Árboles de Decisión

Algoritmia

2019/2020

Íñigo Olcoz

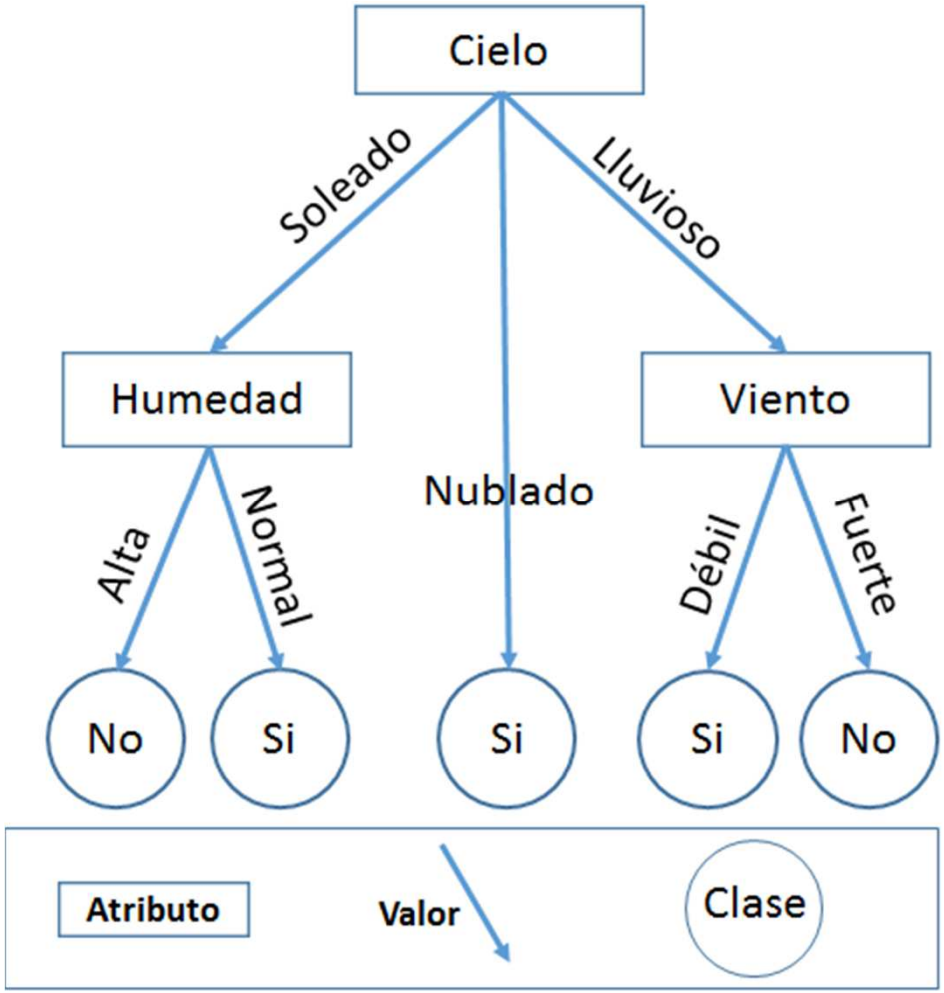
Árboles de Decisión

- Contexto

- Aprendizaje
- Heurística
- Discretización
- Inferencia
- Sobre-aprendizaje

- Problemas
  - Problema 1: Clasificación de tipo de estrella
  - Problema 2: Clasificación de tipo de vidrio
  - Problema 3: Diagnóstico de cáncer de mama
  - Problema 4: Predicción de vida de personaje literario

Contexto


Cielo
Soleado
Lluvioso
Humedad
Viento
Alta
Normal
Nublado
Débil
Fuerte
No
Si
Si
Si
No
Atributo
Valor
Clase

