Instituto Tecnológico de Estudios Superiores de Monterrey

Reporte del modelo: Árbol de decisión

Inteligencia artificial avanzada para ciencia de datos

Grupo 102

Alumno

Luis Gabriel Martínez Rentería

A01651812

Introducción

A lo largo de este bloque se trabajaron dos modelos de machine learning, una implementación en la que no usáramos ningún Framework y otra en la que si se usara un Framework. Tras haber analizado ambas implementaciones, opté por desarrollar más el árbol de decisión, ya que este es el que mejores resultados me daba a la hora de entrenar el modelo y mejor interfaz de implementación tenía.

Entre los cambios que tuve que hacerle a la entrega está el agregar un set de validación, ya que este no se encontraba en el original, como también las opciones en el menú para que se usara este. También agregue la manera de ver el sesgo y la varianza del modelo, como otras cosas que ya tenía para el set de prueba que agregue para el set de validación.

Entre los diferentes parámetros que se le pueden cambiar al árbol de decisión, no tuve que cambiar o ajustar nada, ya que al correr con menos datos de entrenamiento aún seguía dándome buenos resultados, ya que en el set de prueba tengo una precisión de 97% mientras que en el set de validación 87%, aunque la diferencia es de 10%, el sesgo es muy bajo.

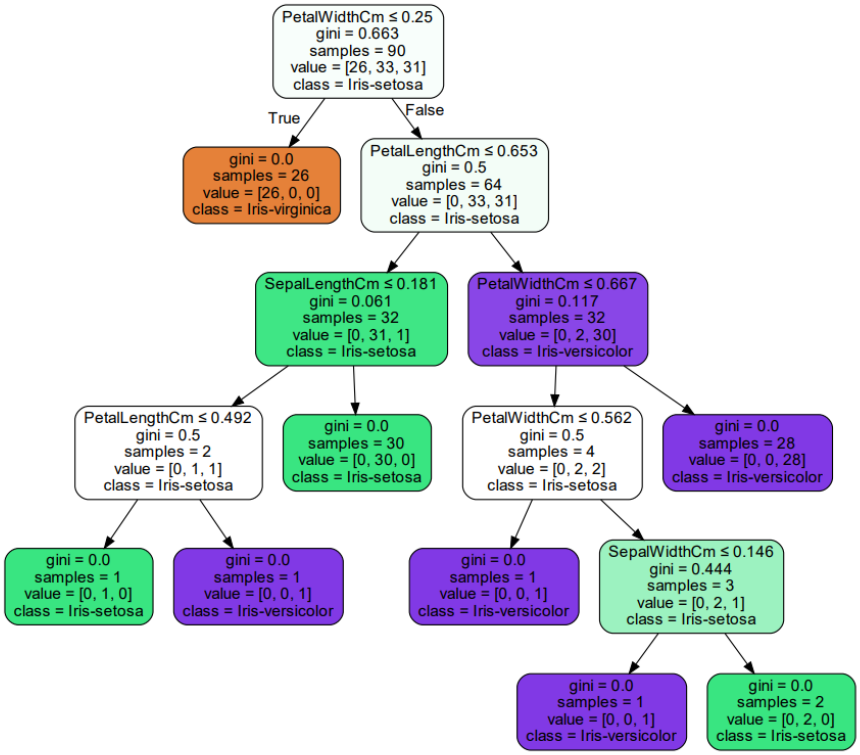
Ejemplos del uso del programa

```
Options:
    1. Predict data
    2. Plot Decision Tree
    3. Print Metrics
    4. Use Validation Set
    5. Print Metrics with validation set
    6. Print Bias and Variance
    7. Exit

Select and option:
```

```
Select and option: 1
Predictions: [1 2 2 0 0 1 0 1 1 1 2 1 0 0 2 1 0 0 0 0 0 2 1 0 1 1 0 0 1 1]
Predicted data with Success!
```

```
Select and option: 3
Metrics:
        Accuracy: 0.9666666666666667
        R^2 :  0.9424184261036468
        MAE : 0.03333333333333333
        RMSE: 0.18257418583505536

Confusion Matrix
                 Iris-virginica  Iris-setosa  Iris-versicolor
Iris-virginica               13            0                0
Iris-setosa                   0           11                0
Iris-versicolor               0            1                5

Classification report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        13
           1       0.92      1.00      0.96        11
           2       1.00      0.83      0.91         6

    accuracy                           0.97        30
   macro avg       0.97      0.94      0.96        30
weighted avg       0.97      0.97      0.97        30
```

Árbol de decisión generado por el modelo:




```
Select and option: 4
Predictions: [2 2 0 1 2 0 2 2 1 1 2 0 2 0 2 0 1 1 0 0 1 0 1 0 1 2 2 0 2 0]
Predicted validation data with Success!
```

}

```
Metrics:
       Accuracy: 0.8666666666666667
       R^2 :  0.8324022346368716
       MAE : 0.13333333333333333
       RMSE: 0.3651483716701107

Confusion Matrix
                Iris-virginica  Iris-setosa  Iris-versicolor
Iris-virginica              11            0                0
Iris-setosa                  0            5                1
Iris-versicolor              0            3               10

Classification report
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        11
           1       0.62      0.83      0.71         6
           2       0.91      0.77      0.83        13

    accuracy                           0.87        30
   macro avg       0.84      0.87      0.85        30
weighted avg       0.89      0.87      0.87        30
```

VARIANZA Y SESGO DEL MODELO

```
Select and option: 6
MSE: 0.023
BIAS: 0.015414999999999996
VAR: 0.007585000000000001
```

Como se puede ver, el sesgo (BIAS) y la Varianza del modelo son muy bajos, lo que nos puede decir dos cosas:

1) El modelo se esta sobre entrenando y se tiene que ajustar los parámetros para que sea más transferible a nuevos datos
2) La información que le estamos dando al modelo es muy buena y por eso se ajusta muy bien a los datos y aun así nos da buenos resultados en los sets de prueba y de validación

Personalmente me voy por la segunda opción, ya que en el caso de que se tratara de un overfitting que no da buenos resultados en las predicciones si se debiese de hacer un cambio, pero el que los registros se parezcan y los datos estén limpios y sean coherentes entre ellos, el modelo va a dar buenos resultados sin la necesidad de que se esté aprendiendo los datos.

En el caso de la varianza, también tenemos una varianza baja, lo cual puede decir que los datos usados para el fit del modelo no cambian mucho, y en el caso de que entre un dato muy fuera de lo que se entrenó, el modelo empezará a dar malos resultados.

En conclusión, se podría decir que el modelo está teniendo un overfitting, pero como los datos de la base de datos se parecen mucho entre ellos no es necesario hacer un ajuste en el modelo porque, para los datos existentes, se dan buenos resultados. Si en un futuro se ampliara la base de datos usada se podría volver a entrenar al modelo con el fin de que este se ajuste a nuevos datos.