Trabajo Final: Utilizando Random Forests

Diseño de Sistemas Inteligentes

Carlos Córdoba Ruiz
UCLM Ciudad Real
Carlos.Cordoba@alu.uclm.es

INTRODUCCIÓN

El objetivo de este trabajo es saber evaluar el rendimiento de un clasificador y compararlo con otras alternativas. Un segundo objetivo es adentrarse en el diseño de nuevos clasificadores a partir de otros existentes.

El clasificador que se va a implementar va a ser el que está descrito en el artículo adjuntado en el Campus Virtual, y se van a realizar distintos hitos evaluando y comparando su rendimiento con el que aparece descrito en el artículo.

1 Procesado de los datos

El primer paso antes de comenzar con la creación del modelo descrito ha sido realizar un preprocesamiento sobre los datos, para permitir la manipulación de éstos mismos. Para ello se ha reutilizado el mismo Código del primer trabajo de la asignatura.

El preprocesamiento realizado ha sido bastante simple, primero se parametrizaron las variables no numéricas y posteriormente, al igual que en el artículo se cargaron en un dataframe una muestra de 100.000 registros para utilizarlos en el estudio realizado.

El conjunto de datos se separó en dos, dejando así un dataset con todas las variables que se van a utilizar para entrenar el modelo y otro dataset con una única columna en las que estaba almacenada la variable <<mode_main>> que es la variable que se espera predecir.

2 Hito 1

La primera tarea realizada es evaluar el clasificador asignado. Para ello se deben emplear los siguientes dos índices para medir su rendimiento:

-Precisión (Accuracy). Es la proporción de aciertos sobre el total de la muestra.

-Sensibilidad. Es la proporción de aciertos dentro de cada alternativa. Por ejemplo, que proporción de viajes realizados en el modo coche son correctamente clasificados, y así con todas las alternativas. En este problema hay cuatro alternativas.

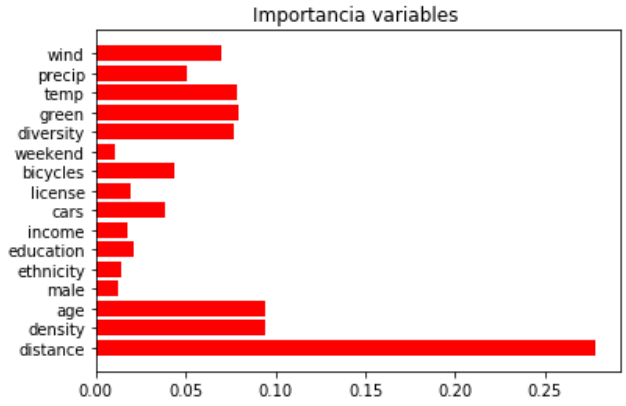
Para mejorar las estimaciones de los indicadores se realizaron dos tareas: una de ellas, balanceamiento de datos, ya que hay una gran diferencia en el número de viajes realizados de distinto tipo, por ejemplo, hay muchos mas registros de viajes realizados en coche que, por ejemplo, caminando. Otra de las tareas realizadas fue la técnica de 10-fold cross-validation, en la que se divide la muestra en 10 submuestras y se crea un nuevo conjunto de entrenamiento 90% train y 10% test, y la estimación de la precisión será la media de los 10 resultados obtenidos.

A la vez que se generan las submuestras del cross-validation se aplica el balanceamiento de los datos aplicando la formula explicada en el enunciado de la práctica, pero con el nº de registros de cada submuestra del cross-validation.

```
Media de los valores: 0.71097
Media de la sensibilidad : 0.6718964773763896
```

El resultado muestra que el modelo que se ha generado y entrenado tiene una precisión de 71,09% de precisión, frente al cerca del 90% que tiene el modelo descrito en el artículo, hay un 20% de diferencia.

Al haberse usado los mismos valores para crear el modelo y utilizado los métodos de cross-validation y balanceo de datos, podría deberse a que en el apartado de procesado de los datos no se han parametrizado bien las variables no numéricas.



Respecto a la importancia de las variables, se puede ver cómo la variable distancia tiene algo mas del 25% de relevancia a la hora de realizar un tipo de viaje u otro.

