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

Actividad Momento de Retroalimentación: Uso de framework o biblioteca de aprendizaje máquina para la implementación de una solución

Presenta:

Abraham Gil Félix

Matrícula A01750884

Profesor Módulo 2:

Dr. Jorge Adolfo Ramirez Uresti

Estado de México septiembre, 2022

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.

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

Modelo: Clasificador Bosque Aleatorio

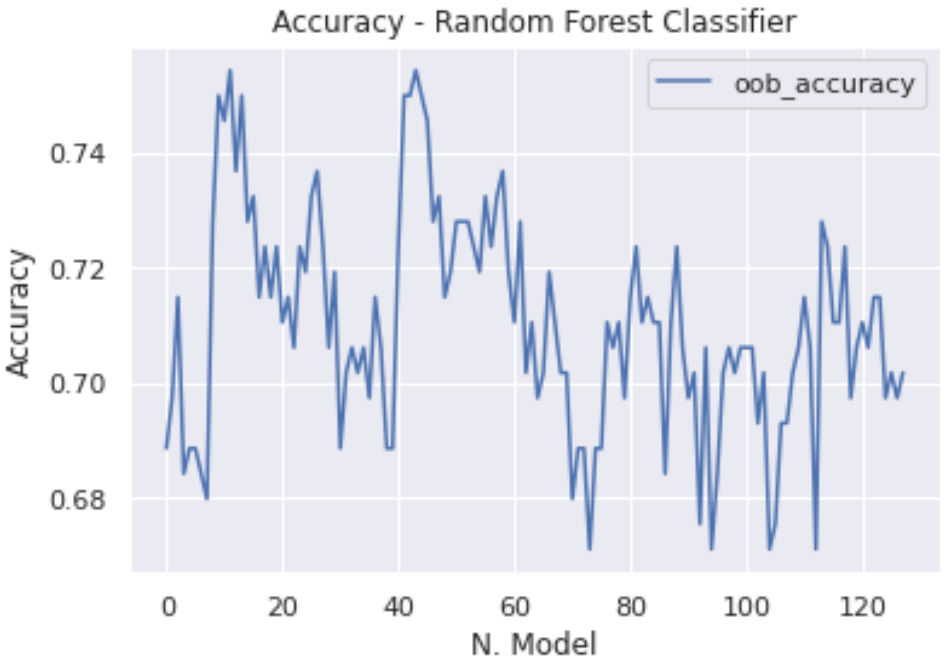
Para el entrenamiento y prueba del modelo clasificador se utilizaron el 80% de los datos en la etapa de validación, el 20% restante se designó para conocer la efectividad del clasificador.

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 los siguientes:

Parámetro	Descripción
n_estimators	Número de árboles en el bosque.
criterion	Función para medir la calidad de divisón (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