Semana_14_Clasificacion

May 21, 2025

0.1 Arboles de decisión

Según [1] un arbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. De acuerdo con [2] se aplican para tareas de clasificación son un subtipo de árbol de predicción para una variable de respuesta categórica.

Tiene una estructura jerárquica de árbol, que consta de un nodo raíz, ramas, nodos internos y nodos hoja.Como puede ver en el siguiente diagrama, un árbol de decisión comienza con un nodo raíz, que no tiene ninguna rama entrante. Las ramas salientes del nodo raíz luego alimentan los nodos internos, también conocidos como nodos de decisión. Con base a [2] este árbol se construye por medio del método de ***división binaria recursiva***.

Tomado de [1]

Para [1] este tipo de estructura en forma de árbol también crea una representación fácil de digerir de la toma de decisiones, lo que permite a los diferentes grupos de una organización comprender mejor por qué se tomó una decisión.

De acuerdo con [1] el aprendizaje de árboles de decisión emplea una estrategia de divide y vencerás realizando una búsqueda codiciosa para identificar los puntos de división óptimos dentro de un árbol. Este proceso de división se repite de forma descendente y recursiva hasta que todos o la mayoría de los registros se hayan clasificado con etiquetas de clase específicas. Para [1] el **sobreajuste** se puede evitar si se emplea la poda(**prunning**), la cual se trata de un proceso que elimina las ramas con características de poca importancia. Sin embargo, para [2] además de la poda tambien se puede evitar limitando el tamaño del arbol con el métod de parada temprana(**early stopping**).

En un ábol de decisión los nodos se vinculan o asocian a atributos del conjunto de datos. Para asociarlos o vincularlos se emplean dos métodos:**la ganancia de información y la impureza de Gini**. Por consiguiente los nodos del árbol actuan como criterios de división dentro del árbol.

La ganancia de información se basa en el concepto de entropía. La entropía mide la impureza de los valores de la muestra. Los valores de entropía pueden estar comprendidos entre 0 y 1. Si todas las muestras del conjunto de datos, pertenecen a una clase, entonces la entropía será igual a cero. Si la mitad de las muestras se clasifican en una clase y la otra mitad en otra, la entropía alcanzará su máximo en 1. Con el fin de seleccionar la mejor característica para dividir y encontrar el árbol de decisión óptimo, se debe utilizar el atributo con la menor cantidad de entropía. **La Ganancia de información** representa la diferencia en la entropía antes y después de una división en un atributo dado.

A juicio de [1] la impureza de GINI es la probabilidad de clasificar incorrectamente un punto de datos aleatorio del conjunto de datos si se etiquetara en función de la distribución de clases. Si pertenece a un clase su impureza es 0 [1] [4].

La ecuación de GINI es:

$Gini(D)=1- _{i=1}^m Pi^2 $ Donde Pi es la Probabilidad de una tupla o registro pertenezca a la clase

$IndiceGini = 1-(probabilidadCategoría1)^{2-(probabilidadCategoría2)}2 $

En [4] tenemos un ejemplo en el que el conjunto de datos tiene un total de 10 registros o tuplas pero de ellos sólo 3 son «demonios» y los 7 restantes son «asesinos», entonces:

- ***La probabilidad de la categoría 1 (demonios) sería de 3/10 = 0.3***
- ***La probabilidad de la categoría 2 (asesinos) sería de 7/10 = 0.7***

En este caso el índice GINI sería igual a: $IndiceGini=1-(3/10)^{2-(7/10)}2=1-0.3^{2-0.7}2=0.42 $

A criterio de [2] la prueba de bondad de ajuste chi-cuadrado ($X^{2} $) tambien se emplea en la división de nodos de un árbol para verificar si existe una diferencia significativa entre los nodos hijos y el nodo parental. Cuando el árbol se crea con estas condiciones se nombra como CHAID(Chi-square automatic interaction detector).

Ventajas de los árboles de decisión: Desde la posición de [1] son útiles para tareas de minería de datos y descubrimiento de conocimiento. Los árboles son fáciles de interpretar, requieren poca preparación en los datos y son muy flexibles. Para [3] las ventajas son:

- El árbol de decisión no tiene suposiciones sobre la distribución de los datos debido a la naturaleza no parámetrica del algoritmo.
- Puede captar patrones no lineales.

Desventajas de los árboles de decisión Desde el punto de vista de [1] los árboles de decisión son propensos al sobreajuste. Sin embargo, esto se puede evitar con procesos de poda(corte) previa a la construcción del árbol o posterior a su generación. Otra desventaja son los estimadores de alta varianza que ocacionan pequeñas variaciones dentro de los datos lo que produce un árbol de decisión diferente. Un método que reduce la varianza de los árboles de decisión es el ***embolsado*** o el ***promedio de estimaciones***. También, otra desventaja es el costo de su construcción. Desde el punto de vista de [3] son sensibles a datos ruidosos.

