ArbolDecision

March 16, 2023

0.1 Arbol decision

Jonathan Andres Pardo

Importar librerias

[ ]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
import matplotlib.pyplot as plt
```

Importar conjunto de datos el cual contiene la base de tumores y busca predecir sin un tumor es benigno o maligno https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[ ]:
```
dataset = pd.read_csv("Arbol/breast_cancer_data.csv")
```

Informacion del data set

[ ]:
```
dataset.head()
```

[ ]:

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	\
0	842302	M	17.99	10.38	122.80	1001.0	
1	842517	M	20.57	17.77	132.90	1326.0	
2	84300903	M	19.69	21.25	130.00	1203.0	
3	84348301	M	11.42	20.38	77.58	386.1	
4	84358402	M	20.29	14.34	135.10	1297.0	

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
0	0.11840	0.27760	0.3001	0.14710	
1	0.08474	0.07864	0.0869	0.07017	
2	0.10960	0.15990	0.1974	0.12790	
3	0.14250	0.28390	0.2414	0.10520	
4	0.10030	0.13280	0.1980	0.10430	

	…	texture_worst	perimeter_worst	area_worst	smoothness_worst	\
0	…	17.33	184.60	2019.0	0.1622	
1	…	23.41	158.80	1956.0	0.1238	
2	…	25.53	152.50	1709.0	0.1444	
3	…	26.50	98.87	567.7	0.2098	

```
4  …          16.67           152.20      1575.0           0.1374

   compactness_worst  concavity_worst  concave points_worst  symmetry_worst  \
0             0.6656           0.7119                0.2654          0.4601
1             0.1866           0.2416                0.1860          0.2750
2             0.4245           0.4504                0.2430          0.3613
3             0.8663           0.6869                0.2575          0.6638
4             0.2050           0.4000                0.1625          0.2364

   fractal_dimension_worst  Unnamed: 32
0                  0.11890          NaN
1                  0.08902          NaN
2                  0.08758          NaN
3                  0.17300          NaN
4                  0.07678          NaN

[5 rows x 33 columns]
```

[ ]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
```

```
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
 32  Unnamed: 32              0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

[ ]:
```
dataset.describe()
```

[ ]:

	id	radius_mean	texture_mean	perimeter_mean	area_mean	\
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	

	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	\
count	569.000000	569.000000	569.000000	569.000000	
mean	0.096360	0.104341	0.088799	0.048919	
std	0.014064	0.052813	0.079720	0.038803	
min	0.052630	0.019380	0.000000	0.000000	
25%	0.086370	0.064920	0.029560	0.020310	
50%	0.095870	0.092630	0.061540	0.033500	
75%	0.105300	0.130400	0.130700	0.074000	
max	0.163400	0.345400	0.426800	0.201200	

	symmetry_mean	…	texture_worst	perimeter_worst	area_worst	\
count	569.000000	…	569.000000	569.000000	569.000000	
mean	0.181162	…	25.677223	107.261213	880.583128	
std	0.027414	…	6.146258	33.602542	569.356993	
min	0.106000	…	12.020000	50.410000	185.200000	
25%	0.161900	…	21.080000	84.110000	515.300000	
50%	0.179200	…	25.410000	97.660000	686.500000	
75%	0.195700	…	29.720000	125.400000	1084.000000	
max	0.304000	…	49.540000	251.200000	4254.000000	

	smoothness_worst	compactness_worst	concavity_worst	\
count	569.000000	569.000000	569.000000	

```
mean            0.132369           0.254265         0.272188
std             0.022832           0.157336         0.208624
min             0.071170           0.027290         0.000000
25%             0.116600           0.147200         0.114500
50%             0.131300           0.211900         0.226700
75%             0.146000           0.339100         0.382900
max             0.222600           1.058000         1.252000

       concave points_worst  symmetry_worst  fractal_dimension_worst  \
count            569.000000      569.000000               569.000000
mean               0.114606        0.290076                 0.083946
std                0.065732        0.061867                 0.018061
min                0.000000        0.156500                 0.055040
25%                0.064930        0.250400                 0.071460
50%                0.099930        0.282200                 0.080040
75%                0.161400        0.317900                 0.092080
max                0.291000        0.663800                 0.207500