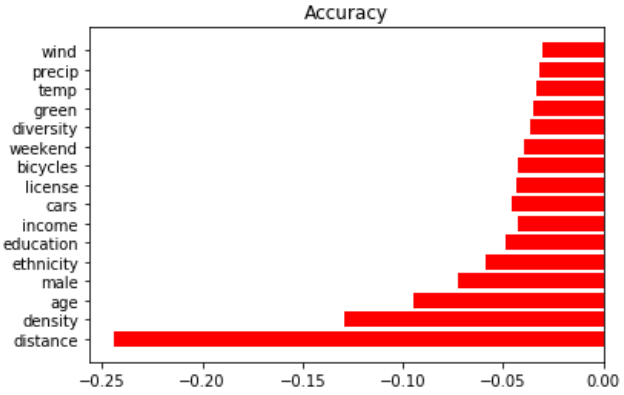
3 Hito 2

En este Segundo Hito, se tratará de medir la importancia de las variables para destacar la importancia que estas tienen para el rendimiento del clasificador.

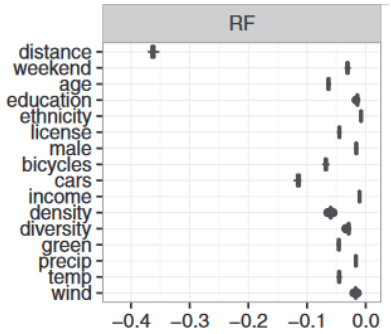
La idea clave del métodoes que si una variable no fuese importante para el problema de clasificación, si cambiáramoss sus valores aleatoriamente, el rendimiento del clasificador,medido a través de la precisión o de la sensibilidad, no variaría sustancialmente.

Para ello se ha realizado este hito en dos partes, uno para medir cómo varia la precision de cada variable respecto con la precision del modelo calculada en el hito anterior, y luego la sensibilidad de cada variable en función del tipo de viaje.

Para medir la precision de cada una de las variables y cómo influye el proceso realizado ha sido el siguiente: se ha cogido cada una de las variables del modelo y se han permutado los valores de esa variable en cada uno del 10-fold-cross-validation, y calcular la precision obtenida y restarla a la precision del modelo calculada en el hito 1 para posteriormente graficarlo y mostrar gráficamente cómo influye realmente esa variable para la precision del modelo.



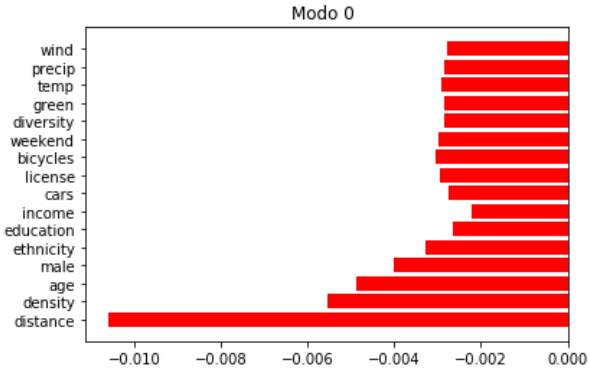
La figura muestra cómo la variable distancia es la que más afecta a la precision del modelo, seguida de la variable density. A continuación se muestra además la gráfica obtenida del artículo que muestra la precision de cada una de las variables cómo influyen al modelo. Se puede ver que la única variable que se puede observar que destaca un poco más es distance, el resto de variables no muestran la misma similitud ni se puede observar que important tanto, esto puede deberse a que la precision de nuestro primer clasificador solo obtuvo un 70% de precision, y el del artículo tiene un 90%, este error puede ser que se siga propagando.

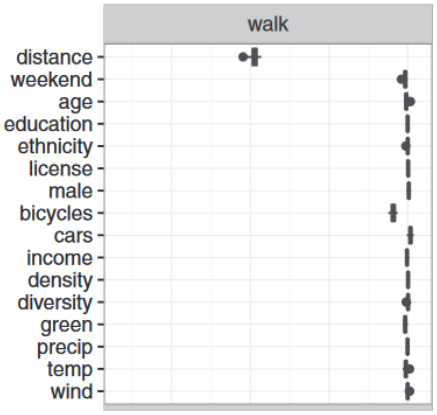


Para medir la sensibilidad de cada una de las variables es algo distinto que en el caso anterior, ahora hay que filtrar por cada modo de transporte, entonces a la hora de permutar el valor de las variables únicamente hay que permutar las variables que pertenezcan al modo de viaje que se está analizando, y una vez hecha esta operación, se hace una diferencia con la sensibilidad calculada en el hito 1 para ver cómo varía esa variable respecto el modo de viaje que se está analizando y afecta a la sensibilidad del modelo.

