TECNOLÓGICO DE MONTERREY

Campus Monterrey

Inteligencia artificial avanzada para la ciencia de datos I

TC3006C.101

Módulo 2 : Análisis y Reporte sobre el desempeño del modelo

Suemy Aquino Zumaya

A00828585

Septiembre 2022

1. Modelo implementado

Se utilizó el data set con características físicas de la cara de personas para clasificarlas en hombres y mujeres según su fisiología.Este consta de 7 variables recopilada más la columna de género.

Se implementó el modelo de árboles de decisión ya que se refiere a un ejercicio de clasificación.

2. Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación (Train/Test/Validation).

Los datos se dividieron primero en la variable dependiente que corresponde la columna "gender"la cuál se convirtió en variable categórica [0,1] y las demás se asignaron como variables independientes.

Se dividieron los datos utilizando la librería *sklearn.model_selection* en secciones de 70 % para entrenamiento del modelo, 25 % para su prueba y el 5 % para su validación.

Después de entrenar el modelo se obtuvo el siguiente resultado.

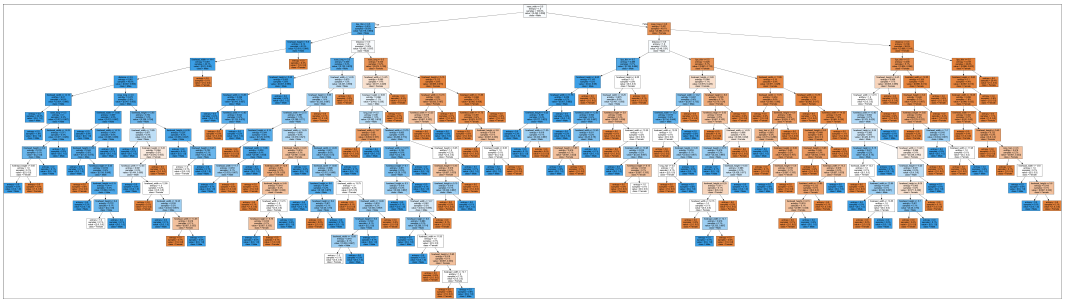

Figura 1: Árbol de decisión inicial

Como se puede observar es un modelo muy extenso y podría ocasionar mayores costos computacionales.

3. Diagnóstico y explicación el grado de bias o sesgo: bajo medio alto

Se obtuvo un MSE con los datos de entrenamiento de 0.0011 mientras que con los datos de prueba fue del 0.0392 de esta manera podemos notar que existe una diferencia notable entre las métricas, podemos definir que el modelo tiene un sesgo *medio-bajo* según el estándar que se quiera tomar, esto nos indica que va a tener menos precisión y exactitud en sus resultados.

4. Diagnóstico y explicación el grado de varianza: bajo medio alto

AL tener una un sesgo no tan alto, se espera también una varianza *media-alta* según los estándares a tomar y esto lo podemos corroborar con la r^2 que es de 0.995 y 0.843 para entrenamiento y prueba respectivamente, esto nos quiere decir que esperamos cambios altos en al cambiar los conjuntos a predecir.

5. Diagnóstico y explicación el nivel de ajuste del modelo: underfitt fitt overfitt

Con las característica de sesgo bajo y varianza alta, podemos notar que existe una tendencia al overfitting del modelo siendo esto que el modelo se "memoriza"los datos de entrenamiento por lo que su predicción falla al probarlo con datos distintos.

Por ello se intentará refinar y simplificar el modelo a manera de reducir el sobre ajuste a los datos y mejorar los valores de sesgo y varianza.

6. Técnicas de regularización para mejorar el desempeño del modelo

6.1. Reducir nodos del árbol

Ya que teníamos overfitting se redujeron los nodos del árbol a solo 3 como máximo para simplificar el modelo con lo que se obtuvo el siguiente diagrama:

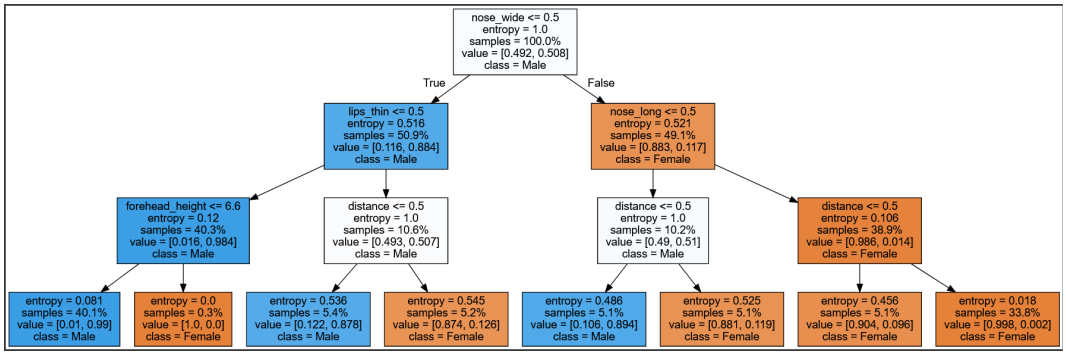

Figura 2: Árbol de decisión de tres nodos

Probando los datos se obtuvieron la métricas que se muestran a continuación con sus respectivos conjuntos de datos.

