CAJAMAR UNIVERSITYHACK 2020

RETO MINSAIT LAND CLASSIFICATION

Equipo *Anacongas*

Escuela Superior de Informática de Ciudad Real (UCLM)

- Rubén Márquez Villalta (Ruben.Marquez@alu.uclm.es)
- Diego Pedregal Hidalgo (Diego.Pedregal@alu.uclm.es)
- Antonio Manjavacas Lucas (Antonio.Manjavacas@alu.uclm.es)

Coordinador: Francisco Pascual Romero Chicharro (FranciscoP.Romero@uclm.es)

1. RESUMEN DEL TRABAJO

Con este trabajo presentamos nuestra participación en el reto *MINSAIT LAND CLASSIFICATION* del *Cajamar UniversityHack 2020*. En esta edición, el desafío planteado es la elaboración de un modelo predictivo orientado a la clasificación automática de tipos de suelo en base a características catastrales y datos extraídos de imágenes vía satélite.

El contenido adjunto incluye los resultados obtenidos y medios empleados a la hora de abordar este proyecto.

2. RESUMEN DEL ANÁLISIS EXPLORATORIO Y LAS CONCLUSIONES

A la hora de abordar el desafío, inicialmente se realizó un análisis de los datos, estudiando tanto su naturaleza como su distribución. Algunas conclusiones obtenidas tras este análisis inicial de los datos fueron las siguientes:

1. Baja cantidad de registros nulos.
2. Gran desbalanceo en el número clases presentes en el dataset, especialmente entre registros RESIDENCIALES y NO RESIDENCIALES, con una amplia mayoría de RESIDENCIALES, hecho que condicionaría nuestra forma de abordar el reto.
3. Dos tipos de datos claramente diferenciados: aquellos referentes a la imagen (canales de color) y otros asociados a características de los edificios (año de construcción, metros cuadrados, etc.).
4. Campos pendientes de ser categorizados, como el identificador de calidad catastral (*CADASTRALQUALITYID*) o la propia clase de suelo (variable objetivo a predecir).

3. RESUMEN DE MANIPULACIÓN DE VARIABLES Y ARGUMENTACIÓN

Sobre esta base, se llevaron a cabo las labores de preprocesamiento y limpieza de los datos, añadiendo campos adicionales que consideramos relevantes con respecto a los colores de la imagen, tales como media, desviación típica, máximos y mínimos de los diferentes deciles.

Una vez llevado a cabo el preprocesamiento de los datos, y dado el alto desbalanceo entre clases RESIDENCIALES y NO RESIDENCIALES, se decidió dividir el proceso de predicción en dos fases diferenciadas:

1. Un primer modelo (**árbol de decisión**) llevaría a cabo la clasificación entre edificios RESIDENCIALES y NO RESIDENCIALES (**clasificación binaria**).
2. Un segundo modelo (**random forest**) trataría de etiquetar los clasificados como NO RESIDENCIALES en sus correspondientes clases (**clasificación multiclase**),

incluyendo aquellos RESIDENCIALES que no habían sido correctamente clasificados por el modelo anterior.

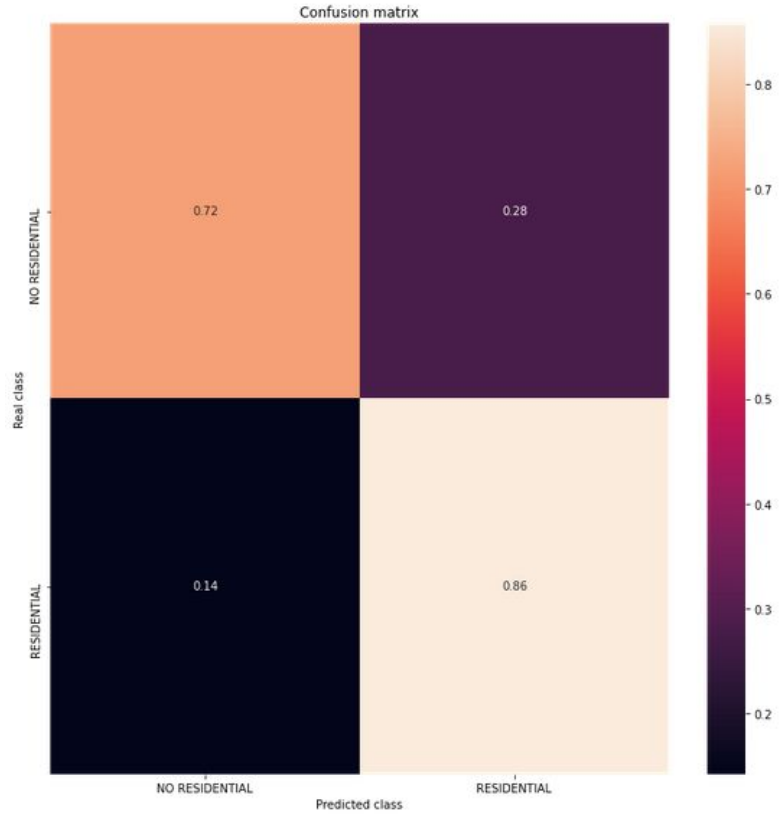
a) MODELO 1: RESIDENCIAL *vs* NO RESIDENCIAL

Con respecto al primero de los modelos (RESIDENCIAL vs NO RESIDENCIAL), inicialmente se llevó a cabo la separación del dataset en datos de entrenamiento (*train*) y validación (*test*), aplicando al mismo tiempo sobremuestreo (*oversampling*) con el fin de equilibrar las instancias de ambas clases. A su vez, los datos presentes en el conjunto de validación fueron igualmente equilibrados con respecto a sus clases.