       Unnamed: 32
count          0.0
mean           NaN
std            NaN
min            NaN
25%            NaN
50%            NaN
75%            NaN
max            NaN

[8 rows x 32 columns]
```

Se limpia el dataset

[ ]:
```
dataset = dataset.drop(["id"], axis = 1)
dataset = dataset.drop(["Unnamed: 32"], axis = 1)
```

Graficos de analisis de los graficos de tumores benignos y malignos

Se crea un dataset con los tumores malignos y otro con los benignos

[ ]:
```
M = dataset[dataset.diagnosis == "M"]
```

[ ]:
```
M.head(5)
```

[ ]:
```
  diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
0         M        17.99         10.38          122.80     1001.0
1         M        20.57         17.77          132.90     1326.0
2         M        19.69         21.25          130.00     1203.0
3         M        11.42         20.38           77.58      386.1
4         M        20.29         14.34          135.10     1297.0
```

```
   smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
0          0.11840           0.27760          0.3001              0.14710
1          0.08474           0.07864          0.0869              0.07017
2          0.10960           0.15990          0.1974              0.12790
3          0.14250           0.28390          0.2414              0.10520
4          0.10030           0.13280          0.1980              0.10430

   symmetry_mean  …  radius_worst  texture_worst  perimeter_worst  \
0         0.2419  …         25.38          17.33           184.60
1         0.1812  …         24.99          23.41           158.80
2         0.2069  …         23.57          25.53           152.50
3         0.2597  …         14.91          26.50            98.87
4         0.1809  …         22.54          16.67           152.20

   area_worst  smoothness_worst  compactness_worst  concavity_worst  \
0      2019.0            0.1622             0.6656           0.7119
1      1956.0            0.1238             0.1866           0.2416
2      1709.0            0.1444             0.4245           0.4504
3       567.7            0.2098             0.8663           0.6869
4      1575.0            0.1374             0.2050           0.4000

   concave points_worst  symmetry_worst  fractal_dimension_worst
0                0.2654          0.4601                  0.11890
1                0.1860          0.2750                  0.08902
2                0.2430          0.3613                  0.08758
3                0.2575          0.6638                  0.17300
4                0.1625          0.2364                  0.07678

[5 rows x 31 columns]
```

[ ]:
```
B = dataset[dataset.diagnosis == "B"]
```

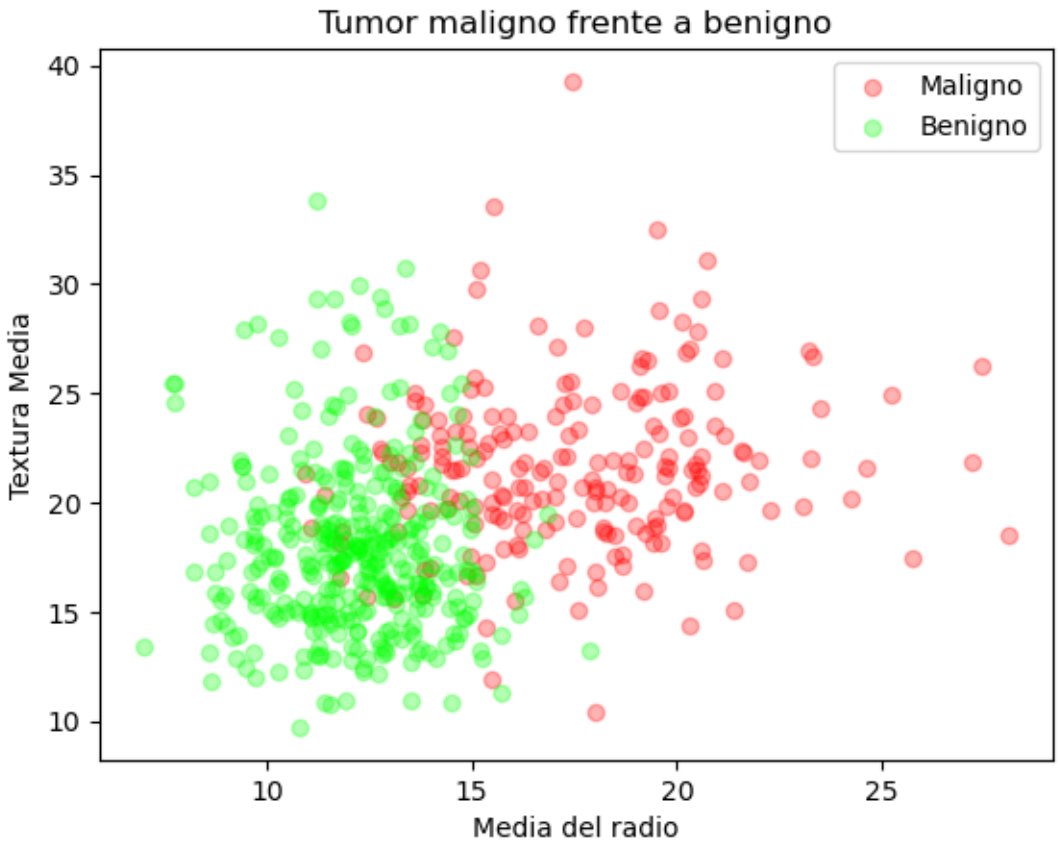
[ ]:
```
B.head(5)
```

[ ]:
```
   diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
19         B       13.540         14.36           87.46      566.3
20         B       13.080         15.71           85.63      520.0
21         B        9.504         12.44           60.34      273.9
37         B       13.030         18.42           82.61      523.8
46         B        8.196         16.84           51.71      201.9

    smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
19          0.09779           0.08129         0.06664              0.047810
20          0.10750           0.12700         0.04568              0.031100
21          0.10240           0.06492         0.02956              0.020760
37          0.08983           0.03766         0.02562              0.029230
```

```
46          0.08600           0.05943         0.01588              0.005917

    symmetry_mean  …  radius_worst  texture_worst  perimeter_worst  \
19         0.1885  …        15.110          19.26            99.70
20         0.1967  …        14.500          20.49            96.09
21         0.1815  …        10.230          15.66            65.13
37         0.1467  …        13.300          22.81            84.46
46         0.1769  …         8.964          21.96            57.26

    area_worst  smoothness_worst  compactness_worst  concavity_worst  \
19       711.2           0.14400            0.17730          0.23900
20       630.5           0.13120            0.27760          0.18900
21       314.9           0.13240            0.11480          0.08867
37       545.9           0.09701            0.04619          0.04833
46       242.2           0.12970            0.13570          0.06880

    concave points_worst  symmetry_worst  fractal_dimension_worst
19               0.12880          0.2977                  0.07259
20               0.07283          0.3184                  0.08183
21               0.06227          0.2450                  0.07773
37               0.05013          0.1987                  0.06169
46               0.02564          0.3105                  0.07409

[5 rows x 31 columns]
```

Se visualiza la dispersion entre los dos tipos de tumores

[ ]:
```
plt.title("Tumor maligno frente a benigno")
plt.xlabel("Media del radio")
plt.ylabel("Textura Media")
plt.scatter(M.radius_mean, M.texture_mean, color = "red", label = "Maligno",␣
 ↪alpha = 0.3)
plt.scatter(B.radius_mean, B.texture_mean, color = "lime", label = "Benigno",␣
 ↪alpha = 0.3)
plt.legend()
plt.show()
```



Normalizar Datos, se normalizan los diganosticos

```
[ ]: dataset.diagnosis = [1 if i == "M" else 0 for i in dataset.diagnosis]
```

```
[ ]: dataset
```

```
[ ]:      diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  \
     0            1        17.99         10.38          122.80     1001.0
     1            1        20.57         17.77          132.90     1326.0
     2            1        19.69         21.25          130.00     1203.0
     3            1        11.42         20.38           77.58      386.1
     4            1        20.29         14.34          135.10     1297.0
     ..         …          …             …              …          …
     564          1        21.56         22.39          142.00     1479.0
     565          1        20.13         28.25          131.20     1261.0
     566          1        16.60         28.08          108.30      858.1
     567          1        20.60         29.33          140.10     1265.0
     568          0         7.76         24.54           47.92      181.0

          smoothness_mean  compactness_mean  concavity_mean  concave points_mean  \
```

```
0           0.11840           0.27760         0.30010              0.14710
1           0.08474           0.07864         0.08690              0.07017
2           0.10960           0.15990         0.19740              0.12790
3           0.14250           0.28390         0.24140              0.10520
4           0.10030           0.13280         0.19800              0.10430
..              …                 …               …                    …
564         0.11100           0.11590         0.24390              0.13890
565         0.09780           0.10340         0.14400              0.09791
566         0.08455           0.10230         0.09251              0.05302
567         0.11780           0.27700         0.35140              0.15200
568         0.05263           0.04362         0.00000              0.00000