	Entrenamiento	Pruebas
MSE	0.034	0.042
r^2	0.864	0.833
Puntaje	0.966	0.958

Cuadro 1: Resultados Árbol 3 nodos

Podemos observar que aunque el rendimiento del modelo bajó un poco, este presenta una varianza mucho menor entre grupos, un sesgo menor en los errores y por lo tanto un ajuste de modelo fit. Si bien se logró mejorar las características del modelo, se mejorar su desempeño sin comprometer el sesgo o la varianza.

6.2. Podar el árbol

Este método consiste en simplificar el árbol de decisión dado que quitando las ramas menos significativas, de esta manera con el comando *cost_complexity_pruning_path* podemos obtener diferentes coeficientes para reducir los nodos del árbol para después crear un loop donde se pruebe el desempeño de cada alpha en con los conjuntos de entrenamiento y pruebas como se ve a continuación.

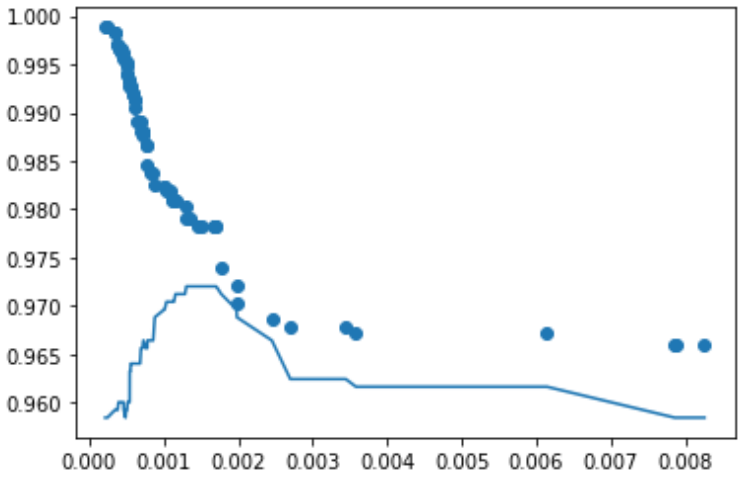

Figura 3: Desempeño de diferentes alphas, puntos = conjunto entrenamiento; línea = conjunto pruebas

De esta manera seleccionamos el alpha con mayor desempeño tanto en entrenamiento como en pruebas para buscar el mejor ajuste del modelo con lo que se obtuvo el siguiente diagrama.

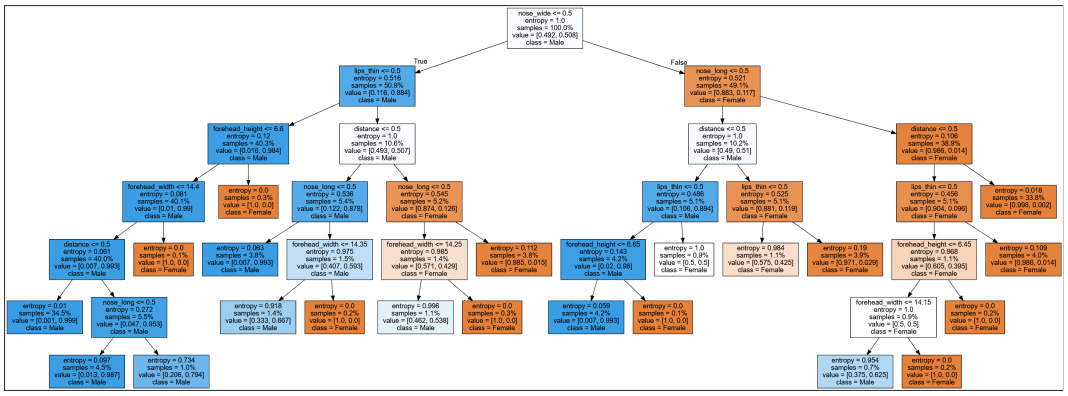

Figura 4: Árbol de decisión podado

El cual tuvo el siguiente desempeño.

	Entrenamiento	Pruebas
MSE	0.028	0.0.030
r^2	0.888	0.878
Puntaje	0.972	0.969

Cuadro 2: Resultados Árbol podado

Con este modelo de árbol de decisión podado, podemos ver que tenemos un modelo por mucho más simple que el inicial pero con un buen nivel de predicción de datos tanto en pruebas como en entrenamiento, inclusive un poco mejor que con el árbol de tres nodos, pero la varianza se redujo en una manera bastante considerable casi al punto de obtener el mismo desempeño con ambos conjuntos de datos, el sesgo aumentó ya que se tienen mejores predicciones de los datos y por lo tanto se obtuvo un modelo que no sufre ni de underfitting ni overfitting, siendo este el fit para los datos.

7. Conclusión

Comparemos el desempeño en el conjunto de pruebas obtenido por cada modelo

	Inicial	3 nodos	Podado
MSE	0.0392	0.042	0.0.030
r^2	0.843	0.833	0.878
Puntaje	0.956	0.958	0.969

Cuadro 3: Comparación de desempeño de los árboles decisión.

Todos los modelos se probaron con 5 valores puntuales y todos predijeron correctamente esos casos de validación solo que cambiaba el nivel de seguridad sobre la decisión a tomar, sin embargo el modelo con mejor ajustes de datos y mejor desempeño de manera simplificada es el árbol obtenido después de la poda de nodos, ya que este demostró tener baja varianza, alto sesgo y tener un ajuste fit con los datos que se estudiaron.

Anexo

Link de GitHub: https://github.com/Suemy-AZ/Modulo2_CL.git