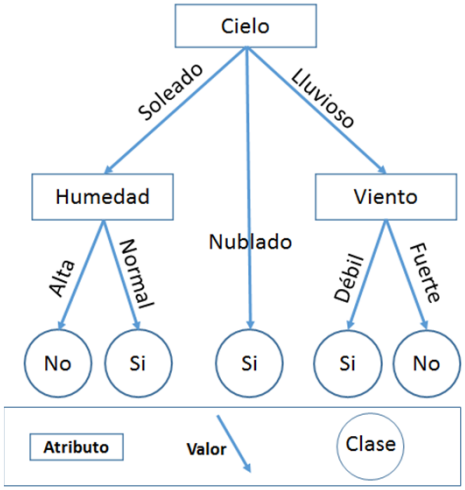
Contexto

- Inteligencia Artificial
- Aprendizaje Automático
- Aprendizaje Supervisado
- Clasificación
- Árboles de Decisión

Aprender del pasado para tomar decisiones en el futuro

Contexto

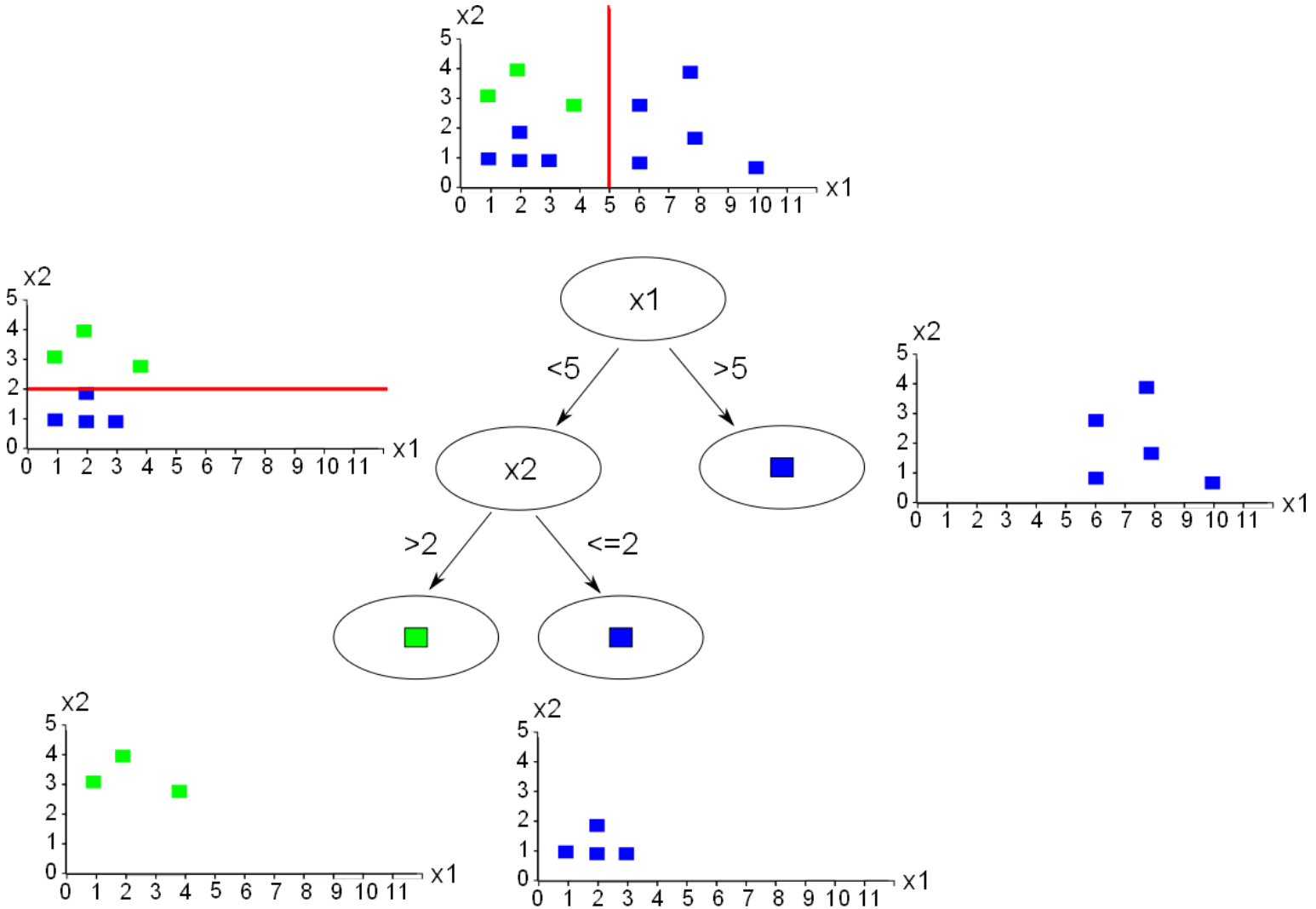
EJEMPLOS	ATRIBUTOS				CLASE
Día	Cielo	Temperatura	Humedad	Viento	Decisión
D1	soleado	calor	alta	débil	NO
D2	soleado	calor	alta	fuerte	NO
D3	nublado	calor	alta	débil	SÍ
D4	lluvia	templado	alta	débil	SÍ
D5	lluvia	frío	normal	débil	SÍ
D6	lluvia	frío	normal	fuerte	NO
D7	nublado	frío	normal	fuerte	SÍ
D8	soleado	templado	alta	débil	NO
D9	soleado	frío	normal	débil	SÍ
D10	lluvia	templado	normal	débil	SÍ
D11	soleado	templado	normal	fuerte	SÍ
D12	nublado	templado	alta	fuerte	SÍ
D13	nublado	calor	normal	débil	SÍ
D14	lluvia	templado	alta	fuerte	NO



