exp_propio

Piñera - Saint-Nom - De Diego

2023-08-22

Ejercicio 3

Introducción: En el ejercicio 1 realizamos una transformación para modificar ciertas categorias de la columna "NObeyesdad" (Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II) por 0 indicando que el individuo no tiene obesidad y (Obesity Type I, Obesity Type II y Obesity Type III) por 1 indicando que el individuo tiene obesidad. Ahora queremos explorar cómo la optimización del punto de corte utilizado para clasificar individuos como "obesos" o "no obesos" puede influir en el rendimiento del modelo.

Para la transofrmación de los datos modificados con el criterio de corte, clasificamos el "Corte 1", siendo el corte más restrictivo (solo Obesity Type III considerado como obesidad), hasta "Corte 5", siendo el corte mas abarcativo (considerando a Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II y Obesity Type III como obesidad).

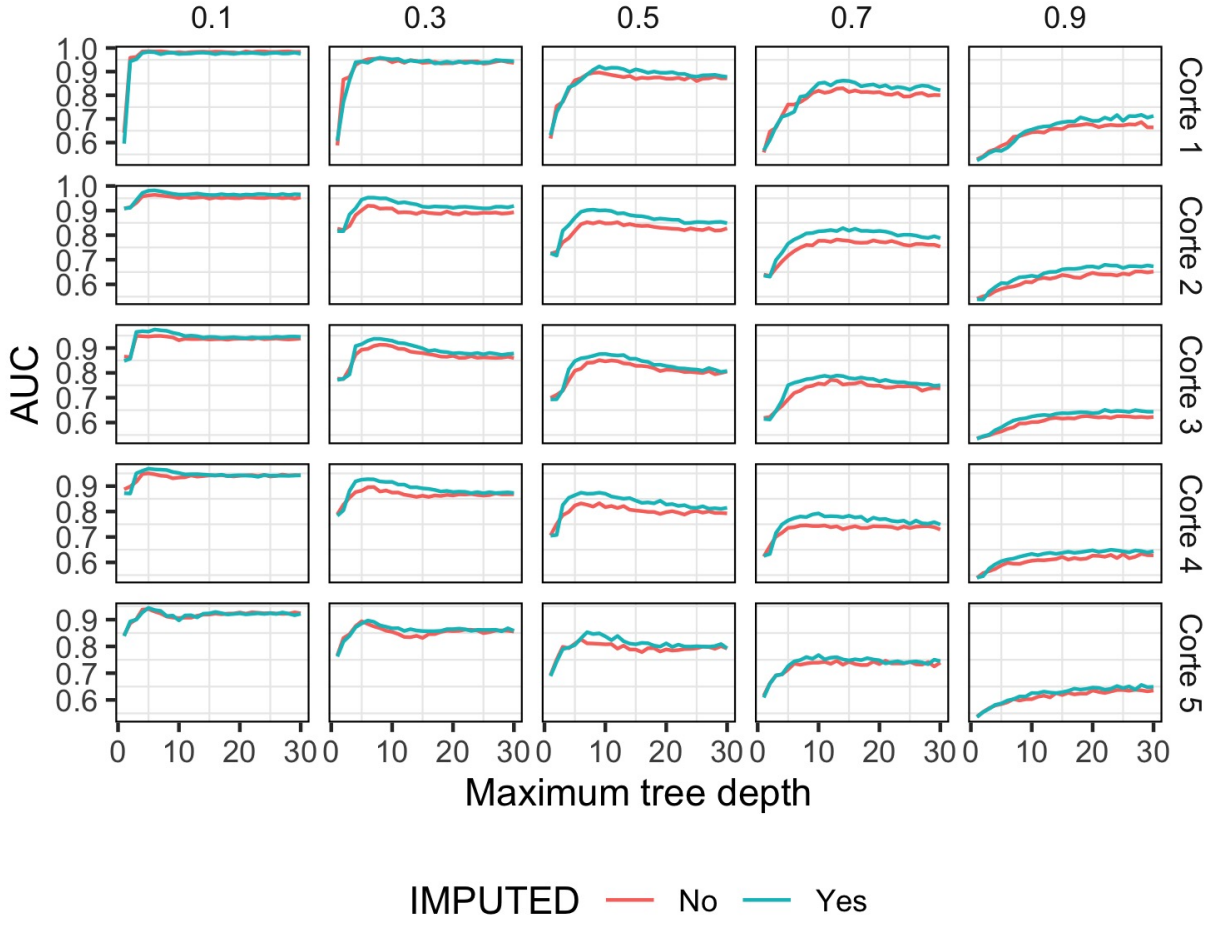
Luego de realizar las transformaciones probando distintos cortes de clasificación, vemos la frecuencia de ceros y unos en cada una.