0.2 Ejemplo 1

Se tiene información de sillas infantiles de 400 tiendas distintas. Para cada una de las tiendas se han registrado 11 variables. Con base a estos datos se pretende generar un modelo de clasificación que permita predecir si una tienda tiene ventas altas **(ventas > 8) o bajas(ventas <=8)** en función de todas las variables disponibles. Los datos de la tienda estan disponibles en los dataset de ejemplo de la librería statmodels.

0.3 1. Carga del conjunto de datos

```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
```

```
2            8      183             64              0        0  23.3
3            1       89             66             23       94  28.1
4            0      137             40             35      168  43.1

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
3                     0.167   21        0
4                     2.288   33        1
```

2.Preprocesado Selección de características.

0.3.1 2.1 División de datos

0.4 3. Construcción del modelo de árbol de decisión

0.5 4. Evaluación del modelo

```
Accuracy: 0.6753246753246753
```

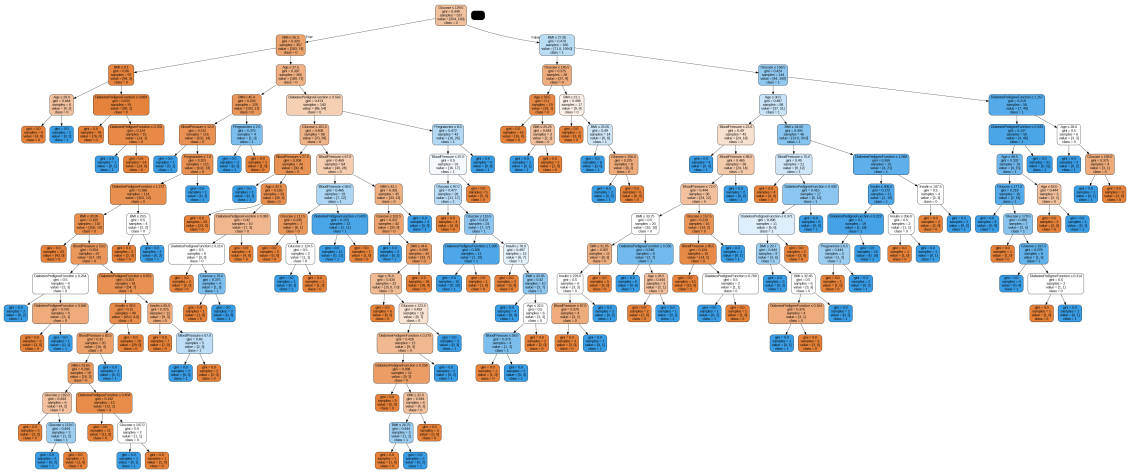
Obtuvimos un índice de clasificación del 67,53%, lo que se considera una buena precisión. Puedes mejorar esta precisión ajustando los parámetros del algoritmo del árbol de decisión.

0.6 5. Visualización del árbol de decisión

Se debe instalar la librería:

- pip install graphviz
- pip install pydotplu

```
Requirement already satisfied: six in /usr/local/lib/python3.11/dist-packages
(1.17.0)
```

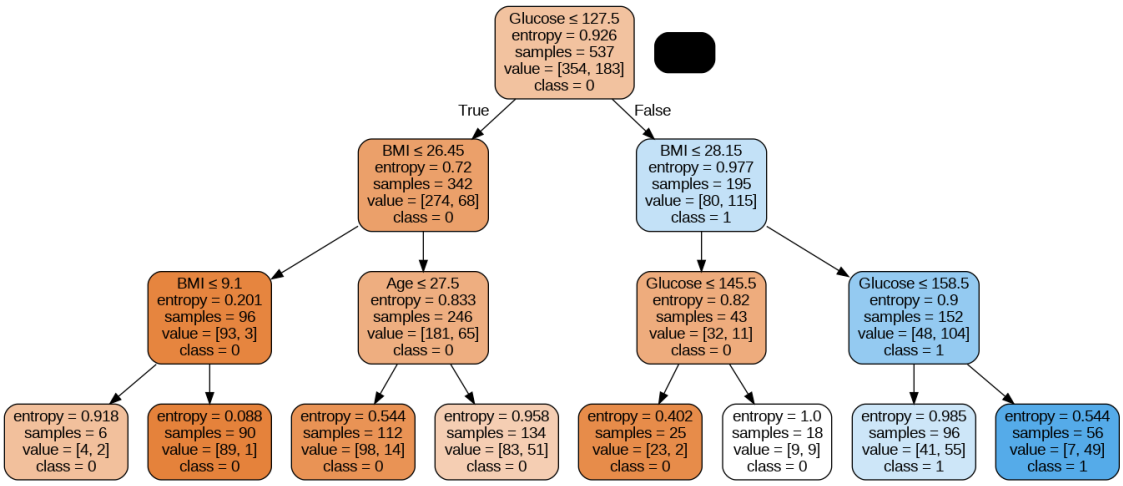


0.7 6. Optimazación del ábol de decisión:

