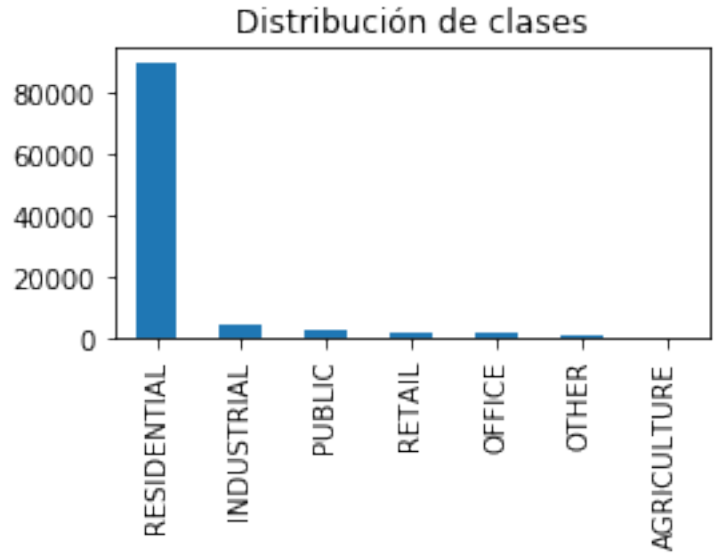
Breve resumen del trabajo desarrollado.

Análisis exploratorio de los datos

En primer lugar comenzamos explorando el conjunto y viendo lascaracterísticas más básicas (numero de instancias, número de clases, etc...). Entre estos valores, cabe destacar la distribución de las clases, ya que el problema presenta un desbalanceo bastante elevado potencialmente hacia una clase concreta (RESIDENTIAL).



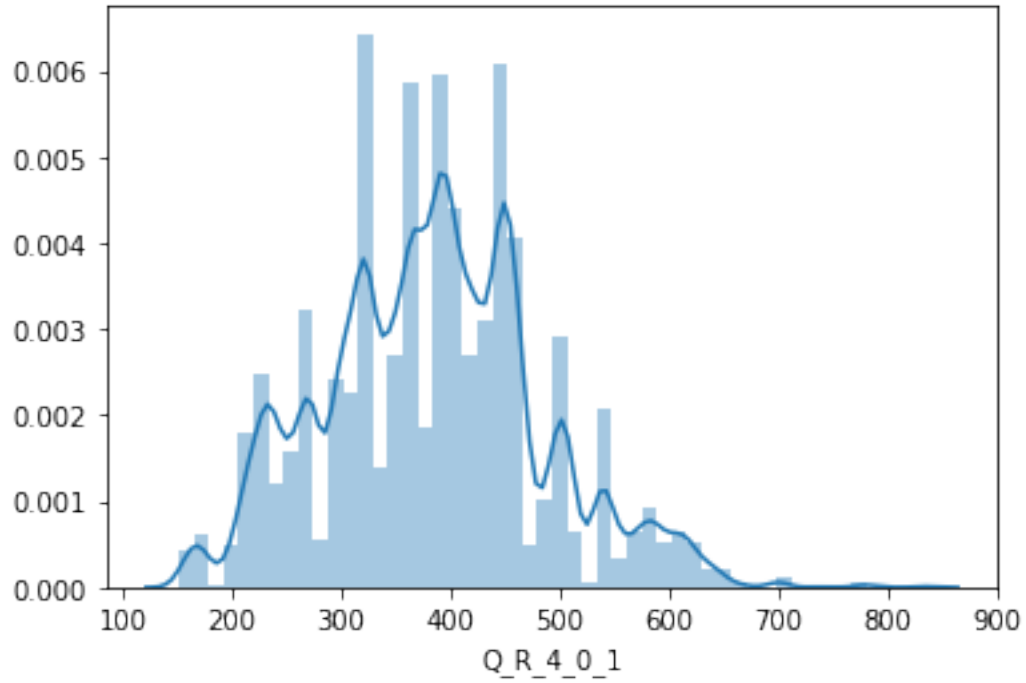
Se puede comprobar claramente en la imagen que tenemos un número de instancias muy superior de la clase *RESIDENTIAL* respecto al resto de clases. También se puede apreciar cómo hay un desbalaceo también bastante notatorio de la clase *AGRICULTURE* respecto al resto. Este problema hay que tratarlo y en posteriores secciones veremos los distintos métodos utilizados y por cuál se ha optado finalmente.