Hipótesis: Nuestra hipótesis es que al ajustar este corte, si la cantidad de ceros y unos llegan a un numero parecido, podríamos mejorar la capacidad del modelo para predecir mejor. Sin embargo, también consideramos que cambios en el corte podrían aumentar los errores de clasificación. Nuestro objetivo es examinar cómo estos ajustes afectan las métricas de evaluación del modelo.

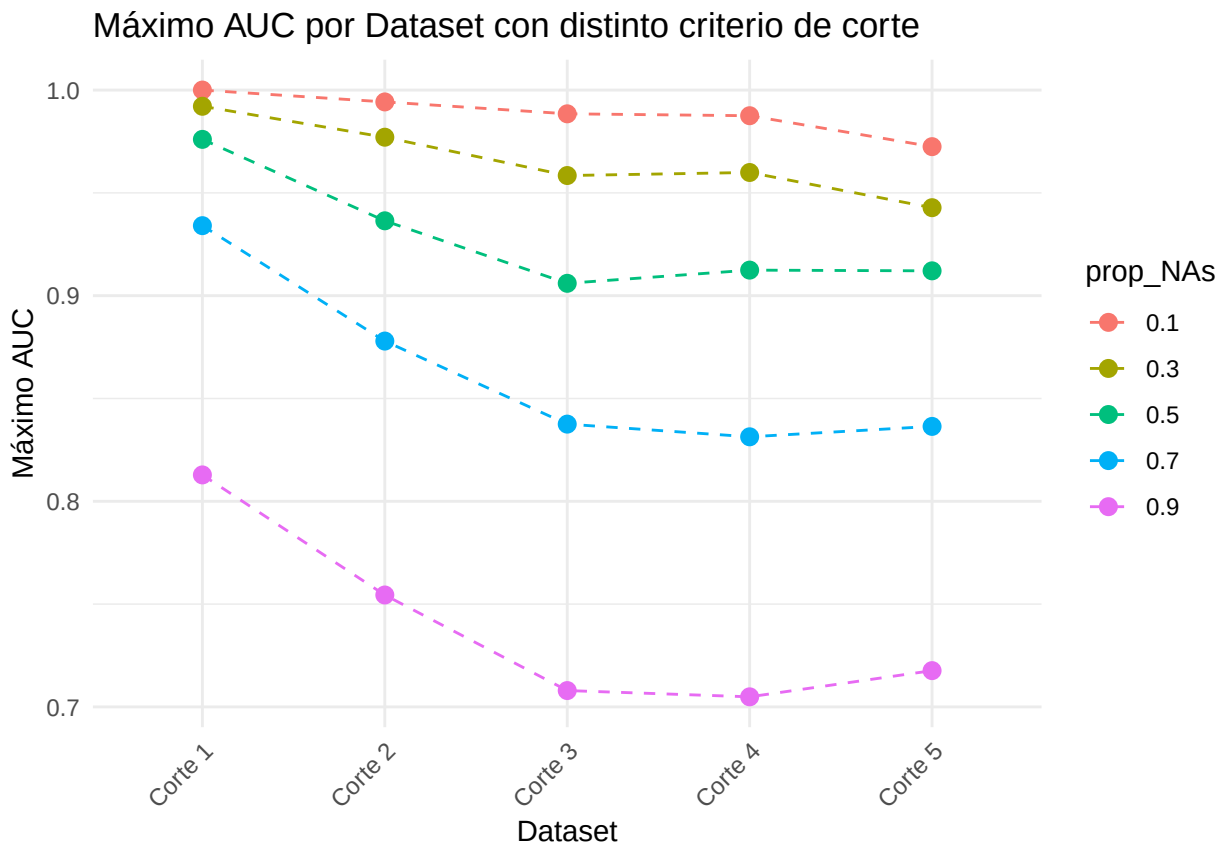
Gráficos: La tabla que viene a continuacion es la proporcion de casos de obesidad positivos segun el corte utilizado. A su vez, En el siguiente gráfico, examinamos cómo varía el valor máximo del AUC en diferentes conjuntos de datos en función de los umbrales de corte utilizados para asi tener una intucion de que esperar. También buscamos visualizar cómo estos umbrales de corte afectan a conjuntos de datos con distintos niveles de valores faltantes. En el caso del Corte 1, solo se consideran como "obesos" (1s) a los individuos con obesidad extrema, mientras que al resto se les clasifica como "no obesos" (0s). Por otro lado, en el otro extremo de Corte 5, incluso los individuos con un nivel 1 de sobrepeso son considerados como "obesos".

```
## Frecuencias para corte 1:

##
##    0    1
## 1787  324

##
## Frecuencias para corte 2:
```

```
##
##    0    1
## 1490  621

##
## Frecuencias para corte 3:

##
##    0    1
## 1139  972

##
## Frecuencias para corte 4:

##
##    0    1
##  849 1262

##
## Frecuencias para corte 5:

##
##    0    1
##  559 1552
```