Para optimizar el arbol de decisíon se pueden ajustar los hiperparámetros(Parámetros configurables por el usuario) del modelo. Algunos de estos parámetros son:

- **Criterio(Medida de selección de atributos):** Si queremos asociar los atributos por el ratio GINI se debe configurar así: **criterion=“gini”**. Si queremos utilizar el método de la ganancia de información debemos escribir: **criterion=“entropy”**.
- **Divisor(Estrategia de división):** configura la estrategia de división del arbol. Las estrategias admitidas son: **“mejor”** para elegir la mejor división y **“aleatoria”** para eligir la mejor decisión aleatoria.
- **Profundidad máxima del árbol:** este parámetro ajusta la expansión del árbol. Las opciones son: **valor_entero o None**. Si la opción es **“None”** los nodos se extienden hasta que todas la hojas contengan menos muestras que el parámetro **min_muetras_split**. Un valor entero muy alto genera **sobreajuste** y un valor entero muy bajo **infraajuste**.

```
Accuracy: 0.7705627705627706
```



0.8 Problema resuelto2:

```
(2008, 24)
```

	NUMERO	GENERO	TIPO_COLEGIO	LOCALIDAD_COLEGIO	CALENDARIO	MUNICIPIO	\
0	1	1	0	20	0	0	
1	2	1	2	2	1	19	
2	3	0	2	20	1	121	
3	4	0	1	19	1	19	
4	5	0	2	8	1	19	

