Inteligencia Artificial Avanzada para la Ciencia de Datos I (TC3006C)

Actividad Momento de Retroalimentación: Análisis y Reporte sobre el desempeño del modelo.

Presenta:

Abraham Gil Félix

Matrícula A01750884

Profesor Módulo 2:

Dr. Jorge Adolfo Ramirez Uresti

Estado de México septiembre, 2022

Modelo: Clasificador Bosque Aleatorio

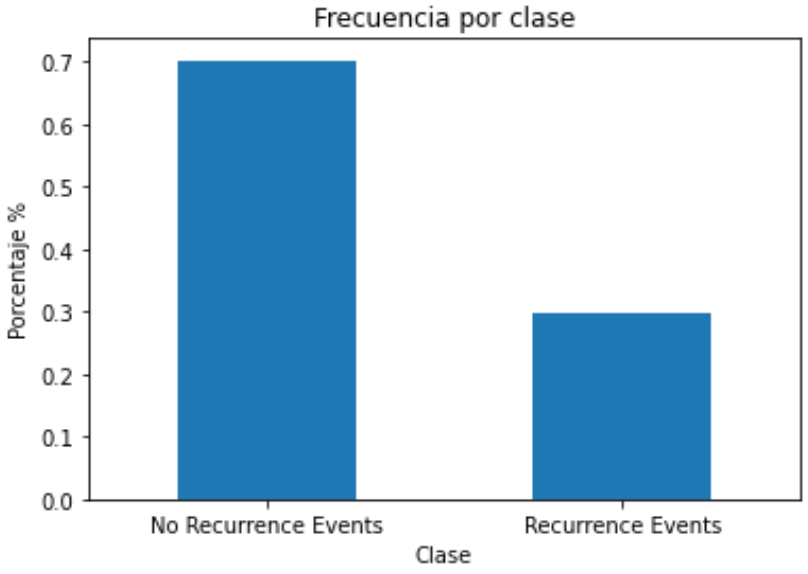
El clasificador de bosque aleatorio es un algoritmo metaestimador que es capaz de ajustar una serie de clasificadores (árboles de decisión) en varias submuestras del conjunto de datos, además de utilizar su promedio para mejorar la precisión predictiva y controlar el sobreajuste. El clasificador puede usar todo el conjunto o submuestras de los datos para construir cada árbol de decisión.

Conjunto de datos

El conjunto de datos con el que se entrenó y probó el modelo se obtuvo del repositorio de aprendizaje automático de la Universidad de California en Irvine, la cual actualmente mantiene 622 conjuntos de datos como un servicio para la comunidad de aprendizaje automático.

Breast-Cancer: conjunto de datos sobre el cáncer de mama. Se emplea como problema de clasificación binaria para predecir aquellos pacientes con eventos recurrentes relacionados a dicha enfermedad.

- Clase 0: personas con eventos no recurrentes relacionados al cáncer de mama.
- Clase 1: personas con eventos recurrentes relacionados al cáncer de mama.



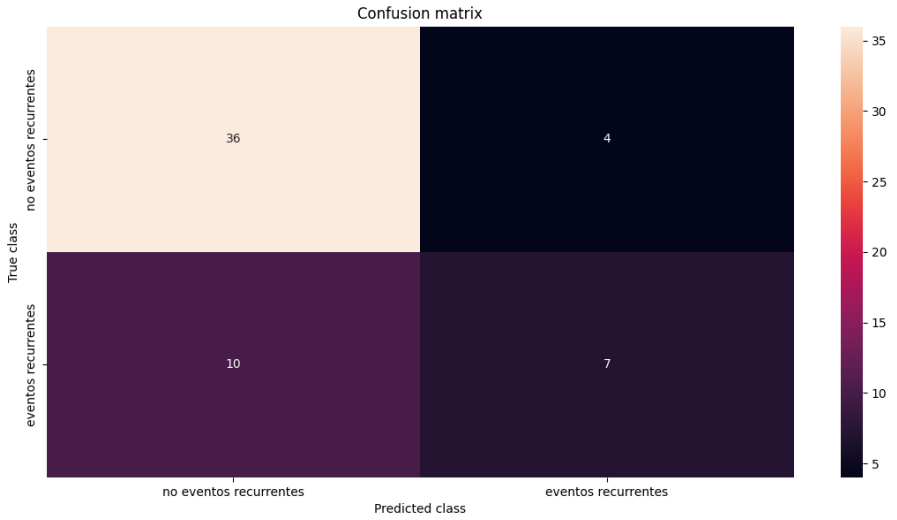
➔ Conjunto de datos multivariante
➔ Número de instancias: 286
➔ Características de los atributos: categóricas
➔ Número de atributos: 9