Aprendizaje

- Los ejemplos de entrenamiento parten del nodo raíz
- Los atributos son categóricos: si son continuos, se discretizan previamente
- El atributo de la decisión se selecciona en base a una medida heurística: **ALGORITMO VORAZ**
- Los ejemplos se dividen recursivamente en base a los atributos elegidos: **ALGORITMO DIVIDE Y VENCERÁS**

Aprendizaje



Aprendizaje

Sea T el conjunto de ejemplos, C el conjunto de clases y A el conjunto de Atributos:

ConstruirÁrbol (T, C, A) :

- Crear un nodo RAÍZ para el árbol
- Si todos los ejemplos en T pertenecen a la misma clase Ci:
  - Devolver el nodo RAÍZ con etiqueta Ci
- Si A es conjunto vacío:
  - Devolver el nodo RAÍZ con etiqueta Ci donde Ci es la clase mayoritaria en T
- Elección heurística de a: Atributo de A que mejor clasifica T
- Etiquetar RAÍZ con a
- Para cada valor a_j de a:
  - Añadir una nueva rama bajo RAÍZ con la comprobación $a==a_j$
  - Sea Tj el subconjunto de T en donde su atributo a es a_j
  - ConstruirÁrbol (Tk, C, A-a)
- Devolver RAÍZ

Heurística

<u>Criterio heurístico</u>: Ganancia de la información en función de la entropía

Se persigue que en un nodo los ejemplos estén distribuidos **no-homogéneamente** entre las clases:

- Si los ejemplos están distribuidos homogéneamente entre las clases se tiene información poco interesante: nodo con alto grado de impureza
- Si los ejemplos están distribuidos no-homogéneamente entre las clases se tiene información muy interesante: nodo con bajo grado de impureza. (Si todos los nodos pertenecen a la misma clase: clasificación perfecta)

Heurística

Entropía: Medida del grado de impureza de un nodo

Entropía para un nodo t:

$$E(t) = \sum_{j=1}^{Clases} -p(j|t)\log_2(p(j|t))$$

- Nodo con entropía baja: bajo grado de impureza. Los ejemplos están muy separados en base a las clases
- Nodo con entropía alta: alto grado de impureza. Los ejemplos están poco separados en base a las clases