	DEPARTAMENTO	PG_ICFES	CON_MAT_ICFES	APT_MAT_ICFES	…	BIOLOGIA_ICFES	\
0	0	474	66	66	…	58	

```
1             5       523             99             99  …             61
2            18       483             64             64  …             63
3             5       529             89             89  …             83
4             5       478             78             78  …             66

   SOCIALES_ICFES  FILOSOFIA_ICFES  IDIOMA_ICFES  LOCALIDAD  DISTANCIA  \
0               0               65            79         20          6
1               0               64            60          1          5
2               0               64            77         20          6
3               0               62            75         19          5
4               0               60            70          8          3

   INSCRIPCION  ESTRATO  ANO_INGRESO  RENDIMIENTO_UNO
0            9        3         2011                2
1            9        3         2011                2
2            9        3         2011                1
3            9        2         2011                1
4            9        3         2011                2

[5 rows x 24 columns]

      GENERO  TIPO_COLEGIO  LOCALIDAD_COLEGIO  CALENDARIO  MUNICIPIO  \
0          1             0                 20           0          0
1          1             2                  2           1         19
2          0             2                 20           1        121
3          0             1                 19           1         19
4          0             2                  8           1         19
…      …           …                 …          …         …
2003       0             0                  0           0          0
2004       1             0                  0           0          0
2005       0             0                  0           0          0
2006       0             0                  0           0          0
2007       1             0                  0           0          0

      DEPARTAMENTO  PG_ICFES  CON_MAT_ICFES  APT_MAT_ICFES  FISICA_ICFES  …  \
0                0       474             66             66            73  …
1                5       523             99             99            70  …
2               18       483             64             64            76  …
3                5       529             89             89            70  …
4                5       478             78             78            70  …
…            …       …             …              …           …  …
2003             0       342             68              0             0  …
2004             0       347             69              0             0  …
2005             0       348             67              0             0  …
2006             0       340             70              0             0  …
2007             0       351             69              0             0  …

      BIOLOGIA_ICFES  SOCIALES_ICFES  FILOSOFIA_ICFES  IDIOMA_ICFES  \
```

```
0                58              0              65              79
1                61              0              64              60
2                63              0              64              77
3                83              0              62              75
4                66              0              60              70
…               …               …              …               …
2003             68             70               0              58
2004             71             69               0              71
2005             67             76               0              68
2006             67             66               0              63
2007             63             78               0              70

      LOCALIDAD  DISTANCIA  INSCRIPCION  ESTRATO  ANO_INGRESO  RENDIMIENTO_UNO
0            20          6            9        3         2011                2
1             1          5            9        3         2011                2
2            20          6            9        3         2011                1
3            19          5            9        2         2011                1
4             8          3            9        3         2011                2
…         …          …          …       …          …             …
2003         11          5            9        3         2018                2
2004         19          5           10        1         2018                2
2005         11          5            9        2         2018                1
2006         11          5            9        3         2018                2
2007          7          5            9        2         2018                3

[2008 rows x 23 columns]
      GENERO  TIPO_COLEGIO  LOCALIDAD_COLEGIO  CALENDARIO  MUNICIPIO  \
0          1             0                 20           0          0
1          1             2                  2           1         19
2          0             2                 20           1        121
3          0             1                 19           1         19
4          0             2                  8           1         19
…       …            …                …          …         …
2003       0             0                  0           0          0
2004       1             0                  0           0          0
2005       0             0                  0           0          0
2006       0             0                  0           0          0
2007       1             0                  0           0          0

      DEPARTAMENTO  PG_ICFES  CON_MAT_ICFES  APT_MAT_ICFES  FISICA_ICFES  …  \
0                0       474             66             66            73  …
1                5       523             99             99            70  …
2               18       483             64             64            76  …
3                5       529             89             89            70  …
4                5       478             78             78            70  …
…             …        …             …              …             …  …
2003             0       342             68              0             0  …
```

```
2004             0       347             69              0             0  …
2005             0       348             67              0             0  …
2006             0       340             70              0             0  …