     symmetry_mean  …  radius_worst  texture_worst  perimeter_worst  \
0           0.2419  …        25.380          17.33           184.60
1           0.1812  …        24.990          23.41           158.80
2           0.2069  …        23.570          25.53           152.50
3           0.2597  …        14.910          26.50            98.87
4           0.1809  …        22.540          16.67           152.20
..             …  …           …               …                …
564         0.1726  …        25.450          26.40           166.10
565         0.1752  …        23.690          38.25           155.00
566         0.1590  …        18.980          34.12           126.70
567         0.2397  …        25.740          39.42           184.60
568         0.1587  …         9.456          30.37            59.16

     area_worst  smoothness_worst  compactness_worst  concavity_worst  \
0        2019.0           0.16220            0.66560           0.7119
1        1956.0           0.12380            0.18660           0.2416
2        1709.0           0.14440            0.42450           0.4504
3         567.7           0.20980            0.86630           0.6869
4        1575.0           0.13740            0.20500           0.4000
..          …                 …                  …                …
564      2027.0           0.14100            0.21130           0.4107
565      1731.0           0.11660            0.19220           0.3215
566      1124.0           0.11390            0.30940           0.3403
567      1821.0           0.16500            0.86810           0.9387
568       268.6           0.08996            0.06444           0.0000

     concave points_worst  symmetry_worst  fractal_dimension_worst
0                  0.2654          0.4601                  0.11890
1                  0.1860          0.2750                  0.08902
2                  0.2430          0.3613                  0.08758
3                  0.2575          0.6638                  0.17300
4                  0.1625          0.2364                  0.07678
..                    …               …                        …
564                0.2216          0.2060                  0.07115
565                0.1628          0.2572                  0.06637
```

```
566          0.1418          0.2218          0.07820
567          0.2650          0.4087          0.12400
568          0.0000          0.2871          0.07039

