**Mejora de la red ciclista de Madrid**

Cuando dices que vas en bici por Madrid, ¿no te preguntan: “no te da miedo”?, ¿Vas por la carretera en bici?, pero… irás por carril bici, ¿no?

A todos los que hemos ido en bici por Madrid, seguro que nos han hecho este tipo de preguntas y por ellas por las que nace nuestro proyecto: “Mejora de la red ciclista de Madrid”

Y os preguntaréis, ¿en qué consiste vuestro proyecto? Nuestro proyecto trata de predecir la forma más segura para movernos de un punto A a un punto B y, para ello, hemos utilizado Google Street View, para analizar los obstáculos que nos podemos encontrar en la calle cuando circulamos en bici.

En primer lugar, se han descargado unas 200 imágenes en formato RGB colors de Google Street View, las cuales se han redimensionado para que todas tengan el mismo tamaño y se ha identificado los elementos de cada una de ellas (carretera, acera, edificio, coche, camión, persona, vegetación, tren, vaya, semáforo, cielo…) mediante el uso de las redes neuronales de MobileNetV2 y Xception65.

Una vez cada elemento de la imagen se ha identificado, se cuantifican, es decir, si la imagen sólo muestra carretera y cielo, ambos elementos tienen un peso del 50%. Adicionalmente, se ha añadido el tipo de vía a cada imagen, ya que se disponía de esta información, y se categorizado como numérica, indicando con el valor 1 si es ciclable y 0 si no es ciclable. Los tipos de vías que hay son:

* Itinerario recomendado: 1
* Senda peatonal ciclable: 0
* Ciclocarril: 1
* Carril bici unidireccional en calzada: 0
* Ciclocarril+carril bus: 0
* Calle 20

Para tener un mayor número de imágenes, se ha aplicado data aumentation…

Adicionalmente, el proceso anteriormente descrito de redimensionado, segmentación y data aumentation, también se ha aplicado a las imágenes en blanco y negro (BW), que previamente se han convertido, para poder analizar resultados.

Una vez obtenidas las imágenes segmentadas en formato RGB y BW, se ha divido cada conjunto de imágenes en 3 set diferentes:

* Set de entranamiento: 65% del total
* Set de test: 20% del total
* Set de validación: 15% del total

A continuación, para cada set de entrenamiento (RGB y BW), se ha aplicado modelos de clasificación de machine learning. Tras el análisis de varios modelos, se han optado por los siguientes:

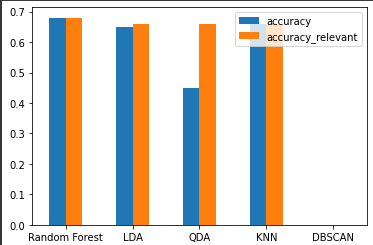
* **Random Forest**: es un algoritmo de aprendizaje automático supervisado basado en un conjunto de árboles de decisión.
* **LDA** (Lineal Discriminant Analysis): es un método de clasificación supervisado de variables cualitativas en el que dos o más grupos son conocidos a priori y nuevas observaciones se clasifican en uno de ellos en función de sus características.
* **KNN**: es un algoritmo de aprendizaje supervisado, es decir, que a partir de un juego de datos inicial su objetivo será el de clasificar correctamente todas las instancias nuevas
* **QDA** (Quadratic Discriminant Analysis): es un método de clasificación supervisado que se encuentra en un punto medio entre el método no paramétrico KNN y los métodos lineales LDA y regresión logística.
* **DBSCAN**: es un algoritmo de clúster o agrupamiento, no supervisado, basado en la densidad que puede ser utilizado para identificar clústeres de cualquier forma en un conjunto de datos que contiene ruido y valores atípicos.

Cada uno de los modelos y conjuntos descritos, se han aplicado sobre 2 conjuntos de datos:

1. Utilizando todos los elementos (features) de las imágenes
2. Utilizando los elementos (features) más significativos de las imágenes. Para ello, se han seleccionado aquellas features cuyo percentil 25% es superior al 50%, obteniéndose un conjunto de datos con las siguientes features (ordenadas por mediana y desviación típica): acera, coche, cielo, vegetación, carretera, edificio

**Imágenes RGB**

Una vez ajustados los parámetros de cada modelo y analizado los resultados, la precisión que obtenemos para las imágenes en RGB para cada modelo y conjunto de datos es:

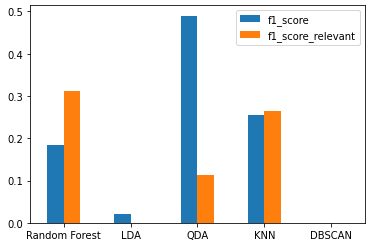


Donde las columnas azules están calculadas con el conjunto 1 y las columnas naranjas con el conjunto 2. Como podemos observar, con ambos conjuntos de datos obtenemos prácticamente la misma precisión para los modelos Random Forest, LDA y KNN. Para el modelo QDA, se obtiene mejor precisión sobre el conjunto de datos 2. Y para el modelo DBSCAN, no se obtiene resultado.

Analizando un poco más en detalle cada modelo y conjunto de datos, el clasificador f1-score, que representa el número de positivos verdaderos (vía ciclable y clasificada como ciclable), positivos falsos (vía ciclable y no clasificada como ciclable), negativos verdaderos (vía no ciclable y clasificada como ciclable), y negativos falsos (vía no ciclable y clasificada como no ciclable), permitirá aceptar o descartar los modelos de clasificación descritos acorde a los siguientes valores:

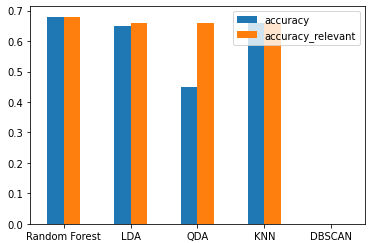
* ≥0.9: muy bueno
* [0.8, 0.9): bueno
* [0.5,0.8): aceptable
* <0.5: no aceptable

En la gráfica siguiente, podemos ver que sólo el modelo QDA es aceptable para el conjunto de datos 1:

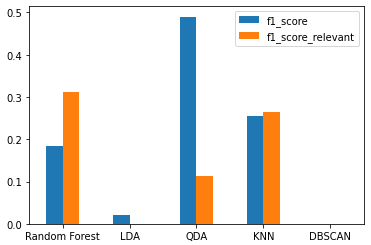


**Imágenes BW**

En el caso de las imágenes BW, los resultados de precisión para cada modelo y cada conjunto de datos son muy similares a los obtenidos con las imágenes RGB, como se muestra a continuación:



Y los resultados de f1-score, también son similares a los obtenidos con las imágenes RGB, como se muestra en la siguiente gráfica:



**Conclusiones**

Tras el análisis realizado sobre el conjunto de las imágenes RGB y BW, se puede concluir que:

* Para predecir si una imagen es ciclable o no, es indiferente aplicar modelos de clasificación sobre imágenes en RGB o BW
* Sobre el conjunto de 200 imágenes, la precisión de los modelos de clasificación sobre el conjunto de todas las features es mejor que sobre el conjunto de las features más significativas. Habría que hacer un estudio sobre una muestra más grande
* Dado el conjunto de 200 imágenes, el modelo de clasificación que mejor funciona es el QDA, dado que es modelo híbrido entre los modelos LDA y KNN