2007             0       351             69              0             0  …

      LITERATURA_ICFES  BIOLOGIA_ICFES  SOCIALES_ICFES  FILOSOFIA_ICFES  \
0                   68              58               0               65
1                   81              61               0               64
2                   65              63               0               64
3                   74              83               0               62
4                   71              66               0               60
…                  …               …               …               …
2003                71              68              70                0
2004                68              71              69                0
2005                69              67              76                0
2006                71              67              66                0
2007                71              63              78                0

      IDIOMA_ICFES  LOCALIDAD  DISTANCIA  INSCRIPCION  ESTRATO  ANO_INGRESO
0               79         20          6            9        3         2011
1               60          1          5            9        3         2011
2               77         20          6            9        3         2011
3               75         19          5            9        2         2011
4               70          8          3            9        3         2011
…              …          …          …           …       …          …
2003            58         11          5            9        3         2018
2004            71         19          5           10        1         2018
2005            68         11          5            9        2         2018
2006            63         11          5            9        3         2018
2007            70          7          5            9        2         2018

[2008 rows x 22 columns]

      RENDIMIENTO_UNO
0                   2
1                   2
2                   1
3                   1
4                   2
…                  …
2003                2
2004                2
2005                1
2006                2
2007                3

[2008 rows x 1 columns]

      GENERO  TIPO_COLEGIO  LOCALIDAD_COLEGIO  CALENDARIO  MUNICIPIO  \
```

```
0        1.0           0.0              20.0         0.0        0.0
1        1.0           2.0               2.0         1.0       19.0
2        0.0           2.0              20.0         1.0      121.0
3        0.0           1.0              19.0         1.0       19.0
4        0.0           2.0               8.0         1.0       19.0
…      …           …                 …          …         …
2003     0.0           0.0               0.0         0.0        0.0
2004     1.0           0.0               0.0         0.0        0.0
2005     0.0           0.0               0.0         0.0        0.0
2006     0.0           0.0               0.0         0.0        0.0
2007     1.0           0.0               0.0         0.0        0.0

      DEPARTAMENTO  PG_ICFES  CON_MAT_ICFES  APT_MAT_ICFES  FISICA_ICFES  …  \
0              0.0     474.0           66.0           66.0          73.0  …
1              5.0     523.0           99.0           99.0          70.0  …
2             18.0     483.0           64.0           64.0          76.0  …
3              5.0     529.0           89.0           89.0          70.0  …
4              5.0     478.0           78.0           78.0          70.0  …
…            …       …             …            …           …  …
2003           0.0     342.0           68.0            0.0           0.0  …
2004           0.0     347.0           69.0            0.0           0.0  …
2005           0.0     348.0           67.0            0.0           0.0  …
2006           0.0     340.0           70.0            0.0           0.0  …
2007           0.0     351.0           69.0            0.0           0.0  …

      LITERATURA_ICFES  BIOLOGIA_ICFES  SOCIALES_ICFES  FILOSOFIA_ICFES  \
0                 68.0            58.0             0.0             65.0
1                 81.0            61.0             0.0             64.0
2                 65.0            63.0             0.0             64.0
3                 74.0            83.0             0.0             62.0
4                 71.0            66.0             0.0             60.0
…                 …               …                …               …
2003              71.0            68.0            70.0              0.0
2004              68.0            71.0            69.0              0.0
2005              69.0            67.0            76.0              0.0
2006              71.0            67.0            66.0              0.0
2007              71.0            63.0            78.0              0.0

      IDIOMA_ICFES  LOCALIDAD  DISTANCIA  INSCRIPCION  ESTRATO  ANO_INGRESO
0             79.0       20.0        6.0          9.0      3.0       2011.0
1             60.0        1.0        5.0          9.0      3.0       2011.0
2             77.0       20.0        6.0          9.0      3.0       2011.0
3             75.0       19.0        5.0          9.0      2.0       2011.0
4             70.0        8.0        3.0          9.0      3.0       2011.0
…             …          …          …            …        …         …
2003          58.0       11.0        5.0          9.0      3.0       2018.0
2004          71.0       19.0        5.0         10.0      1.0       2018.0
2005          68.0       11.0        5.0          9.0      2.0       2018.0
```

```
2006          63.0       11.0        5.0          9.0      3.0       2018.0
2007          70.0        7.0        5.0          9.0      2.0       2018.0

[2008 rows x 22 columns]

[2 1 3 4]
[0 1 2 3]

c:\ProyectosPython\cienciadedatos\.venv\Lib\site-
packages\sklearn\preprocessing\_label.py:110: DataConversionWarning: A column-
vector y was passed when a 1d array was expected. Please change the shape of y
to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)

      GENERO  TIPO_COLEGIO  LOCALIDAD_COLEGIO  CALENDARIO  MUNICIPIO  \
0        1.0           0.0               20.0         0.0        0.0
1        1.0           2.0                2.0         1.0       19.0
2        0.0           2.0               20.0         1.0      121.0
3        0.0           1.0               19.0         1.0       19.0
4        0.0           2.0                8.0         1.0       19.0