Para el entrenamiento y prueba del modelo clasificador se utilizaron el 80% de los datos en la etapa de validación, el 20% restante se designó para conocer la efectividad del clasificador. Se utilizó la función *train_test_split* del módulo model_selection de la librería Scikit-Learn para la partición del conjunto de datos.

Grado de bias o sesgo

Dado que, el modelo busca clasificar correctamente aquellos pacientes con eventos recurrentes relacionados con el cáncer de mama, es sumamente importante visualizar la exhaustividad (recall) del modelo en cada clase binaria. Es aquí donde se expone el sesgo del clasificador, ya que a pesar de obtener un *accuracy_score* mayor al 75% en la etapa de prueba, el modelo solo es capaz de clasificar correctamente 7 personas con eventos recurrentes de cáncer de mama de un total de 17 pacientes. No ocurre lo mismo cuando el clasificador predice para aquellas personas con eventos no recurrentes, puesto que solo falla el 10% de dichas predicciones.

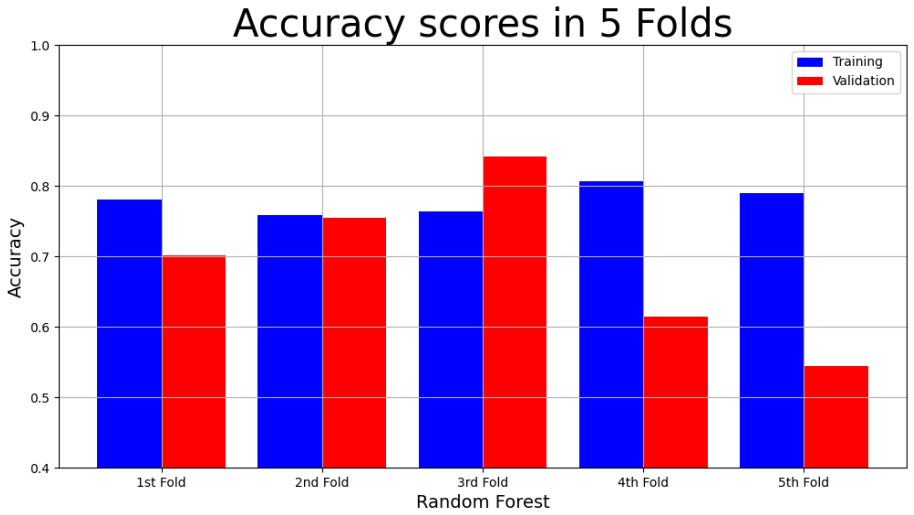
El modelo sufre de un problema conocido como clases no balanceadas, muy recurrente en problemas de clasificación binaria donde el principal objetivo es predecir aquellos eventos con menor frecuencia. A continuación, se muestra una matriz de correlación que permite observar el sesgo en la clase minoritaria.



En la actualidad, existen diversas técnicas para tratar con el desbalance de clases en los problemas de clasificación binaria, algunas incluyen la generación de datos sintéticos de datos o la pérdida estratégica de información en la clase mayoritaria. Sin embargo, se optó por la implementación del parámetro *class_weight* con el valor 'balanced' del algoritmo *RandomForestClassifier.* No obstante, el incremento del porcentaje de exhaustividad (recall) para la clase minoritaria después de dicha implementación fue únicamente del 12%, obteniendo así 53%. Por lo que, si se plantea el siguiente cuestionamiento; ¿qué porcentaje de los pacientes con eventos recurrentes de cáncer de mama es capaz de identificar el modelo? La respuesta más sensata es que las predicciones del clasificador serían similares a tirar una moneda al aire.

Grado de varianza

En las etapas de entrenamiento y prueba del modelo clasificador, se implementó 5-fold Cross-Validation con el propósito de conocer su estabilidad y el score promedio variando el conjunto de datos. A continuación, se presenta una gráfica de barras que permite visualizar lo anterior.