[569 rows x 31 columns]
```

[ ]:
```
x = dataset.drop(["diagnosis"], axis = 1)
y = dataset.diagnosis.values
```

Se separa el conjunto en dos partes el 80% se usara para entrenar el modelo el 20% para realizar pruebas

[ ]:
```
# Dividir el conjunto de datos en conjuntos de entrenamiento y prueba
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3,␣
 ↪random_state = 42)
```

Crear el modelo de árbol de decisión

[ ]:
```
dt = DecisionTreeClassifier()
```

Se entrena el modelo

[ ]:
```
dt.fit(x_train, y_train)
```

[ ]:
```
DecisionTreeClassifier()
```

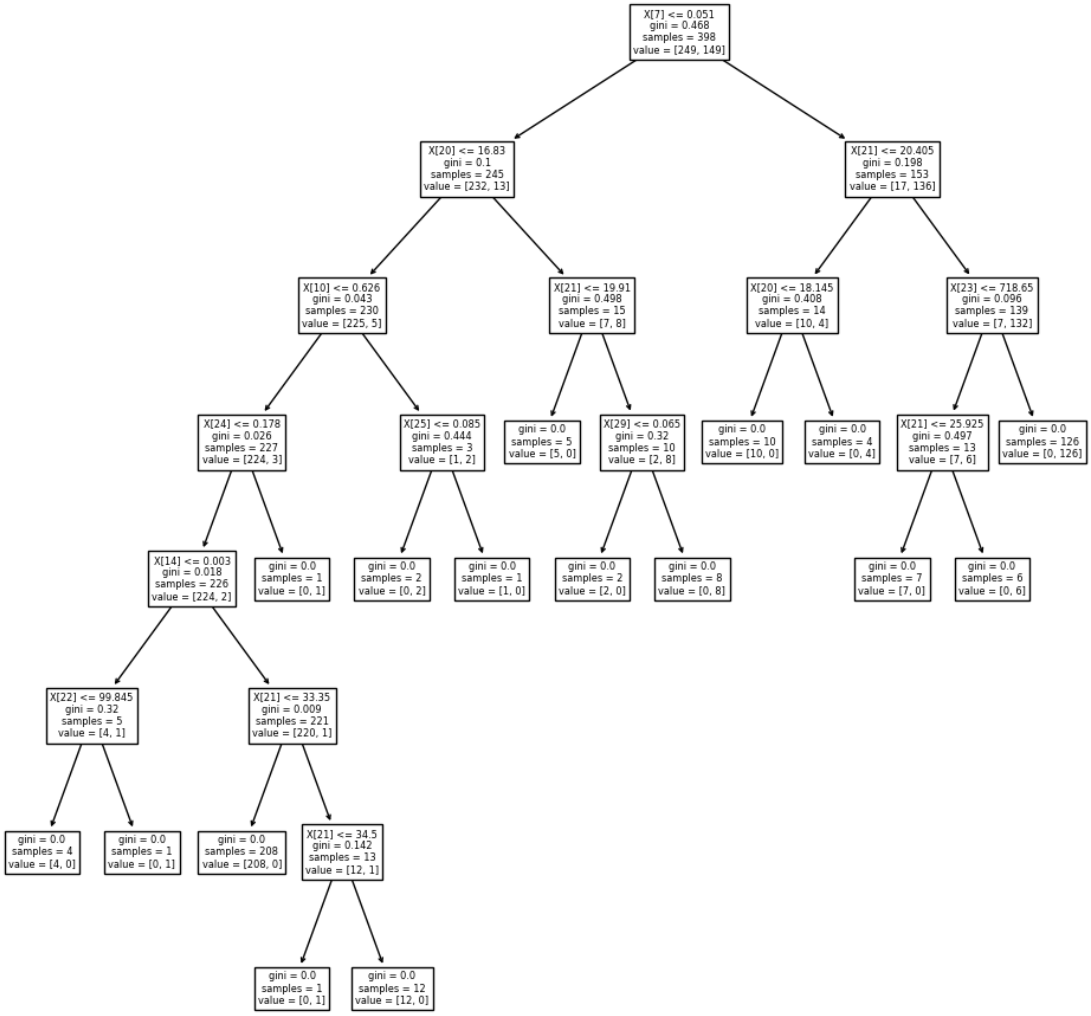
Prueba de prediccion con el dataset original

[ ]:
```
dt.score(x_test, y_test)
```

[ ]:
```
0.9239766081871345
```

Graficacion del arbol con el modelo entrenado

[ ]:
```
plt.figure(figsize=(12, 12))
tree.plot_tree(dt);
```


X[7] <= 0.051
gini = 0.468
samples = 398
value = [249, 149]
X[20] <= 16.83
gini = 0.1
samples = 245
value = [232, 13]
X[21] <= 20.405
gini = 0.198
samples = 153
value = [17, 136]
X[10] <= 0.626
gini = 0.043
samples = 230
value = [225, 5]
X[21] <= 19.91
gini = 0.498
samples = 15
value = [7, 8]
X[20] <= 18.145
gini = 0.408
samples = 14
value = [10, 4]
X[23] <= 718.65
gini = 0.096
samples = 139
value = [7, 132]
X[24] <= 0.178
gini = 0.026
samples = 227
value = [224, 3]
X[25] <= 0.085
gini = 0.444
samples = 3
value = [1, 2]
gini = 0.0
samples = 5
value = [5, 0]
X[29] <= 0.065
gini = 0.32
samples = 10
value = [2, 8]
gini = 0.0
samples = 10
value = [10, 0]
gini = 0.0
samples = 4
value = [0, 4]
X[21] <= 25.925
gini = 0.497
samples = 13
value = [7, 6]
gini = 0.0
samples = 126
value = [0, 126]
X[14] <= 0.003
gini = 0.018
samples = 226
value = [224, 2]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 2
value = [2, 0]
gini = 0.0
samples = 8
value = [0, 8]
gini = 0.0
samples = 7
value = [7, 0]
gini = 0.0
samples = 6
value = [0, 6]
X[22] <= 99.845
gini = 0.32
samples = 5
value = [4, 1]
X[21] <= 33.35
gini = 0.009
samples = 221
value = [220, 1]
gini = 0.0
samples = 4
value = [4, 0]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 208
value = [208, 0]
X[21] <= 34.5
gini = 0.142
samples = 13
value = [12, 1]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.0
samples = 12
value = [12, 0]