…       …            …                 …           …         …
2003     0.0           0.0                0.0         0.0        0.0
2004     1.0           0.0                0.0         0.0        0.0
2005     0.0           0.0                0.0         0.0        0.0
2006     0.0           0.0                0.0         0.0        0.0
2007     1.0           0.0                0.0         0.0        0.0

      DEPARTAMENTO  PG_ICFES  CON_MAT_ICFES  APT_MAT_ICFES  FISICA_ICFES  …  \
0              0.0     474.0           66.0           66.0          73.0  …
1              5.0     523.0           99.0           99.0          70.0  …
2             18.0     483.0           64.0           64.0          76.0  …
3              5.0     529.0           89.0           89.0          70.0  …
4              5.0     478.0           78.0           78.0          70.0  …
…             …       …             …              …             …  …
2003           0.0     342.0           68.0            0.0           0.0  …
2004           0.0     347.0           69.0            0.0           0.0  …
2005           0.0     348.0           67.0            0.0           0.0  …
2006           0.0     340.0           70.0            0.0           0.0  …
2007           0.0     351.0           69.0            0.0           0.0  …

      LITERATURA_ICFES  BIOLOGIA_ICFES  SOCIALES_ICFES  FILOSOFIA_ICFES  \
0                 68.0            58.0             0.0             65.0
1                 81.0            61.0             0.0             64.0
2                 65.0            63.0             0.0             64.0
3                 74.0            83.0             0.0             62.0
4                 71.0            66.0             0.0             60.0
…                …              …               …              …
2003              71.0            68.0            70.0              0.0
2004              68.0            71.0            69.0              0.0
```

```
2005              69.0            67.0            76.0              0.0
2006              71.0            67.0            66.0              0.0
2007              71.0            63.0            78.0              0.0

      IDIOMA_ICFES  LOCALIDAD  DISTANCIA  INSCRIPCION  ESTRATO  ANO_INGRESO
0             79.0       20.0        6.0          9.0      3.0       2011.0
1             60.0        1.0        5.0          9.0      3.0       2011.0
2             77.0       20.0        6.0          9.0      3.0       2011.0
3             75.0       19.0        5.0          9.0      2.0       2011.0
4             70.0        8.0        3.0          9.0      3.0       2011.0
…              …          …          …            …        …          …
2003          58.0       11.0        5.0          9.0      3.0       2018.0
2004          71.0       19.0        5.0         10.0      1.0       2018.0
2005          68.0       11.0        5.0          9.0      2.0       2018.0
2006          63.0       11.0        5.0          9.0      3.0       2018.0
2007          70.0        7.0        5.0          9.0      2.0       2018.0

[2008 rows x 22 columns]

[[ 1.4259148  -0.9306264   1.2263107  …  0.21117978  0.9157391
  -1.5626022 ]
 [ 1.4259148   1.3240983  -1.1091399  …  0.21117978  0.9157391
  -1.5626022 ]
 [-0.70130426  1.3240983   1.2263107  …  0.21117978  0.9157391
  -1.5626022 ]
 …
 [-0.70130426 -0.9306264  -1.5720342  …  0.21117978 -0.57410145
   1.5225084 ]
 [-0.70130426 -0.9306264  -1.5720342  …  0.21117978  0.9157391
   1.5225084 ]
 [ 1.4259148  -0.9306264  -1.5720342  …  0.21117978 -0.57410145
   1.5225084 ]]

c:\ProyectosPython\cienciadedatos\.venv\Lib\site-
packages\numpy\_core\_methods.py:194: RuntimeWarning: overflow encountered in
multiply
  x = um.multiply(x, x, out=x)
c:\ProyectosPython\cienciadedatos\.venv\Lib\site-
packages\numpy\_core\_methods.py:205: RuntimeWarning: overflow encountered in
reduce
  ret = umr_sum(x, axis, dtype, out, keepdims=keepdims, where=where)

        GENERO  TIPO_COLEGIO  LOCALIDAD_COLEGIO  CALENDARIO  MUNICIPIO  \
0     1.425915     -0.930626           1.226311   -0.946527  -0.932239
1     1.425915      1.324098          -1.109140    1.024870   0.871811
2    -0.701304      1.324098           1.226311    1.024870   1.512984
3    -0.701304      0.708158           1.124151    1.024870   0.871811
4    -0.701304      1.324098          -0.155777    1.024870   0.871811
…         …            …                 …          …          …
```

```
2003 -0.701304     -0.930626          -1.572034   -0.946527  -0.932239
2004  1.425915     -0.930626          -1.572034   -0.946527  -0.932239
2005 -0.701304     -0.930626          -1.572034   -0.946527  -0.932239
2006 -0.701304     -0.930626          -1.572034   -0.946527  -0.932239
2007  1.425915     -0.930626          -1.572034   -0.946527  -0.932239

      DEPARTAMENTO  PG_ICFES  CON_MAT_ICFES  APT_MAT_ICFES  FISICA_ICFES  …  \
0        -0.935053  1.594288      -0.785448       0.659432      1.003832  …
1         0.881765  2.496176       2.968962       1.092583      0.977482  …
2         1.578680  1.762976      -1.023572       0.629451      1.029269  …
3         0.881765  2.603949       1.861875       0.971664      0.977482  …
4         0.881765  1.669436       0.615144       0.829086      0.977482  …
…              …       …              …              …          …  …
2003     -0.935053 -1.071123      -0.548742      -1.338546     -1.097676  …
2004     -0.935053 -0.962498      -0.430904      -1.338546     -1.097676  …
2005     -0.935053 -0.940858      -0.666921      -1.338546     -1.097676  …
2006     -0.935053 -1.114772      -0.313405      -1.338546     -1.097676  …
2007     -0.935053 -0.876105      -0.430904      -1.338546     -1.097676  …

      LITERATURA_ICFES  BIOLOGIA_ICFES  SOCIALES_ICFES  FILOSOFIA_ICFES  \
0             0.572083       -0.776711       -0.781967         1.027201
1             2.143884       -0.428105       -0.781967         1.017159
2             0.175104       -0.192069       -0.781967         1.017159
3             1.325292        2.316401       -0.781967         0.996666
4             0.955135        0.167292       -0.781967         0.975599
…                …              …               …               …
2003          0.955135        0.410327        1.287576        -1.096359
2004          0.572083        0.779947        1.284630        -1.096359
