Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Escuela de Ingeniería y Ciencias

TC3006C.101

Inteligencia artificial avanzada para la ciencia de datos I

Momento de retroalimentación Módulo 2:

Análisis y reporte sobre el desempeño del modelo

Alumno:

Alan Alcántara Ávila

A01753505

Profesor:

Jorge Adolfo Ramírez Uresti

Fecha de entrega:

11/09/2024

Algoritmo de Machine Learning

Árboles de decisión

Resumen del dataset

El dataset contiene información relacionada con estudiantes inscritos en títulos de pregrado. Tiene información del estudiante, así como su desempeño académico en los dos semestres anteriores. El objetivo es predecir si un estudiante abandonó, sigue inscrito o se gradúa.

Justificación del dataset para el AML

El dataset contiene una variedad de características de variables categóricas y numéricas representadas ya de forma numérica, relacionadas con el desempeño académico y socioeconómico que el árbol de decisión puede manejar.

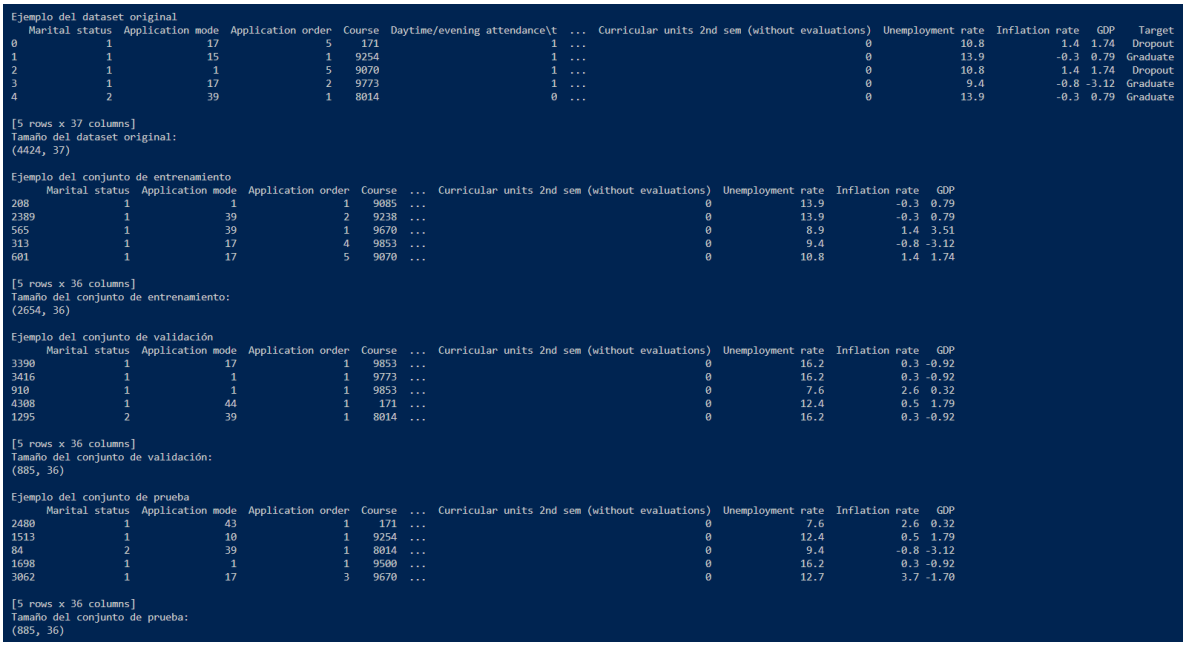
Igualmente, la variable que buscamos predecir tiene múltiples categorías: "Dropout", "Enrolled" y "Graduate". Esto hace que los árboles nos permitan manejar mejor el problema al ser de clasificación multiclase.

Finalmente considero que el dataset contiene suficientes instancias que permiten realizar un buen análisis de desempeño en conjuntos de datos grandes, así como dividir los datos en subconjuntos de entrenamiento, validación, y prueba.