Las características seleccionadas y empleadas por este primer modelo fueron seleccionadas y ajustadas de forma iterativa tras comprobar sus correspondientes relevancias en la predicción.

Una vez preparados los datos, se llevó a cabo la parametrización (*tuning*) del modelo: un proceso iterativo que supuso un total de 7 horas de ejecución, permitiendo el ajuste de hiperparámetros a utilizar.

Finalmente, los resultados obtenidos por este primer modelo (tras múltiples ajustes, tanto de los hiperparámetros como de las características empleadas) alcanzaron un *accuracy* medio del 79%, obteniéndose los resultados mostrados en la Figura 1.



	precision	recall	f1-score	support
0	0.83	0.72	0.77	3911
1	0.75	0.86	0.80	3911
accuracy			0.79	7822
macro avg	0.79	0.79	0.79	7822
weighted avg	0.79	0.79	0.79	7822

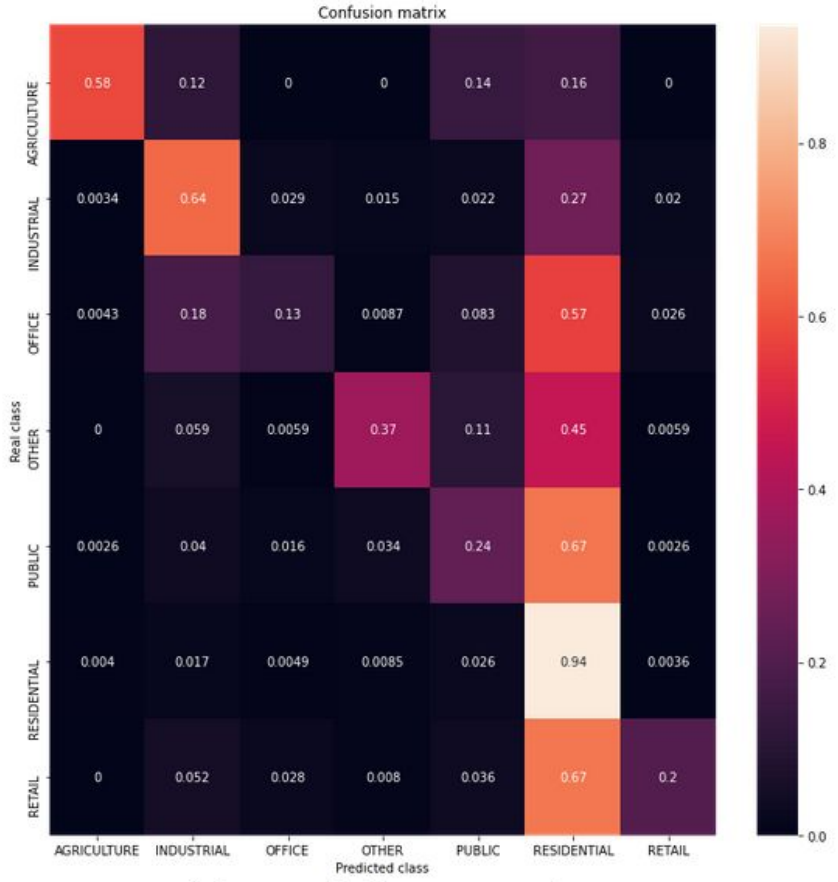
Figura 1. Resultados del primer modelo.

b) MODELO 2: NO RESIDENCIALES

Una vez predicha la naturaleza residencial de nuestro conjunto de datos, se extrajo el conjunto de datos etiquetado como NO RESIDENCIAL, haciéndolo pasar por un segundo modelo de clasificación encargado de predecir de qué tipos de edificios no residenciales se trataban.

Al igual que en el caso anterior, las características empleadas por el modelo fueron consideradas basándonos en su relevancia a la hora de predecir los datos de validación. También se llevó a cabo un proceso de parametrización similar al anterior, adecuado a los hiperparámetros de los random forests, esta vez con una duración de aproximadamente 10 horas.

Tras el entrenamiento, esta vez utilizando validación cruzada, los resultados obtenidos en esta clasificación multiclase fueron los detallados en la Figura 2.



	precision	recall	f1-score	support
0	0.66	0.58	0.62	43
1	0.76	0.64	0.70	592
2	0.42	0.13	0.20	230
3	0.58	0.37	0.45	170
4	0.42	0.24	0.30	379
5	0.72	0.94	0.82	2227
6	0.65	0.20	0.31	250
accuracy			0.70	3891
macro avg	0.60	0.44	0.49	3891
weighted avg	0.67	0.70	0.66	3891

Figura 2. Resultados del segundo modelo.

4. JUSTIFICACIÓN DE LA SELECCIÓN DEL MODELO

La elección de ambos modelos, tanto del árbol de decisión empleado en la clasificación binaria (RESIDENCIAL vs NO RESIDENCIAL), como del *random forest* utilizado en la clasificación multiclase (NO RESIDENCIALES) se llevó a cabo tras probar la eficacia de árboles de decisión, *random forests* y redes neuronales en ambos casos, siendo los modelos elegidos los que mejores resultados arrojaron.

Aun así, nótese cómo la robustez de los árboles de decisión y *random forests* frente a datos altamente desbalanceados tiende a ser mayor en comparación con las redes neuronales, que presentan más dificultades asociadas al *overfitting* (sobreentrenamiento).