Observaciones: Luego de realizar los experimentos, observamos un patrón interesante: el modelo obtuvo su mejor rendimiento cuando el punto de corte fue más restrictivo, clasificando a un grupo más pequeño como "obeso". En este contexto, notamos que esta mejora podría relacionarse con la presencia de patrones extremos en los estilos de vida de los individuos clasificados como obesos. Esta agrupación de casos extremos podría ser más coherente con la forma en que el modelo está aprendiendo y generalizando, lo que a su vez mejora su capacidad predictiva para estos casos específicos.
Cabe a destacar que no hay gran diferencias entre que sea imputados los datos faltantes o no a lo que vimos enel experimento 1, cumple con los mismos patrones y tendencias, mostrandose siempre superior los casos en los que estan imputados.



Nuestra exploración en este gráfico nos ha proporcionado una perspectiva valiosa sobre cómo los umbrales de corte influyen en el rendimiento del modelo en diferentes conjuntos de datos. Hemos demostrado que el conjunto de datos con el umbral de corte más extremo, en particular en el caso de obesidad (Dataset Corte 1), logra el máximo rendimiento en términos de predicción, sin importar la proporción de valores faltantes que pueda tener. Este hallazgo respalda la idea del "entrenamiento con valores extremos", donde las instancias más notables y marcadas desempeñan un papel esencial en mejorar la capacidad de predicción del modelo.

Conclusión: Nuestro análisis sobre los umbrales de corte y su impacto en el rendimiento del modelo destaca la importancia de elegir cuidadosamente estos umbrales al clasificar individuos como "obesos" o "no obesos". Encontramos que un enfoque más restrictivo, que identifica casos extremos como obesidad, mejora la precisión del modelo. Esto nos ha llevado a comprender que nuestras decisiones previas respecto a los umbrales de corte podrían no haber maximizado el potencial de predicción.

La consistencia de estos resultados en diferentes conjuntos de datos muestra la efectividad de este enfoque, independientemente de los valores faltantes. Sin embargo, es importante señalar que estos hallazgos pueden no ser aplicables en todos los conjuntos de datos, ya que cada problema puede tener características únicas a considerar.

En resumen, esta exploración mejora nuestras estrategias de modelado y nos guía en la clasificación de casos similares en el futuro.