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

La división de los datos en conjuntos de entrenamiento, validación, y prueba permite evaluar la capacidad del modelo para generalizar a datos no vistos, para aprender de un subconjunto de datos y luego ser evaluado con en datos que no ha visto y para poder realizar ajustes al modelo sin darle a conocer toda la información.

En mi caso, decidí realizar la separación dando un 60% para entrenamiento y del 40% restante lo dividí a la mitad, finalizando con 20% para validación y 20% para pruebas del dataset total.

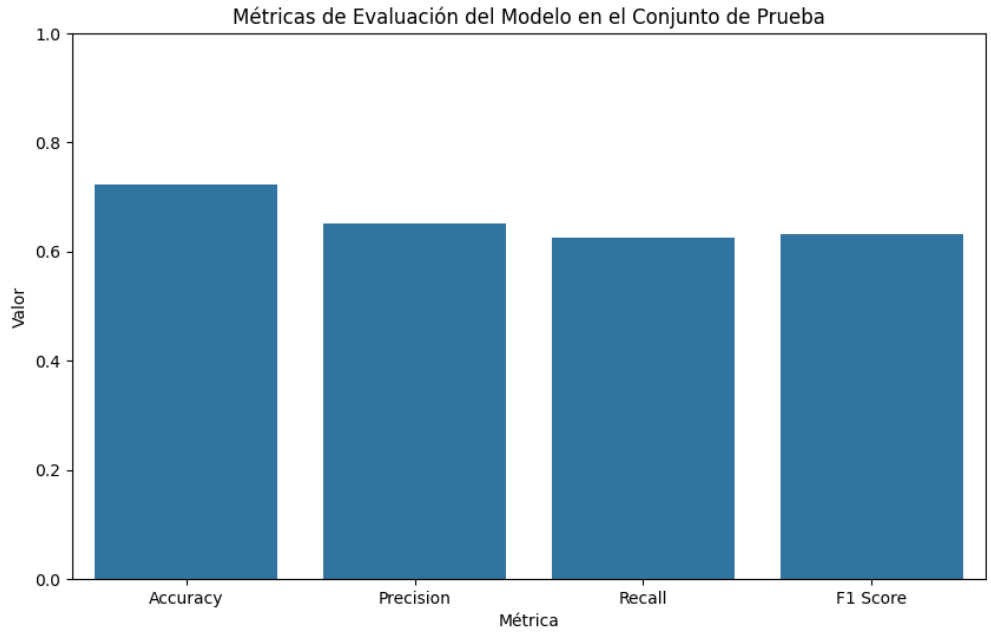


```
Ejemplo del dataset original
   Marital status  Application mode  Application order  Course  Daytime/evening attendance\t  ...  Curricular units 2nd sem (without evaluations)  Unemployment rate  Inflation rate   GDP    Target
0               1                17                  5     171                             1  ...                                              0               10.8             1.4  1.74   Dropout
1               1                15                  1    9254                             1  ...                                              0               13.9            -0.3  0.79  Graduate
2               1                 1                  5    9070                             1  ...                                              0               10.8             1.4  1.74   Dropout
3               1                17                  2    9773                             1  ...                                              0                9.4            -0.8 -3.12  Graduate
4               2                39                  1    8014                             0  ...                                              0               13.9            -0.3  0.79  Graduate

[5 rows x 37 columns]
Tamaño del dataset original:
(4424, 37)

Ejemplo del conjunto de entrenamiento
      Marital status  Application mode  Application order  Course  ...  Curricular units 2nd sem (without evaluations)  Unemployment rate  Inflation rate   GDP
208                1                 1                  1    9085  ...                                              0               13.9            -0.3  0.79
2389               1                39                  2    9238  ...                                              0               13.9            -0.3  0.79
565                1                39                  1    9670  ...                                              0                8.9             1.4  3.51
313                1                17                  4    9853  ...                                              0                9.4            -0.8 -3.12
601                1                17                  5    9070  ...                                              0               10.8             1.4  1.74

[5 rows x 36 columns]
Tamaño del conjunto de entrenamiento:
(2654, 36)

Ejemplo del conjunto de validación
      Marital status  Application mode  Application order  Course  ...  Curricular units 2nd sem (without evaluations)  Unemployment rate  Inflation rate   GDP
3390               1                17                  1    9853  ...                                              0               16.2             0.3 -0.92
3416               1                 1                  1    9773  ...                                              0               16.2             0.3 -0.92
910                1                 1                  1    9853  ...                                              0                7.6             2.6  0.32
4308               1                44                  1     171  ...                                              0               12.4             0.5  1.79
1295               2                39                  1    8014  ...                                              0               16.2             0.3 -0.92

[5 rows x 36 columns]
Tamaño del conjunto de validación:
(885, 36)

Ejemplo del conjunto de prueba
      Marital status  Application mode  Application order  Course  ...  Curricular units 2nd sem (without evaluations)  Unemployment rate  Inflation rate   GDP
2480               1                43                  1     171  ...                                              0                7.6             2.6  0.32
1513               1                10                  1    9254  ...                                              0               12.4             0.5  1.79
84                 2                39                  1    8014  ...                                              0                9.4            -0.8 -3.12
1698               1                 1                  1    9500  ...                                              0               16.2             0.3 -0.92
3062               1                17                  3    9670  ...                                              0               12.7             3.7 -1.70

[5 rows x 36 columns]
Tamaño del conjunto de prueba:
(885, 36)
```