2005          0.701259        0.288467        1.304144        -1.096359
2006          0.955135        0.288467        1.275438        -1.096359
2007          0.955135       -0.192069        1.309282        -1.096359

      IDIOMA_ICFES  LOCALIDAD  DISTANCIA  INSCRIPCION   ESTRATO  ANO_INGRESO
0         1.263007   1.353436   1.295224     0.211180  0.915739    -1.562602
1        -0.309421  -2.362879   0.550878     0.211180  0.915739    -1.562602
2         1.094287   1.353436   1.295224     0.211180  0.915739    -1.562602
3         0.926277   1.201965   0.550878     0.211180 -0.574101    -1.562602
4         0.509468  -0.682326  -0.889589     0.211180  0.915739    -1.562602
…              …          …          …            …         …
2003     -0.470669  -0.117093   0.550878     0.211180  0.915739     1.522508
2004      0.592454   1.201965   0.550878     3.609206 -1.975565     1.522508
2005      0.344075  -0.117093   0.550878     0.211180 -0.574101     1.522508
2006     -0.065920  -0.117093   0.550878     0.211180  0.915739     1.522508
2007      0.509468  -0.883844   0.550878     0.211180 -0.574101     1.522508

[2008 rows x 22 columns]

        GENERO  TIPO_COLEGIO  LOCALIDAD_COLEGIO  CALENDARIO  MUNICIPIO  \
```

```
0     1.425915     -0.930626           1.226311   -0.946527  -0.932239
1     1.425915      1.324098          -1.109140    1.024870   0.871811
2    -0.701304      1.324098           1.226311    1.024870   1.512984
3    -0.701304      0.708158           1.124151    1.024870   0.871811
4    -0.701304      1.324098          -0.155777    1.024870   0.871811
…         …            …                  …          …          …
2003 -0.701304     -0.930626          -1.572034   -0.946527  -0.932239
2004  1.425915     -0.930626          -1.572034   -0.946527  -0.932239
2005 -0.701304     -0.930626          -1.572034   -0.946527  -0.932239
2006 -0.701304     -0.930626          -1.572034   -0.946527  -0.932239
2007  1.425915     -0.930626          -1.572034   -0.946527  -0.932239

      DEPARTAMENTO  PG_ICFES  CON_MAT_ICFES  APT_MAT_ICFES  FISICA_ICFES  …  \
0        -0.935053  1.594288      -0.785448       0.659432      1.003832  …
1         0.881765  2.496176       2.968962       1.092583      0.977482  …
2         1.578680  1.762976      -1.023572       0.629451      1.029269  …
3         0.881765  2.603949       1.861875       0.971664      0.977482  …
4         0.881765  1.669436       0.615144       0.829086      0.977482  …
…              …       …              …              …             …  …
2003     -0.935053 -1.071123      -0.548742      -1.338546     -1.097676  …
2004     -0.935053 -0.962498      -0.430904      -1.338546     -1.097676  …
2005     -0.935053 -0.940858      -0.666921      -1.338546     -1.097676  …
2006     -0.935053 -1.114772      -0.313405      -1.338546     -1.097676  …
2007     -0.935053 -0.876105      -0.430904      -1.338546     -1.097676  …

      BIOLOGIA_ICFES  SOCIALES_ICFES  FILOSOFIA_ICFES  IDIOMA_ICFES  \
0          -0.776711       -0.781967         1.027201      1.263007
1          -0.428105       -0.781967         1.017159     -0.309421
2          -0.192069       -0.781967         1.017159      1.094287
3           2.316401       -0.781967         0.996666      0.926277
4           0.167292       -0.781967         0.975599      0.509468
…               …               …                …             …
2003        0.410327        1.287576        -1.096359     -0.470669
2004        0.779947        1.284630        -1.096359      0.592454
2005        0.288467        1.304144        -1.096359      0.344075
2006        0.288467        1.275438        -1.096359     -0.065920
2007       -0.192069        1.309282        -1.096359      0.509468

      LOCALIDAD  DISTANCIA  INSCRIPCION   ESTRATO  ANO_INGRESO  \
0      1.353436   1.295224     0.211180  0.915739    -1.562602
1     -2.362879   0.550878     0.211180  0.915739    -1.562602
2      1.353436   1.295224     0.211180  0.915739    -1.562602
3      1.201965   0.550878     0.211180 -0.574101    -1.562602
4     -0.682326  -0.889589     0.211180  0.915739    -1.562602
…         …          …            …         …           …
2003  -0.117093   0.550878     0.211180  0.915739     1.522508
2004   1.201965   0.550878     3.609206 -1.975565     1.522508
2005  -0.117093   0.550878     0.211180 -0.574101     1.522508
```

```
2006  -0.117093   0.550878     0.211180  0.915739     1.522508
2007  -0.883844   0.550878     0.211180 -0.574101     1.522508

      RENDIMIENTO_UNO
0                   2
1                   2
2                   1
3                   1
4                   2
…                 …
2003                2
2004                2
2005                1
2006                2
2007                3

[2008 rows x 23 columns]
      PG_ICFES  CON_MAT_ICFES  FISICA_ICFES  QUIMICA_ICFES  IDIOMA_ICFES  \
0     1.594288      -0.785448      1.003832       1.045102      1.263007
1     2.496176       2.968962      0.977482       1.078446     -0.309421
2     1.762976      -1.023572      1.029269       1.102515      1.094287
3     2.603949       1.861875      0.977482       1.070248      0.926277
4     1.669436       0.615144      0.977482       0.964111      0.509468
…        …             …             …              …             …
2003 -1.071123      -0.548742     -1.097676      -1.097415     -0.470669
2004 -0.962498      -0.430904     -1.097676      -1.097415      0.592454
2005 -0.940858      -0.666921     -1.097676      -1.097415      0.344075
2006 -1.114772      -0.313405     -1.097676      -1.097415     -0.065920
2007 -0.876105      -0.430904     -1.097676      -1.097415      0.509468

      LOCALIDAD  RENDIMIENTO_UNO
0      1.353436                2
1     -2.362879                2
2      1.353436                1
3      1.201965                1
4     -0.682326                2
…         …                  …
2003  -0.117093                2
2004   1.201965                2
2005  -0.117093                1
2006  -0.117093                2
2007  -0.883844                3

[2008 rows x 7 columns]

[0 1 2 3]

c:\ProyectosPython\cienciadedatos\.venv\Lib\site-
```

```
packages\sklearn\preprocessing\_label.py:110: DataConversionWarning: A column-
vector y was passed when a 1d array was expected. Please change the shape of y
to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