Heurística

Clase	Nº ejemplos de cada clase en el nodo
Clase A	0
Clase B	6

P(C1 | t) = 0/6 = 0 P(C2 | t) = 6/6 = 1

E = -0*log2(0)+-1*log2(1)= 0+0= 0

Clase	Nº ejemplos de cada clase en el nodo
Clase A	1
Clase B	5

P(C1 | t) = 1/6 P(C2 | t) = 5/6

E = -1/6*log2(1/6)+-5/6*log2(5/6)= 0.43+0.22 = 0.65

Clase	Nº ejemplos de cada clase en el nodo
Clase A	2
Clase B	4

P(C1 | t) = 2/6 P(C2 | t) = 4/6

E = -2/6*log2(2/6)+-4/6*log2(4/6)= 0.53+0.39= 0.92

Clase	Nº ejemplos de cada clase en el nodo
Clase A	3
Clase B	3

P(C1 | t) = 3/6 P(C2 | t) = 3/6

E = -3/6*log2(3/6)+-3/6*log2(3/6)= 0.5+0.5= 1

Heurística

Ganancia de la información: Cantidad de información que se gana (entropía que se pierde) si se divide por el atributo A con respecto a predecir las clases C

$$\boldsymbol{Gain(A) = E(C) - E(A)}$$

$$E(C) = \sum_{j=1}^{Clases} -\frac{n_j}{n} \log_2 \left(\frac{n_j}{n}\right)$$

$$E(A) = \sum_{i=1}^{particiones} \frac{n_i}{n} E(nodo_i)$$

Heurística

EJEMPLOS	ATRIBUTOS				CLASE
Día	Cielo	Temperatura	Humedad	Viento	Decisión
D1	soleado	calor	alta	débil	NO
D2	soleado	calor	alta	fuerte	NO
D3	nublado	calor	alta	débil	SÍ
D4	lluvia	templado	alta	débil	SÍ
D5	lluvia	frío	normal	débil	SÍ
D6	lluvia	frío	normal	fuerte	NO
D7	nublado	frío	normal	fuerte	SÍ
D8	soleado	templado	alta	débil	NO
D9	soleado	frío	normal	débil	SÍ
D10	lluvia	templado	normal	débil	SÍ
D11	soleado	templado	normal	fuerte	SÍ
D12	nublado	templado	alta	fuerte	SÍ
D13	nublado	calor	normal	débil	SÍ
D14	lluvia	templado	alta	fuerte	NO

Atributo Cielo		
Soleado: 5	Sí: 2	E_soleado = 0,971
	No: 3	
Nublado: 4	Sí: 4	E_nublado = 0
	No: 0	
Lluvia: 5	Sí: 3	E_lluvia = 0,971
	No: 2	

$$E(C) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0{,}940$$

$$E(Cielo) = \frac{5}{14}E_{soleado} + \frac{4}{14}E_{nublado} + \frac{5}{14}E_{lluvia} = 0{,}694$$

