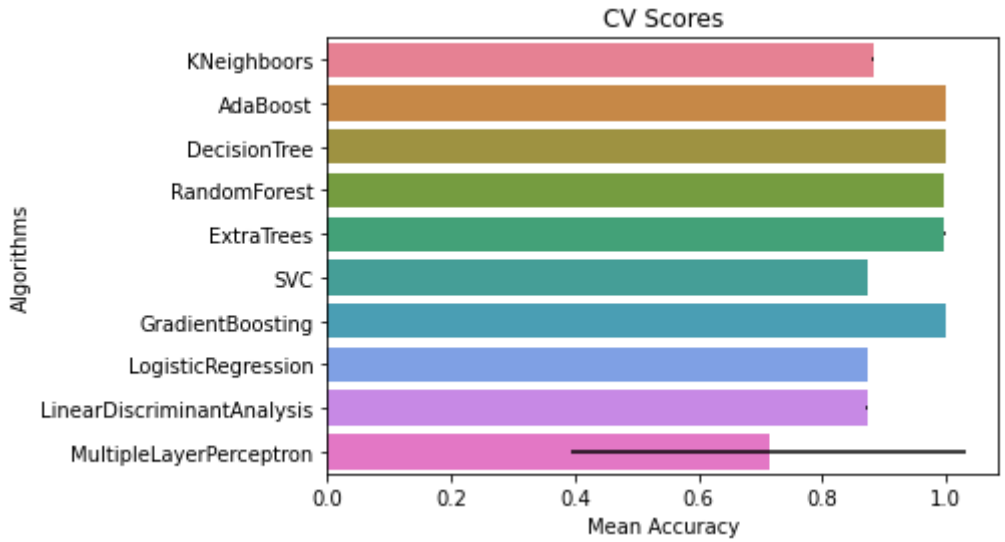

Figura 8: Diagrama de barras de la exactitud de cada uno de los algoritmos de clasificación utilizados para hacer la validación cruzada.

El ajuste de hiperparámetros se realizó sobre aquellos algoritmos de clasificación que dieron mejores resultados, es decir, en los que se obtuvo una puntuación más alta, que fueron máquinas de vectores soporte y bosques aleatorios. En el caso de la clasificación mediante SVM, el ajuste de hiperparámetros con validación cruzada se realizó únicamente con tres iteraciones puesto que un número mayor de iteraciones aumentaba considerablemente el coste computacional. En cuanto al kernel, se introdujo únicamente el kernel gaussiano rbf (del inglés, radial basis function kernel), para no aumentar demasiado la duración de la ejecución. Los valores del parámetro de regularización, C, que controla el balance entre sesgo y varianza, que se introdujeron en el ajuste fueron 1,10,50,100,200 y 300. Los valores del coeficiente del kernel, gamma, fueron 1, 10, 50, 100, 200 y 300. Una vez realizado la técnica de búsqueda por cuadrícula seguido de validación cruzada (del inglés, GridSearchCV) se obtuvieron los mejores parámetros. Así, se estimó que el parámetro C debe ser 1, gamma 0.0001 y la mejor puntuación fue 0.87 (tomando la exactitud como método de puntuación).

A continuación, también se realizó el ajuste de hiperparámetros para el caso del clasificador de bosques aleatorios. Se establecieron 5 iteraciones puesto que valores más altos elevaban la duración de la ejecución del código en exceso. No se estableció profundidad máxima del árbol (del inglés, max Depth). El número mínimo de muestras antes de la separación en nuevos nodos (min simples Split) se estableció como 2, 6 y 20 y el número máximo (max samples leaf) como 100, 200, 300 y 400. Como evaluador de la calidad de la separación de utilizó el criterio de Gini. Los mejores resultados fueron un número mínimo de muestras de 6, un número máximo de 1 y un número de estimadores (n estimators) de 200. La puntuación más alta fue de 0.97 (atendiendo a exactitud como criterio de puntuación).

	Bosques aleatorios			SVM		
	P	E	F	P	E	F
0						
1	0.89	0.96	0.93	0.89	0.96	0.93
2	0.75	0.86	0.80	0.75	0.86	0.86
3	1.00	0.25	0.40	1.00	0.25	0.25
4	0.92	0.71	0.80	0.92	0.71	0.71
5	0.98	1.00	0.99	0.98	1.00	1.00
6	1.00	0.44	0.61	1.00	0.44	0.44
exactitud	0.97			0.97		

12 IMPLEMENTACIÓN

La predicción de la variable dependiente, en este caso la variable CLASE, que hace referencia al tipo de suelo, se realizó sobre el archivo "Estimar.txt". Para ello, se utilizaron los clasificadores de bosques aleatorios y SVM, que son con los que se obtuvieron un score más alto en la fase de entrenamiento y validación. Las predicciones fueron guardadas en un archivo con extensión ".csv".