```
Datos: son 1405 datos para entrenamiento y 603 datos para prueba
```

```
{'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': None,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'monotonic_cst': None,
 'random_state': None,
 'splitter': 'best'}
```

```
{'max_depth': [2, 3, 4, 5, 6, None],
 'min_samples_split': [2, 4, 6, 8, 10, 15],
 'min_samples_leaf': [1, 3, 5, 7, 9],
 'max_features': [2, 3, 4, 5, 6, 7, 'sqrt'],
 'splitter': ['better', 'random'],
 'random_state': [1, 5, 10],
 'criterion': ('gini', 'entropy')}
```

```
Resultado de GridSearchCV para el modelo
Mejor valor de exatitud usando kfold: 0.6248936170212767
Mejor valor de Hipérparametro usando parametros: {'criterion': 'gini',
'max_depth': 5, 'max_features': 5, 'min_samples_leaf': 3, 'min_samples_split':
8, 'random_state': 10, 'splitter': 'random'}
```

```
Rendimiento del modelo con datos de entrenamiento
Promedio de exactitud(accuracy) usando k-fold: 60.84650455927052 %
Desviación estandar de exactitud(accuracy) usando k-fold: 3.9514921199879462 %

Rendimiento del modelo con datos de prueba
Promedio de exactitud(accuracy) usando k-fold: 60.37158469945355 %
Desviación estandar de exactitud(accuracy) usando k-fold: 3.3452978860172213 %
```

0.9 Problema por resolver:

Debe resolver los siguiente problemas computacionales por medio de árboles de decisión. Los algoritmos propuestos deben mostrar claramente el modelo, las pruebas aplicadas sobre este y la interpretación de las métricas usadas para evaluar su rendimiento. Los códigos y la documentación del proceso deben deben entregarse mediante un PDF generado por jupyternotebook. Este docu-

mento debe llevar como anexo la gráfica del árbol generado.

1. Clasifique el conjunto de datos wine.Visite al documentación de lo dato en el siguiente enlace..
2. Clasifique el conjunto de datos Diabetes. La documentación reposa en el siguiente enlace. En este problema debe aplicar las siguientes **imposiciones**:

- Es obligatorio dividir este conjunto de datos en conjunto de entrenamiento y prueba.
- Debe predecir con datos del conjunto de prueba.
- Debe evaluar el modelo mediante métricas e interpretar los resultados.

3. Clasifique el conjunto de datos **breask_cancer**.Consulte su documentación en el siguiente enlace. Debe aplicar las anteriores imposiciones en este conjunto de datos.

Notas: 1. Debe estar presente el estudiante para resolverle las dudas del docente. 2. Debe escribir código con la sintaxis del lenguaje python y ordenar los código aplicando programación orientada a objetos.

1 Referencias

[1] IBM, «Qué es un árbol de decisión?» Accedido: 16 de mayo de 2025. [En línea]. Disponible en: https://www.ibm.com/es-es/think/topics/decision-trees

[2] J. Amat, «Árboles de decisión con Python: regresión y clasificación», https://cienciadedatos.net/. Accedido: 17 de mayo de 2025. [En línea]. Disponible en: https://cienciadedatos.net/documentos/py07_arboles_decision_python

[3] Datacamp, «Tutorial de Clasificación en Árbol de Decisión en Python». Accedido: 17 de mayo de 2025. [En línea]. Disponible en: https://www.datacamp.com/es/tutorial/decision-tree-classification-python

[4] M. Sotaquirá, «Codificando Bits», Codificando Bits. Accedido: 18 de mayo de 2025. [En línea]. Disponible en: https://codificandobits.com/blog/clasificacion-arboles-decision-algoritmo-cart/