Los resultados obtenidos del análisis son los siguientes:

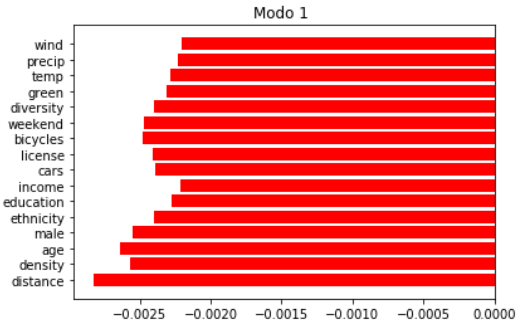
Resultados de la sensibilidad del modo de viaje walk fue la siguiente:

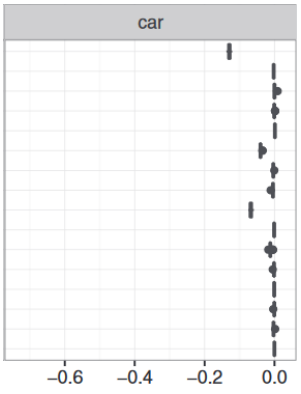




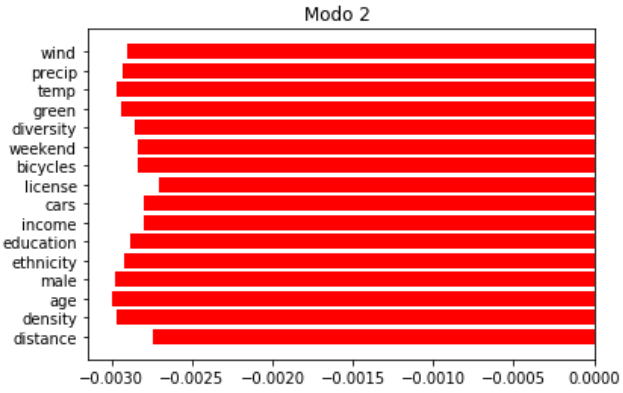
Se observa que la variable distance al igual que en la figura del artículo destaca respecto al resto, que se mantienen cercanas al 0, pero, eso es debido a la escala de la gráfica creada, realmente no hay una gran diferencia en la sensibilidad de la variable distance conforme el resto, de 0.010 a 0.004.

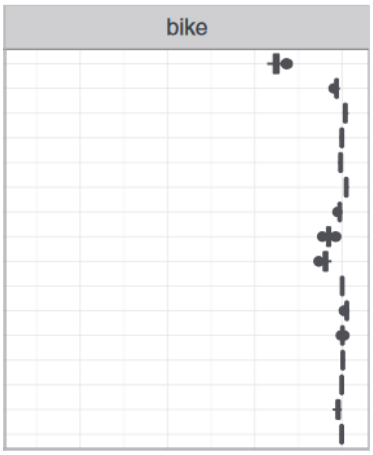
Resultados de la sensibilidad del modo de viaje car fue la siguiente:



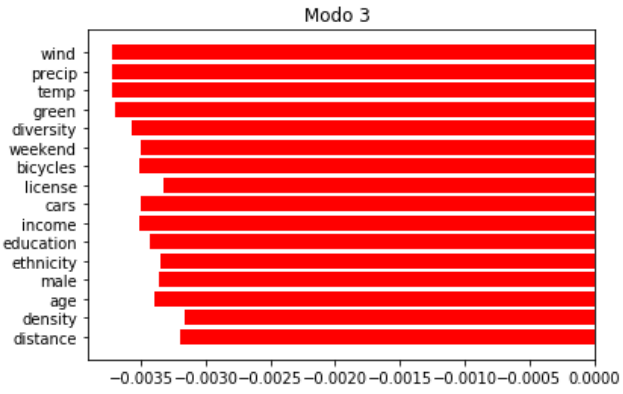


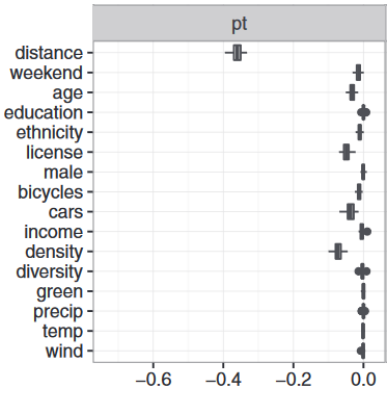
Resultados de la sensibilidad del modo de viaje bike fue la siguiente:



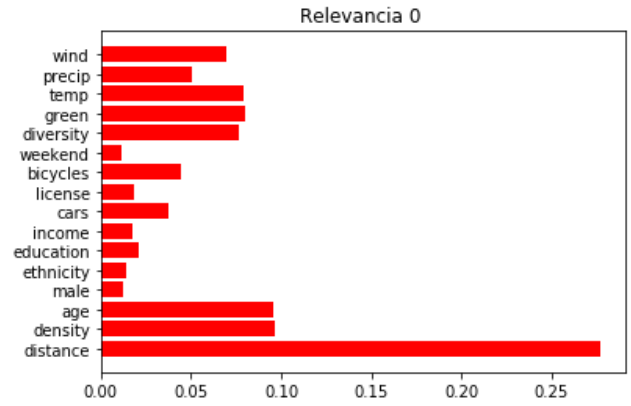


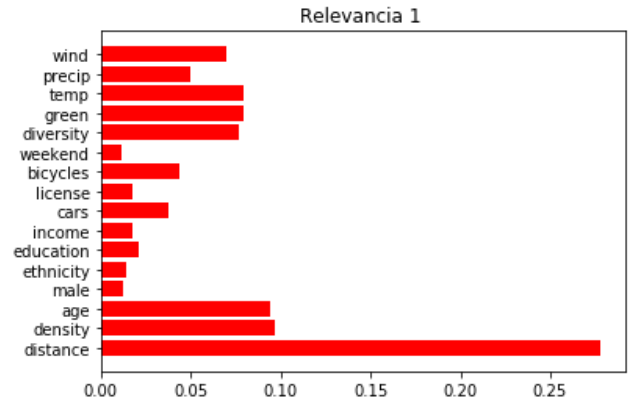
Resultados de la sensibilidad del modo de viaje pt fue la siguiente:

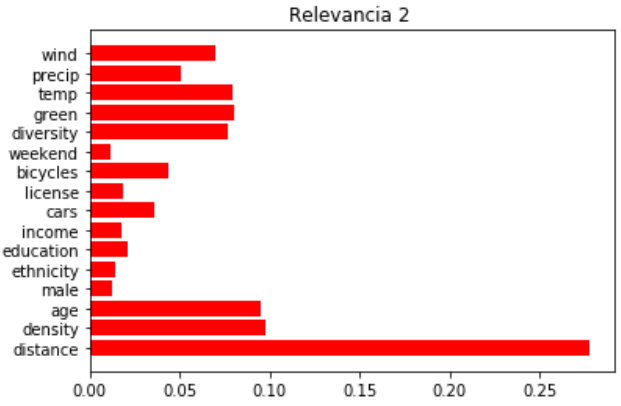




Se puede observar que en general en cada paso, los resultados no se adaptan con los resultados provistos por el artículo, ya que ninguna variable influye mucho en la sensibilidad del modelo, puede deberse a algún problema realizado a la hora de realizar los calculos, para ello se han generado además distintas gráficas mostrando la importancia de las variables del modelo en cada caso para ver la relevancia de estas:







Las gráficas muestran una gran similitude y no parece que varíen mucho de un tipo de viaje a otro, por lo tanto, es possible que se trate un error del Código realizado que no se calculen correctamente la sensibilidad.

La mayor dificultad encontrada en este hito del trabajo ha sido la de similar el mismo funcionamiento en el Código y que funcionase correctamente

4 Código

El Código se puede encontrar en la carpeta adjuntada con el documento o en el repositorio: https://github.com/CarlosCordoba96/FinalWork-DSI