$$Gain(Cielo) = 0{,}940 - 0{,}694 = 0{,}246$$

Heurística

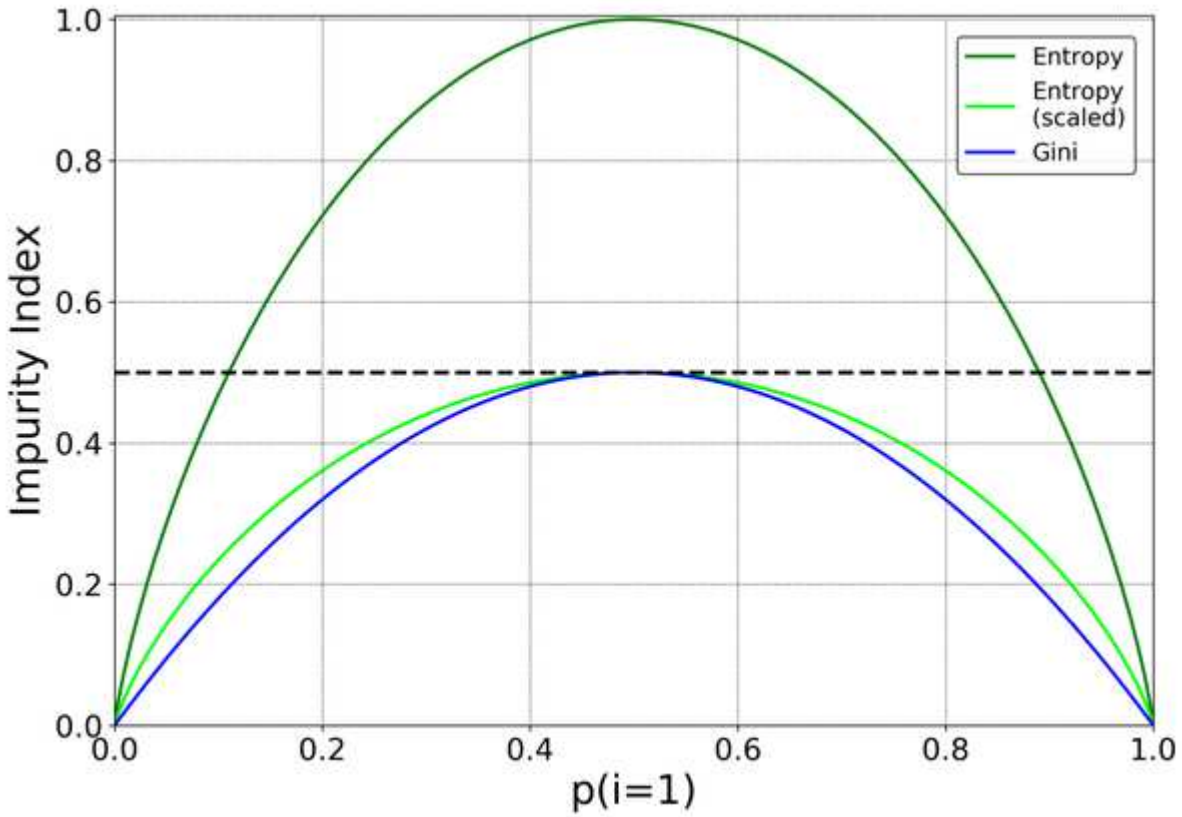
Ratio de Ganancia: Intenta evitar favorecer a los atributos con más valores. Penaliza la creación de muchas particiones pequeñas al ajustar la ganancia de información por la entropía de la división

$$\boldsymbol{GainRatio(A)} = \frac{\boldsymbol{Gain(A)}}{\boldsymbol{SplitInfo(A)}}$$

$$SplitInfo(A) = \sum_{j=1}^{particiones} -\frac{n_j}{n}\log_2\left(\frac{n_j}{n}\right)$$

Heurística

- Ratio de Ganancia de Información: Árbol de Decisión C4.5
- Índice de Gini: Árbol de Decisión CART



Discretización

Los atributos son categóricos. Si son continuos, se discretizan previamente:

- Ordenar los valores del atributo
- Calcular la entropía para cada pareja de valores en los que la clase cambie
- Elegir como umbral la media del par de valores que minimice la entropía