Los notebooks utilizados para la realización del presente proyecto están nombrados como:

- Proyecto_EDA
- Proyecto_Modelos
- Proyecto_Imágenes

En el cuaderno denominado "EDA" queda recogido el análisis exploratorio de los datos mientras que en el cuaderno "Modelos" quedan detallados todos los modelos que se han probado, así como el procedimiento de la predicción sobre el archivo "Estimar.txt". Se ha guardado en un archivo de formato pickle los datos una vez tratados (tras el EDA). Así mismo, también se ha guardado en formato pickle el modelo que dio mejores resultados.

En el cuaderno "Imágenes" se hace un estudio más detallado de las variables relativas a imágenes espectrales, esto es, a las variables que recogen las longitudes de onda de RGB y NIR.

A su vez, para la posterior producción del modelo óptimo, se han creado dos archivos ejecutables:

Main.py

Preprocessing.py

13 CONCLUSIONES Y PERSPECTIVAS FUTURAS

El rápido avance de la teledetección y el desarrollo de sensores que permitan recoger imágenes hiperespectrales (mejorando así sustancialmente la calidad y resolución de las imágenes en comparación con las imágenes multiespectrales) está permitiendo el estudio a gran escala de la superficie terrestre. Sin embargo, el gran tamaño que supone la recogida de este tipo de imágenes, así como el posterior tratamiento y almacenamiento supone un reto de gran envergadura. Por este motivo, el aprendizaje automático e inteligencia artificial juega un papel muy importante en el desarrollo de la teledetección.

A pesar de la creciente popularidad del aprendizaje profundo, los clasificadores de aprendizaje supervisado siguen siendo populares. Dentro de la comunidad de la teledetección, los bosques aleatorios y la máquina de vectores soportes son los clasificadores más utilizados. En el presente trabajo se han utilizado, además de estos dos, otros clasificadores tales como el árbol de decisión, regresión logística o aumento de gradiente, entre otros. En este contexto, en concordancia con la bibliografía, el clasificador que dio como resultado una puntuación más elevada (0.99) fue el de bosques aleatorios (RF, por sus siglas en inglés, Random Forest). Los clasificadores de máquinas de vectores de soporte (SVM, por sus siglas en inglés, Support Vector Machine) y análisis discriminante lineal (LDA, por sus siglas en inglés, Linear Discriminant Analysis) también dieron como resultado una puntuación alta, de 0.89 y 0.87, respectivamente. Sin embargo, no fueron capaces de clasificar las clases. Se obtuvieron valores de precisión, exhaustividad y valor-F mucho más altos en el caso del clasificador de bosques aleatorios.

La presencia de una clase mayoritaria hizo necesaria la utilización de técnicas de rebalanceo. De la aplicación de estas técnicas se ha podido concluir que la que dio una puntuación más alta (0.92) fue SMOTE-Tomek. Además, mediante esta técnica, se pudo obtener una precisión superior a 0.80 en todas las clases. Sin embargo, hay dos clases en las que se obtiene una precisión de uno por lo que ha hecho necesario el ajuste de hiperparámetros. La otra estrategia seguida para corregir el desbalanceo es la separación de la clase mayoritaria del resto. En este contexto, se obtuvo una puntuación de 0.99 mediane el clasificador de bosques aleatorios y una precisión superior a 0.80 en todas las clases. En este caso no se obtuvieron valores de uno o cero en ninguna clase por lo que esta línea de trabajo se podría considerar más efectiva que la aplicación de técnicas de rebalanceo.

En cuanto a posibles líneas de trabajo futuras que puedan complementar el presente trabajo se encuentra la recogida de más imágenes a través de la API de Copernicus (https://scihub.copernicus.eu/). Además, mediante la recogida de datos del Instituto Nacional de Estadística (www.ine.es), se podría aumentar el número de variables relativas demografía y población.

Dada la complejidad de las características de los datos de imágenes hiperespectrales (baja relación lineal y alta dimensionalidad), el aprendizaje profundo constituye también una excelente herramienta para la clasificación de este tipo de imágenes. En el campo de imágenes hiperespectrales se han utilizado con anterioridad las redes neuronales convolucionales (CNN, por sus siglas en inglés) para la extracción de características espectrales de los pixeles en este tipo de imágenes. Este tipo de clasificación, aunque es

exitosa, también conlleva una serie de limitaciones como son la gestión de la memoria, la preparación y carga de los datos, así como la calidad de los mismos.

La utilización de imágenes satelitales de Sentinel 2 conlleva numerosas ventajas como una mayor resolución espacial y la recolección de imágenes con nuevas bandas espectrales. Los clasificadores de bosques aleatorios y SVM han demostrado ser muy útiles para este tipo de clasificación, donde se obtienes resultados de exactitud mayores que con las imágenes recogidas con otros satélites como Landsat, de menor resolución. En contrapartida, el satélite Sentinel no ha recogido tantas imágenes como Landsat. En definitiva, el satélite Sentinel-2 ofrece nuevas oportunidades para la comunidad científica, tanto en el sector público como privado y con el creciente aumento de cobertura de tierras constituye una técnica con bastante perspectiva de futuro.