Análisis estadístico de las variables

En este apartado lo que hicimos fue obtener distintos estadísticos de las variables (media, moda, cuartiles, etc...) para analizarlos y buscar si existe algún tipo de carácteristica que nos permita entender mejor cada una de ellas. En primer lugar observamos que las variables X e Y aunque su valor estaba transformado, la relación entre ellas seguía existiendo siendo posible el ploteo del mapa concreto (se verá en posteriores secciones). Por otra parte vimos que las variables correspondientes a los colores que representan los deciles, vienen ordenadas de menos a mayor siendo *Q-{RGB/NIR}-{n1}-{0}* la variable con menor valor y *Q-{RGB/NIR}-{1}-{0}* la variable con mayor valor. Por otra parte vimos que la variable *AREA* tiene un 3^{er} cuartil ≈ 353 mientras que el máximo era ≈ 238058, lo que hizo que revisasemos esos valores tan grandes. Pudimos observar que se referían a instancias del tipo *RETAIL* o *AGRICULTURE* y que se situaban en zonas externas de la densidad del mapa, lo que nos hizo tomar la decisión de que los valores son factibles. Estudiando las variables geométricas llegamos a la conclusión de que son unas variables construidas a través de otra información ya sea utilizando PCA u otro algoritmo de reducción, por lo que de éstas variable no pudimos sacar unas conclusiones muy concretas. Por otra parte, la variable *CONSTRUCTIONYEAR* estaba bastante clara y no aporatba mucha información estos estadísticos. Por último, vimos que la variable *MAXBUILDINGFLOOR* tenía algunos valores que valían 0. En un primer lugar pensamos en cambiar este valor y sustituirlo ya sea por media o mediana o utilizando métodos más sofisticados de imputación. Finalmente decidimos asumir que 0 era un valor posible para esta variable. Referente a la variable *CADASTRALQUALITYID* simplemente se utilizó una transformación a variable numérica siguiendo un orden establecido.

Distribución de las clases

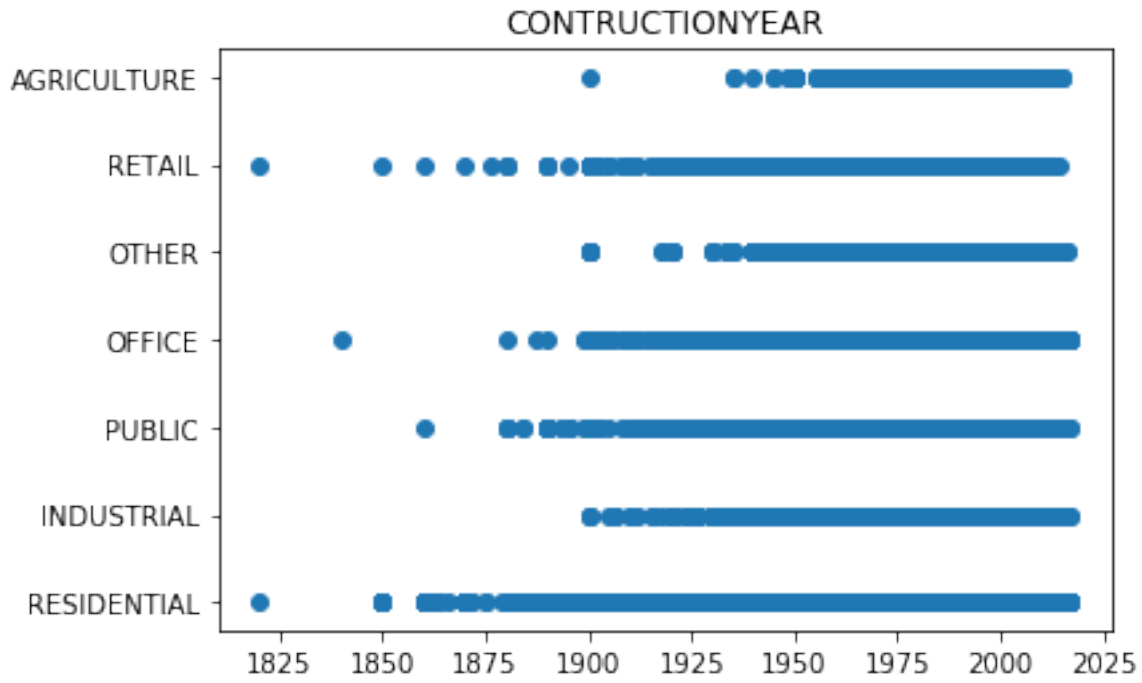
Se hizo un estudio de las distribuciones de las clases para comprobar la existencia de valores extraños, distribuciones anómalas o posibles distribuciones bimodales o multimodales. Sin embargo, comprobamos que la mayoría de las variables tenían una distribución normal y se comportaban como tal. Algunas variables como *Q_R_4_0_0* no aportaban información alguna:



En secciones posteriores veremos las transformaciones realizadas, por lo que podremos ver el procedimiento realizado para la eliminación o extracción de información de estas variables.

Relación de las variables respecto a la clase

La idea de este estudio era comprobar si había ciera relación y si podíamos sacar alguna conclusion respecto a la importancia de las variables. Por ejemplo, pudimos observar que la variable *CONTRUCTIONYEAR* entiende que cuando el valor de ésta variable es bajo, la clase no será *AGRICULTURE*.



Estudio de la correlación entre variables

Hicimos un estudio de correlación entre distintas variables y comprobamos la existencia de correlación entre variables de color entre sí y variables inflarrojas entre sí. Estuvimos estudiando qué decision tomar para finalmente decidir que no íbamos a eliminarlas por dos motivos. El primero, que tomamos estas variables como conjunto, y que si falta una de ellas, el conjunto pierde información, aunque esten muy correladas. Por otra parte, por una futura extracción de información y posterior transformación de todas estas varaibles. El resto de variables no tenían correlación.

Valores nulos

Comprobamos los valores nulos en el conjunto tanto de entrenamiento como de test. Existía un numero insignificante de nulos (27 sobre los mas de 100000 datos que tenemos) y por lo tanto no decidimos utilizar ninguna estrategia avanzada para la imputación. Vimos que los valores nulos existían en la variable *MAXBUILDINGFLOOR* y *CADASTRALQUALITYID* y que pertenecen exactamente a la misma instacia. Decidimos por tanto imputar *MAXBUILDINGFLOOR* por la mediana y *CADASTRALQUALITYID* por la moda.

Conclusión

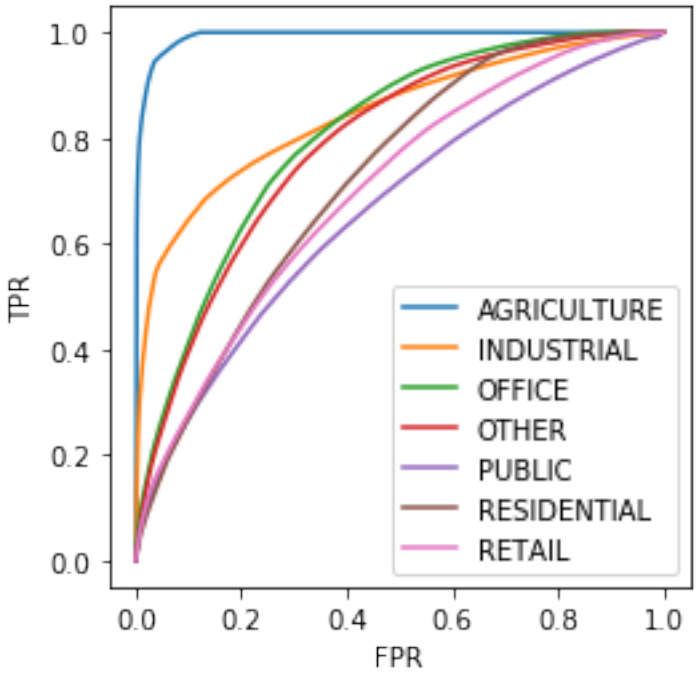
Después de todo éste análisis, entendimiento y comprensión del conjunto, pasamos al apartado de transformaciones donde buscaremos extraer información de las variables que tenemos para maximizar el aprendizaje.