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

En este caso, la gráfica nos permite ver que el bias de nuestro modelo es medio. Al no estar las métricas del accuracy, precision, recall y F1 score tan cercanas al uno, esto nos indica que puede que el modelo este perdiendo ciertos patrones en los datos, afectando un poco su rendimiento.

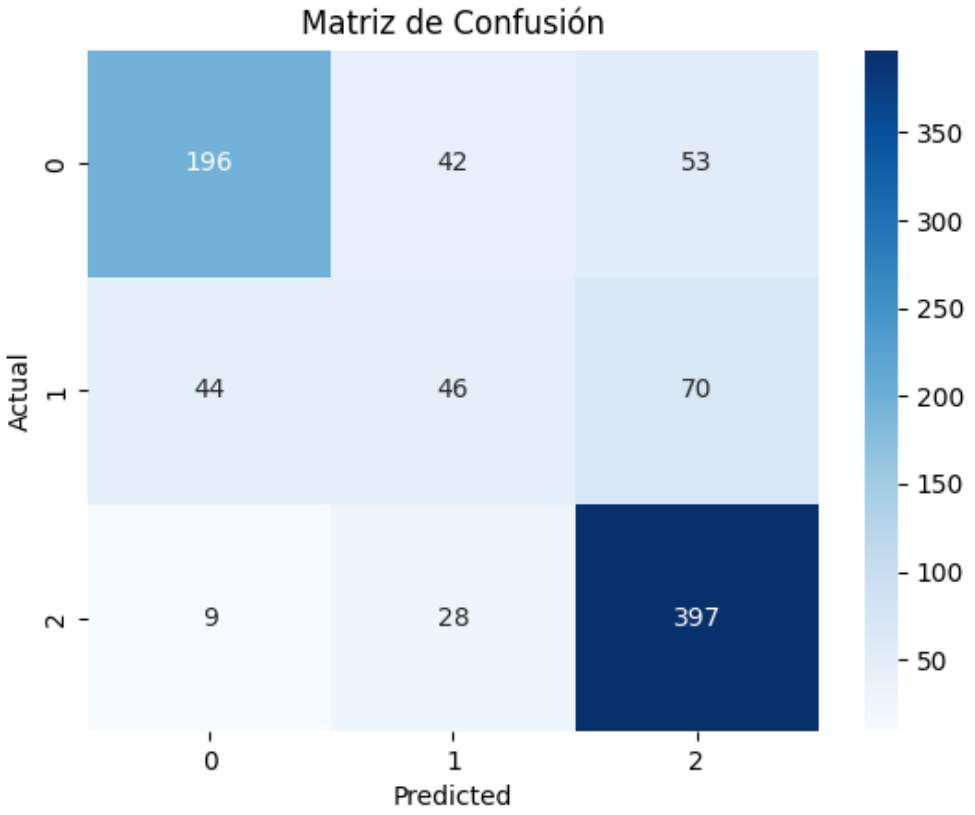
```
Métricas en el conjunto de prueba
Accuracy: 0.7390
Precision: 0.6926
Recall: 0.6593
F1 Score: 0.6690
```