Respecto a la estabilidad del modelo, se muestra la siguiente tabla donde se observa que el modelo obtuvo 0.04 para la evaluación de la desviación estándar. Lo cual nos asegura la un modelo realmente estable, ya que el rango en el que puede variar el accuracy es realmente pequeño.

Etapa	**Desviación Estándar**
Entrenamiento	0.0352
Prueba	0.0363

Nivel de ajuste del modelo

En principio, el modelo base padeció de overfitting en entrenamiento, ya que no se tuneó ningún hiper parámetro; entre ellos *max_depth*. Dicho parámetro controla la profundidad máxima del árbol, en caso de que se establezca el valor por defecto, los nodos se expanden hasta que las hojas sean puras o incluso, hasta que todas las hojas contengan el número mínimo de muestras requeridas para dividir un nodo interno.

Modelo	**Score en entrenamiento**	**Score en prueba**
RandomForestClassifier	0.9868	0.7192
RandomForestClassifier Tuning Hyperparameters	0.7763	0.7544

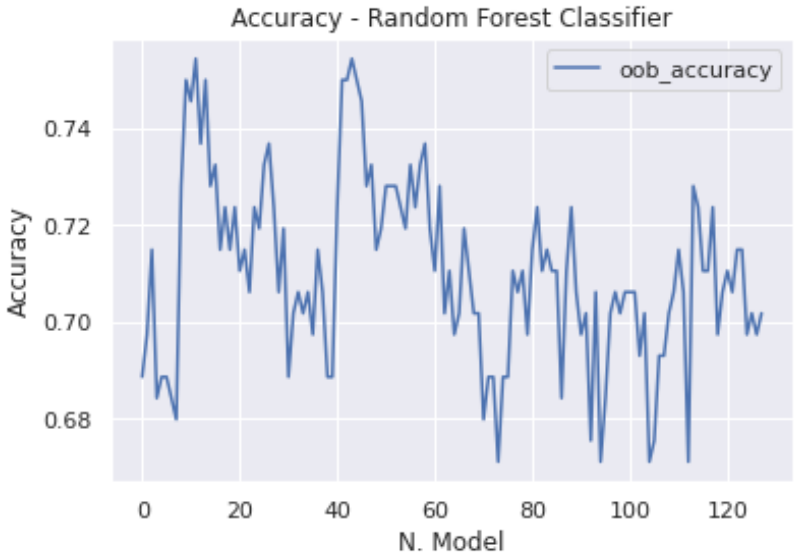
Ajuste de parámetros para mejorar el desempeño del modelo

Con el principal objetivo de obtener el mejor modelo clasificador de bosque aleatorio se implementó *GridSearch*. Es un método de ajuste de parámetros; búsqueda exhaustiva: entre todas las selecciones de parámetros candidatos, a través de bucles y probando todas las posibilidades. Dado que, una desventaja de este método es el tiempo de cómputo cuando la existen bastantes parámetros a evaluar, se optó por ajustar únicamente los siguientes:

Parámetro	Descripción
n_estimators	Número de árboles en el bosque.
criterion	Función para medir la calidad de división (grado de pureza).
max_depth	La profundidad máxima del árbol. Si es Ninguno, los nodos se expanden hasta que todas las hojas sean puras o hasta que todas las hojas contengan menos de min_samples_split samples.
max_features	El número de características a considerar al buscar la mejor división
class_weigth	Pesos asociados a clases en el formulario . Si no se proporciona, se supone que todas las clases tienen peso uno.

- ➔ Número de estimadores: 100 - 1000
- ➔ Máximos atributos: 3, 5, 7 o 9
- ➔ Máxima profundidad: ninguna, 3, 6 o 9
- ➔ Criterio: Gini o Entropía
- ➔ Peso de la clase: ninguna o balanceada

En total, se corrieron 128 modelos con la intención de obtener el accuracy más alto. A continuación, se muestra un gráfica que analiza el accuracy variando algunos hiper parámetros:



Mejor modelo:

N. Modelo	Accuracy	Criterio	N. Estimadores	Máximos atributos	Máxima profundidad	Peso de la clase
11	0.754386	gini	1000	5	3.0	None
43	0.754386	entropy	1000	5	3.0	None

Link Colab:
https://colab.research.google.com/drive/18m9WhDZ8-CXXaacZUROJl0AjZh6l31Jq?usp=sharing