Creación de variables

Generamos de forma sintética múltiples variabels que intentan facilitar el aprendizaje de los modelos explicitando el contenido de los datos en información de valor.

Transformación de los deciles RGBNir

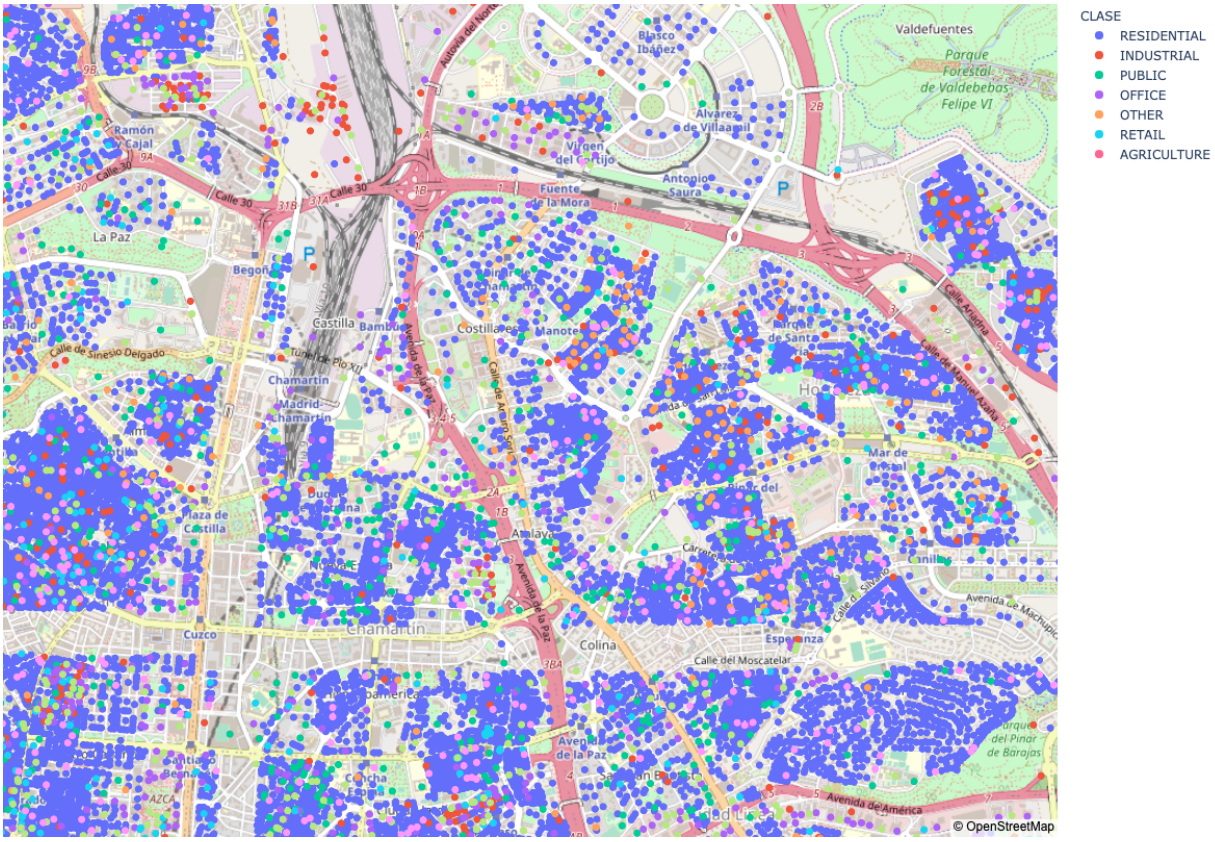
Tras un análisis de los datos observamos que sólo existen 240 configuraciones distintas para las variables que describen los deciles del espectrograma. Por lo tanto transformamos éstas a 7 nuevas variables que codifican la probabilidad de pertenecer a cada una de las clases, en el caso de haber sólo una instancia lo asignamos a AGRICULTURA, pues es la que tienes valores más dispersos. A continuación se puede ver la curva roc para cada una de estas variables, respecto a su clase, de forma individual.