Diagnóstico y explicación el grado de varianza: bajo, medio, alto

Con la matriz de confusión podemos comparar el rendimiento que tiene el modelo para predecir entre las tres clases que nosotros tenemos. Podemos ver que un gran numero de las predicciones son correctas, sin embargo, aún hay muchos datos que los identifica de una clase diferente.

Considerando también las métricas de evaluación, se ve que los valores no son tan dispersos entre ellos, lo que sugiere que la varianza de nuestro modelo es baja o media. Esto indica que el modelo tiene un rendimiento consistente en diferentes conjuntos de datos.



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

Basado tanto en las métricas de evaluación del modelo, así como el grado de sesgo y varianza que se ha calculado previamente, podría decir que el modelo tiene un nivel de ajuste cercano al fitt.

Al tener un sesgo medio y una varianza baja, se da a entender que el modelo tiene un rendimiento decente en diferentes conjuntos, por lo que si está generalizando. Sin embargo, aún hay un margen relativamente grande para mejorar la clasificación que nos brinda el problema de este dataset.

Técnicas de regularización o ajuste de parámetros

Hiperparámetros para "podar el árbol"

Usar hiperparámetros como max_depth, y min_samples_leaf nos permite reducir el tamaño del árbol eliminando ramas que potencialmente tengan poca importancia. Podemos de esta forma limitar la profundidad máxima del árbol, por ejemplo, con el fin de que no se capture tanto ruido.

```
# Hiperparámetros para optimización
param_grid = {
    'criterion': ['gini', 'entropy'],
    'max_depth': [4, 5, 6, 7],
    'min_samples_split': [4, 5, 6, 7],
    'min_samples_leaf': [2, 3, 4],
    'max_features': [None, 15, 25, 35]
}
```

Validación Cruzada

En lugar de usar únicamente una división de entrenamiento y prueba, con validación cruzada hacemos múltiples particiones para evaluar el modelo, generando una estimación más precisa del rendimiento del modelo. Esto significa que, durante el entrenamiento, el modelo se entrena y evalúa varias veces con diferentes particiones del conjunto de entrenamiento, asegurando que el rendimiento no depende de una única división de datos.

```
cv=5,
```

Ajuste de Hiperparámetros

Utilizando una técnica de búsqueda (GridSearch) para encontrar la combinación óptima de hiperparámetros. Así podemos encontrar la mejor configuración para mejorar el rendimiento del modelo en términos de las métricas de evaluación.

```
grid_search = GridSearchCV(tree, param_grid, cv=6, scoring='accuracy')
grid_search.fit(X_train, y_train)
```

ANTES

```
Reporte de clasificación (Validation Set)
              precision    recall  f1-score   support

     Dropout       0.85      0.69      0.76       297
    Enrolled       0.50      0.30      0.37       166
    Graduate       0.72      0.93      0.81       422

    accuracy                           0.73       885
   macro avg       0.69      0.64      0.65       885
weighted avg       0.72      0.73      0.71       885


Métricas en el conjunto de prueba (Test Set)
Accuracy: 0.7492
Precision: 0.7117
Recall: 0.6514
F1 Score: 0.6640
```

DESPUÉS

```
Reporte de clasificación (Validation Set):
              precision    recall  f1-score   support

     Dropout       0.66      0.67      0.67       297
    Enrolled       0.37      0.33      0.35       166
    Graduate       0.76      0.78      0.77       422

    accuracy                           0.66       885
   macro avg       0.60      0.59      0.60       885
weighted avg       0.65      0.66      0.66       885


Métricas en el conjunto de prueba (Test Set):
Accuracy: 0.6893
Precision: 0.6293
Recall: 0.6265
F1 Score: 0.6276
```