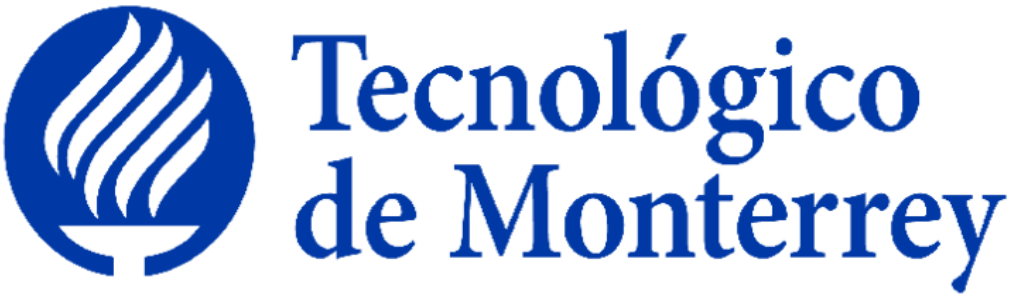

Campus Monterrey

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

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo.

Aldo Daniel Villaseñor Fierro A01637907

Introducción

En este reporte se presenta el desarrollo de un modelo de aprendizaje automático para la clasificación de tumores en dos categorías: malignos y benignos. Esta clasificación se basa en datos médicos obtenidos a partir de diversos estudios diagnósticos.

El proceso implica los siguientes pasos clave:

Adquisición de datos: El paso inicial consiste en importar el conjunto de datos desde un archivo CSV. El conjunto de datos contiene atributos médicos y etiquetas, donde "M" representa los tumores malignos y "B" los benignos. Para facilitar el modelado, se crea una variable binaria "lbl_n" con valores "1" para maligno y "0" para benigno.

Selección de características: Para mejorar la eficiencia y eficacia del modelo, sólo se seleccionan para el análisis las características con una correlación superior a 0,5 con la variable "lbl_n". Este paso ayuda a reducir la dimensionalidad y a centrarse en los atributos más relevantes.

Preprocesamiento de datos: Las características seleccionadas se escalan utilizando el escalado Min-Max para garantizar que todas las variables tienen la misma escala, lo cual es importante para el rendimiento de los algoritmos de aprendizaje automático.

Construcción del modelo: Se elige un clasificador de árbol de decisión como modelo de aprendizaje automático para esta tarea. Se optimizan varios hiperparámetros del árbol de decisión utilizando Optuna, una biblioteca de ajuste de hiperparámetros. Los mejores hiperparámetros se seleccionan en función de los resultados de cross-validation.

Evaluación del modelo: El rendimiento del modelo se evalúa en los conjuntos de datos de entrenamiento y validación. Se utilizan métricas como la precisión, la puntuación F1 y la puntuación ROC AUC para evaluar el rendimiento de clasificación del modelo. También se examinan las matrices de confusión para comprender mejor el comportamiento del modelo.

Importancia de las características: El informe incluye un gráfico de barras que ilustra la importancia de las características asignadas por el modelo de árbol de decisión. Esto ayuda a identificar qué atributos tienen la influencia más significativa en la clasificación de tumores.

Predicciones del modelo: Por último, el informe presenta un DataFrame que contiene las predicciones del modelo y los valores reales para un subconjunto de los conjuntos de datos de prueba y entrenamiento. Esto permite una inspección visual de lo bien que se alinean las predicciones del modelo con las clasificaciones reales de tumores.

Visualización del árbol de decisión: El informe concluye con una representación visual del modelo de árbol de decisión optimizado, que muestra su estructura y la importancia de las características.

El objetivo de este informe es proporcionar información sobre el desarrollo y la evaluación de un modelo de aprendizaje automático para la clasificación de tumores, haciendo hincapié en la transparencia y la interpretabilidad de los resultados.

Selección de Hiperparametros (Optuna)

Se utilizó la librería de python Optuna para la obtención de los parámetros óptimos de una función de sci-kit learn para entrenar un Árbol de decisión para clasificación. Optuna utiliza cross-validation con 3 folds al momento de entrenar y evaluar el modelo, se utilizó la métrica accuracy como métrica a optimizar por optuna. También se realizaron otras pruebas con un set de validación adicional el cual no fue incluido en el proceso de optuna.

En Optuna se declaran todas las variaciones que se deseen hacer para cada modelo y la librería se encarga realizar varias pruebas ("trials") para cada configuración de los parámetros y presenta como mejor candidato al modelo con el mejor rendimiento relativo a una métrica, accuracy en este caso.

Además de las pruebas hechas con optuna también se guarda un subset de datos para realizar un validación el cual no es utilizado en el proceso hecho por optuna para obtener las métricas Accuracy, F1 score y ROC AUC. La accuracy indica la proporción de predicciones que el modelo hizo correctamente, es la más simple de estas 3. El ROC, también conocido como AUC, es el área bajo la curva, valores más cercanos a 1 indican un mejor rendimiento del modelo, mientras que un valor de 0.5 (o cercano a él) puede indicar que el modelo no es mejor que una elección aleatoria. El F1 score es una métrica que combina precisión y sensibilidad (recall) en una sola cifra. Es especialmente útil cuando se tiene un conjunto de datos con una distribución de clases desequilibrada. Un F1 score perfecto sería 1, mientras que el peor valor posible es 0.

Evaluación del modelo

Árbol de decisión

Parámetros elegidos basándose:

```
sklearn.DecisionTreeClassifier(criterion='entropy',splitter='best',
max_features='auto',max_depth=8,min_samples_split=10,
min_samples_leaf=1, ccp_alpha=0.0011457371973159118)
```

Sesgo: Medio

La elección de un valor bajo para el parámetro C, que representa la regularización inversa, implica que el modelo puede hacer ciertas suposiciones sobre los datos, incluso en áreas donde los datos son escasos. Esto podría introducir un cierto grado de sesgo en las predicciones del modelo. Sin embargo, esta tendencia se equilibra en parte con la disminución de la varianza.

Varianza: Baja
La regularización inversa, al tener un valor bajo, contribuye a reducir la varianza del modelo. Una varianza baja significa que el modelo es más consistente y menos propenso a sobre ajustarse a los datos de entrenamiento. La consistencia de los resultados se evidencia en las múltiples iteraciones de Optuna (alrededor de 1000), que dieron resultados similares en términos de precisión (accuracy). Esto indica que la estadística de precisión es confiable y constante.

Nivel de Ajuste: Adecuado
El nivel de ajuste del modelo se considera adecuado gracias a la limpieza de datos previa realizada en el primer entregable. Los datos limpios y la selección de un conjunto de variables relevantes permiten que los modelos, incluido el DecisionTreeClassifier, funcionen de manera óptima. El rendimiento del modelo muestra que se logra una precisión de poco más del 94%, lo que sugiere que se ha logrado un ajuste adecuado a los datos disponibles.

En resumen, el modelo DecisionTreeClassifier con los parametros ya mencionados es una elección sólida para problemas de clasificación, con un equilibrio entre sesgo y varianza que se logra mediante la optimización cuidadosa de los hiperparámetros. Su rendimiento adecuado en el conjunto de datos limpios respalda su utilidad en el proceso de toma de decisiones y clasificación de datos.
Resultados para el set de validation:

```
Accuracy:  0.945054945054945
roc auc: 0.9401515151515151
F1 score: 0.9295774647887323

array([[53,  2],
       [ 3, 33]], dtype=int64)
```