De aquí vemos cómo esta variable es crítia para predecir la clase de agricultura, y tiene sentido que la fotografía area de una parcela con calificación de AGRICULTURA sera significativamente distinta al resto.

Transformación de las coordenadas a coordenadas reales

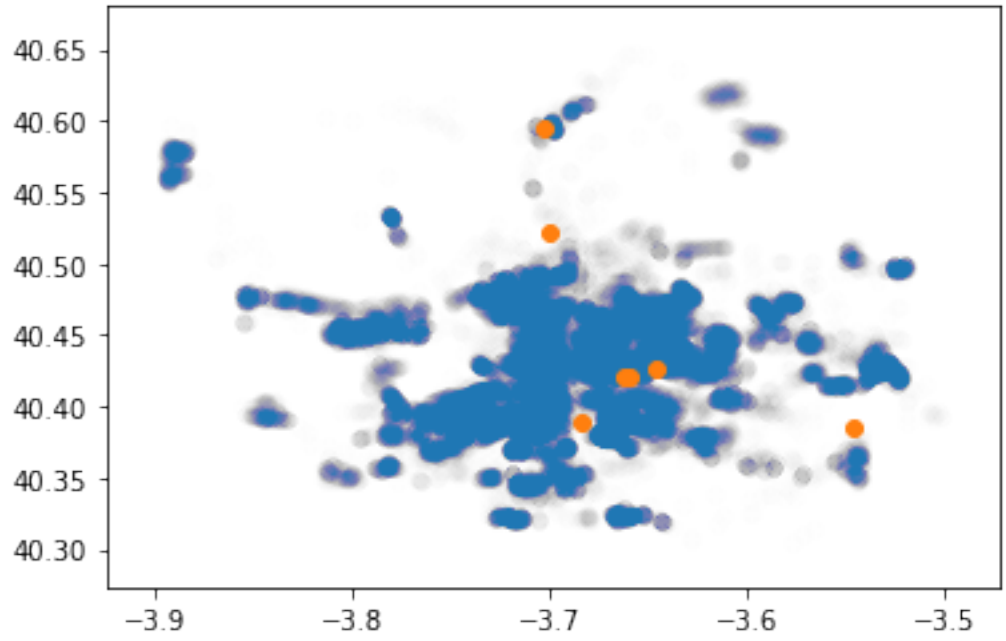
Para facilitarnos el análisis exploratorio de datos tradujimos los valores originales de las variabels X, Y a unos valores aproximados al mapa de Madrid. Para esto ploteamos los valores en un mapa, y a modo de una huella dactilar, fuimos buscando patrones características que identificábamos posteriormente en un mapa con coordenadas, en 12 zonas distintas identificamos 4 puntos próximos (para minimizar el error) y aplicamos mínimos cuadrados para aprender una traducción robusta.



El resultado no se ajusta sobre la parcela exacta por la acumulación de errores, pero nos permite identificar la zona con un dos o tres parcelas de error en el centro y algo más en la periferia.

