Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Imagen que contiene Texto

Descripción generada automáticamente

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

TC3006C.101

Profesores:

Jorge Adolfo Ramírez Uresti

**Momento de Retroalimentación: Módulo 2 Uso de framework o biblioteca de aprendizaje máquina para la implementación de una solución.**

Abner Maximiliano Lecona Nieves | A01753179

09 de Septiembre 2024

**Documentación de Random Forest Classifier**

Este script implementa un modelo de clasificación usando Random Forest para predecir una variable objetivo en un conjunto de datos de jugadores de béisbol, predeciremos si un beisnlista entra o no entra al salón de la fama.

El proceso incluye la preparación de datos, escalado, entrenamiento y evaluación del modelo.

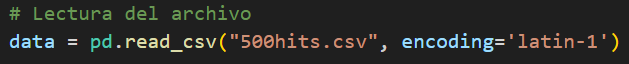
Importamos las librerías

* pandas: Para manipulación y análisis de datos.
* sklearn.model\_selection: Para dividir los datos en conjuntos de entrenamiento y prueba.
* sklearn.preprocessing: Para escalar los datos usando Min-Max Scaling.
* sklearn.ensemble: Para implementar el modelo de Random Forest.
* sklearn.metrics: Para evaluar el rendimiento del modelo mediante reportes de clasificación. las librerías.

Texto

Descripción generada automáticamente

Cargamos el conjunto de datos del archivo csv llamado “500hits”.



Eliminaremos las columnas ‘PLAYER’ y ‘CS’. No serán útiles para que nuestro modelo funcione de manera eficiente.



Separamos las columnas que nos ayudaran a predecir, si es que entrara o no al salón de la fama. Siendo “X” las características del jugador y “y” si es que este esta o no esa en el salón de la fama.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Dividimos el conjunto de datos. Se dividen en conjuntos de entrenamiento y prueba en una proporción estándar (75% entrenamiento, 25% prueba) con train\_test\_split.

Texto

Descripción generada automáticamente

Modelo con los datos escalados

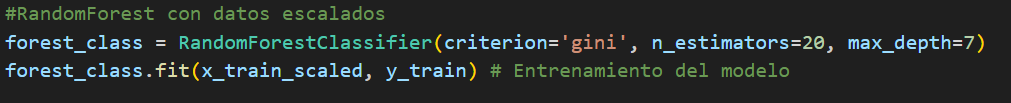
Escalaremos los datos con Min-Max Scaling, esto para los datos de prueba y los datos de testeo. (este paso es opcional, ya que el modelo de RandomForest no es sensible a datos escalados, lo estamos haciendo por fines demostrativos).

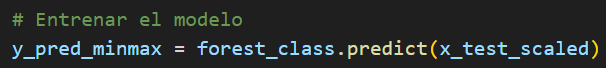
Texto

Descripción generada automáticamente

Se entrena el modelo con los datos escalados.

* Usamos 'gini' por su eficiencia y porque es adecuado para conjuntos de datos donde queremos maximizar la precisión de la clasificación de forma rápida.
* Elegimos 20 árboles para equilibrar la precisión y la eficiencia computacional. Más árboles pueden mejorar la precisión, pero también aumentan el tiempo de cómputo y la complejidad del modelo.
* Seleccionamos una profundidad de 7 para evitar el sobreajuste, manteniendo los árboles lo suficientemente profundos para capturar patrones significativos en los datos sin hacerlos demasiado complejos y ajustados a ruido.





Evaluamos el rendimiento del modelo con librerías de sklearn.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Estas son las métricas del modelo con los datos escalados.

Calendario

Descripción generada automáticamente

El modelo tiene una alta precisión en el conjunto de entrenamiento, lo cual indica que es capaz de capturar bien los patrones en los datos de entrenamiento.

La precisión en el conjunto de prueba es menor que en el conjunto de entrenamiento, pero sigue siendo razonablemente buena.

En conclusión, el modelo funciona bien en general, pero es más preciso para la clase 0. Esto podría ser un problema si la clase 1 es más importante o si los falsos negativos tienen un alto costo para el modelo.

Modelo con los datos no escalados

Ahora utilizaremos los datos sin escalar. Haremos el mismo procedimiento pero con los datos normales.

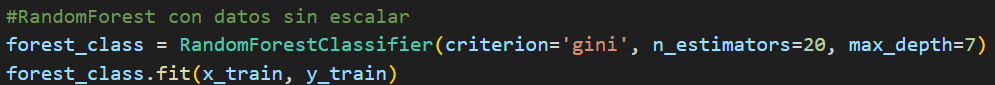


Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza baja

Texto

Descripción generada automáticamente

Calendario

Descripción generada automáticamente

El modelo tiene una alta precisión en el conjunto de entrenamiento (0.9339), lo cual indica que es capaz de capturar bien los patrones en los datos de entrenamiento. La precisión en el conjunto de prueba es también alta (0.8632), lo cual demuestra que el modelo generaliza bien a nuevos datos y mantiene un buen rendimiento.

Conclusión General

Los datos sugieren que el modelo de Random Forest funciona mejor sin escalar los datos. Esto confirma la teoría de que los árboles de decisión y, por ende, los Random Forest, no son sensibles a la escala de los datos debido a la naturaleza de cómo se construyen los umbrales de decisión. Dado que el modelo sin escalar muestra mejor rendimiento en términos de precisión y recall en ambas clases, sería recomendable usar los datos sin escalar para este caso específico.