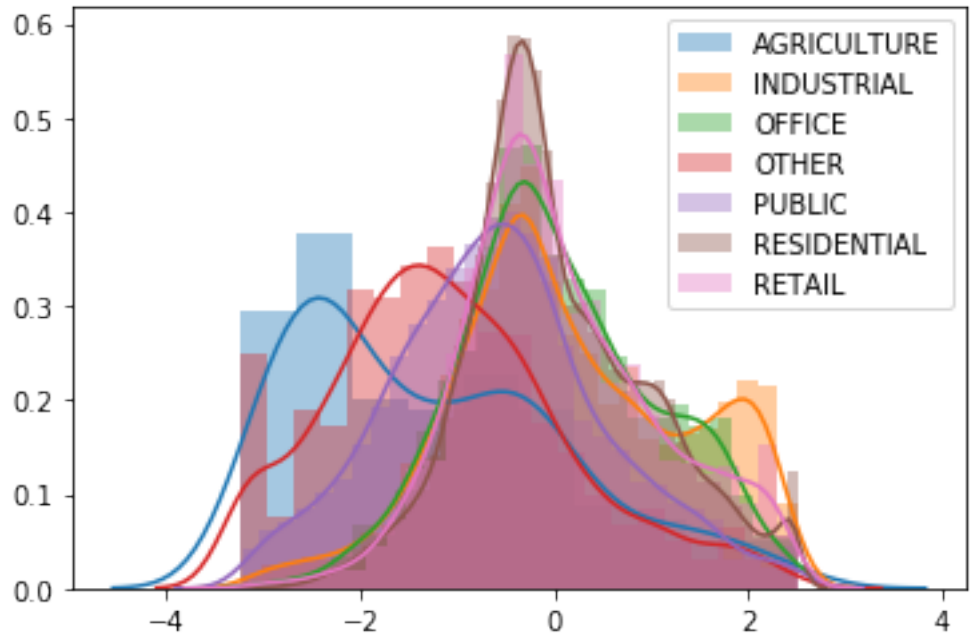
Definición de distintas áreas

Mediante un algoritmo basado en poblaciones (*differential evolution*) buscamos puntos en el mapa que definan un diagrama de Voronoi donde se maximice dentro de cada celda, la variedad en el histograma a la vez que se buscan grupos de tamaño parejo. En la siguiente imagen se puede ver en naranja los puntos que minimizan nuestra función de coste.

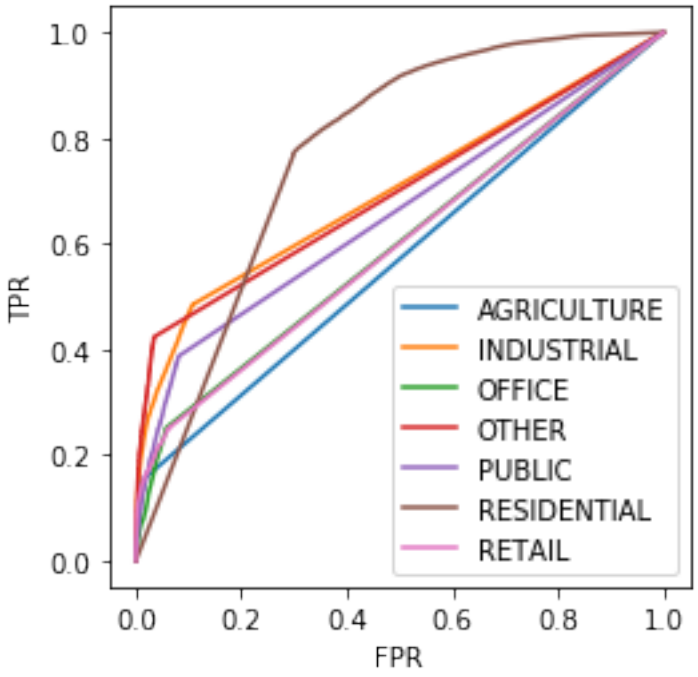


Transformación de los atributos de geometría

Existen cuatro variables con el prefijo "GEOM_" cuyo contenido puede sernos de ayuda pero la información estaba altamente ofuscada. A continuación podemos ver las distinas distribuciones de GEOM_1 según la clase objetivo



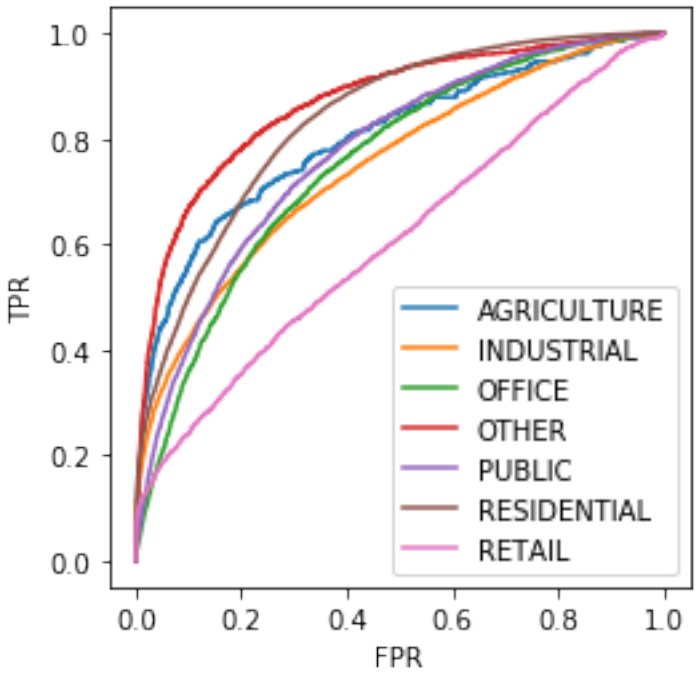
Para compensar la cola de las variables GEOM_2,3 y 4 aplicamos logaritmo y obtenemos cuatro distribuciones aproximadamente normales. A continuación la curva de ROC para un knn sólo con estas cuatro variables. Vemos cómo la información no parece despreciable.



Aplicamos dos transformaciones distintas:

- 7 variables que describen la distancia media a los cuatro vecinos más cercanos de cada una de sendas clases.
- Transformación a *probabilidades*: Corregimos la distribución con cola de las distancias mediante el logaritmo, escalamos entre 0 y 1 siendo **0** el de mayor distancia y **1** el de menor.

Con esta segunda transformación conseguimos mejorar la gráfica anterior de forma significativa como se puede apreciar en la siguiente gráfica que coge para cada instancia la mayor de las *probabilidades*.



Estas variables se crean teniendo en cuenta no el total de puntos, sino los puntos dentro del mismo área antes descrita.

Transformación de la información espacial

Las coordenadas, a juzgar por la visualización del mapa son altamente informativas, pero la granularidad de la información es tan fina que sería mejor extraer unas variables más informativas.

Para ello generamos por cada clase 5 variables, que indican la distancia euclídea al enésimo vecino más cercano de dicha clase dividido por la distancia media a los cinco vecinos más cercanos sin importar su clase. En definitiva queremos ponderar las distancias por distintas densidades en distintas zonas del mapa.

Modelos

OVA

La primera aproximación que decidimos testear ya que tenía bastante sentido era un modelo OVA, es decir, crearíamos un modelo para cada clase en la que se compara una clase determinada frente al resto y así el problema pasa a ser un problema binario. La potencia de este método es que entrenaríamos distintos modelos para cada OVA y se ajustarían concretamente a ese problema. Una vez entrenados todos los modelos, se hace una agregación con la salida y se obtiene el mejor resultado. Este modelo fué el que nos hizo obtener el mejor resultado hasta hacer el cambio de variables y el ajuste para un modelo multiclase como comentaremos a continuación.

OVO

Una de las estrategias consideradas ha sido la de One-vs-One. Esta estrategia consiste en dividir el problema original de clasificación multiclase de m clases en m*(m-1)/2 subproblemas de clasificación binaria, de forma que en cada uno de ellos se tratará de clasificar cada pareja de clases, generando de esta forma un modelo para cada pareja. Una vez realizado esto, se deberán de agregar los resultados de los diferentes clasificadores para obtener una predicción final.

Esta estrategia nos presenta una serie de ventajas teóricas que pensamos que coinciden con las necesidades del problema en cuestión. En primer lugar, es una de las técnicas recurrentes en la literatura, y se ha demostrado que puede obtener buenos resultados en los problemas de clasificación no balanceada [1]. Al tratarse nuestro problema de una problema altamente desbalanceado, pensamos que es una estrategia que nos pueda hacer obtener un buen resultado.

En segundo lugar, la descomposición del problema en subproblemas nos ofrece la ventaja de preprocesar y tratar los datos de forma diferente para cada pareja de clases. De esta forma podremos intentar encontrar el preprocesamiento óptimo para cada par. En este sentido, se ha llevado a cabo una selección y creación de características, pruebas con diferentes algoritmos de clasificación y ajuste de parámetros para cada par de clases.

La selección de características llevada a cabo en este caso consiste en una selección mediante un algoritmo greedy, el cual tratará de dar con un subconjunto de características que optimice la separación de ambas clases. Por otra parte, se han tratado de crear nuevas características a partir de aquellas que parecen más prometedoras, haciendo uso también de la técnica greedy, y aplicando transformaciones polinómicas y logarítmicas.

Para conseguir una predicción final, tendremos que decantarnos por una agregación de los resultados arrojados por cada clasificador y generar la predicción final. Para ello, se ha hecho uso de algunas de las posibles agregaciones expuestas en [2], entre las cuales podemos ver estrategias de voto simple, voto ponderado en función de la probabilidad de pertenencia a cada clase, y otras agregaciones más complejas.

Una vez ajustados los modelos y realizada la agregación, nos damos cuenta que algunos de los clasificadores binarios utilizados como base del modelo OVO tienen un mal comportamiento cuando tratan de clasificar una instancia de una clase que no coincide con ninguna de las dos clases con las que este modelo binario ha sido entrenado. Por lo tanto, la idea que tenemos es intentar darle una importancia a cada clasificador binario en función de su desempeño. Para ello, la estrategia seguida ha sido asignar un valor real a cada uno de nuestros clasificadores, el cual multiplicará a las probabilidades de pertenecer a cada clase generadas por estos, y dichos valores serán optimizados mediante el algoritmo Differential Evolution, tratando de maximizar la métrica que se le indique, en nuestro caso accuracy o f1_score.

A pesar de todas las ventajas teóricas que esta técnica nos ofrece, los resultados obtenidos no consiguen mejorar de forma sustancial los resultados obtenidos con otras técnicas más simples en lo que a número de modelos necesarios se refiere, como One-vs-All o ensambles de árboles, por lo tanto esta no ha sido nuestra solución final.

Modelo multiclase

En cuanto a algoritmos ensemble hemos considerado la aplicación de Random Forest y XGBoost, dos de los algoritmos basados en ensemble de árboles de decisión más reconocidos en problemas de clasificación.

Los resultados que obtenemos en una primera experimentación con ambos algoritmos nos revelan el mejor funcionamiento, con una ventaja sustancial, del algoritmo XGBoost aplicado sobre nuestro conjunto de datos, por lo que en la siguiente fase nos centraremos en el y trataremos de optimizar su funcionamiento.

Ante el elevado número de datos de los que disponemos, y el extenso tiempo de cómputo empleado por XGBoost, la opción de hacer una búsqueda de hiperparámetros exhaustiva es descartada. Optamos por hacer una búsqueda aleatoria de un número fijado de combinaciones de hiperparámetros.

Balanceo: Asignación de pesos a clases para compensar la falta de balanceo

Para el balanceo de las clases utilizamos una estrategia de aprendizaje semisupervisado, dado que el conjunto de test presenta una distribución distinta a la del conjunto de entrenamiento hacemos la siguiente analogía común en la Ciencia de datos aplicada a Medicina:

- El conjunto de test, cuyas etiquetas deconocemos, representa la distribución real, en nuestra analogía la población general.
- El conjunto de entrenamiento, cuyas etiquetas sí conocemos, representa una distribución sesgada, en nuestra metáfora sería la gente que acude al hospital con una sintomatología concreta. A esta población se le realiza una prueba médica que debe interpretar un médico.

Es inmediato ver cómo en la población que acude al médico es más probable que esa prueba dé positivo, pues es una probabilidad condicionada a la sintomatología que presenta. ¿Cómo podríamos estimar qué proporción de la población presenta esa enfermedad, es decir, cuál es la probabilidad no condicionada? Como sabemos la proporción con la que sobreestimamos o infraestimamos cada una de las clases en el conjunto de entrenamiento podemos predecir el conjunto de test, compensar nuestro sesgo y estimar la probabilidad (o en nuestro caso el número de ejemplos de cada clase) en la población objetivo.

De igual forma nosotros, mediante validación cruzada podemos saber cómo nos equivocamos aunque no sepamos dónde, aplicamos este ajuste y estimamos cómo de sobrerrepresentada está una clase en el conjunto de entrenamiento.