Atributo	20	23	23	27	30	40	52	63	65	70
Clase	NO	SÍ	SÍ	SÍ	NO	NO	SÍ	SÍ	SÍ	SÍ

Discretización

Atributo	20	23	23	27	30	40	52	63	65	70
Clase	NO	SÍ	SÍ	SÍ	NO	NO	SÍ	SÍ	SÍ	SÍ

- $E(23) = \frac{1}{10} * 0 + \frac{9}{10} * \left(\frac{7}{9} * log2\left(\frac{9}{7}\right) + \frac{2}{9} * log2\left(\frac{9}{2}\right)\right) = 0.6878$
- $E(30) = \frac{4}{10} * \left(\frac{3}{4} * log2\left(\frac{4}{3}\right) + \frac{1}{4} * log2\left(\frac{4}{1}\right)\right) + \frac{6}{10} * \left(\frac{4}{6} * log2\left(\frac{6}{4}\right) + \frac{2}{6} * log2\left(\frac{6}{2}\right)\right) = 0.8755$
- $E(52) = \frac{6}{10} * \left(\frac{3}{6} * log2\left(\frac{6}{3}\right) + \frac{3}{6} * log2\left(\frac{6}{3}\right)\right) + \frac{4}{10} * \left(\frac{4}{4} * log2\left(\frac{4}{4}\right) + \frac{0}{4} * log2\left(\frac{4}{0}\right)\right) = 0.6$

$$Umbral = Umbral\ (52) = \frac{40 + 52}{2} = 46$$

Inferencia

Hacer uso del Árbol de Decisión entrenado para predecir la clase de nuevos ejemplos:

Ejemplo Tk con atributos A → Árbol de Decisión → Clase Ck

Importancia de la selección y preparación de los datos:

- Calidad de los datos
- Exploración de los datos
- Pre-procesamiento
- Conjunto de ejemplos de Entrenamiento
- Conjunto de ejemplos de Validación

Sobre-aprendizaje

Los árboles de decisión pueden presentar tendencia al sobre-entrenamiento:

- El árbol refleja anomalías del conjunto de entrenamiento (ruido, outliers)
- El árbol resulta más complejo de lo que debería ser
- Disminuye la precisión del clasificador en nuevas inferencias

Técnica de poda:

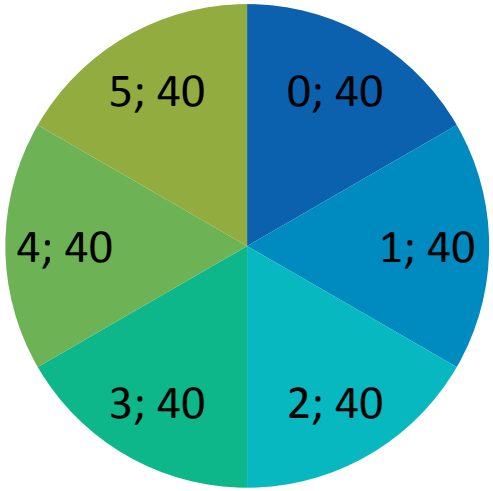
- Parte de un árbol de decisión completamente desarrollado y elimina las partes de poca calidad en función del error
- Recorre el árbol examinando los nodos desde las hojas hasta la raíz y reemplaza un subárbol por una hoja etiquetada con las clase mayoritaria, en caso de que el error en el reemplazo sea menor que el error sin poda

Problemas

- Problema 1: Clasificación de tipo de estrella
- Problema 2: Clasificación de tipo de vidrio
- Problema 3: Diagnóstico de cáncer de mama
- Problema 4: Predicción de vida de personaje literario

Clasificación tipo de estrella

- **Clases:** Tipo de estrella
  - Brown Dwarf -> Tipo = 0
  - Red Dwarf -> Tipo = 1
  - White Dwarf-> Tipo = 2
  - Main Sequence -> Tipo = 3
  - Supergiant -> Tipo = 4
  - Hypergiant -> Tipo = 5



- **Atributos:**
  1. Temperatura (K): Temperatura de la superficie
  2. Luminosidad (L/Lo): Luminosidad calculada respecto al sol
  3. Radio (R/Ro): Radio calculado respecto al sol
  4. Magnitud Visual Absoluta: Medida del brillo si el cuerpo celeste estuviera a 10 párseces
  5. Clase espectral: Clasificación Morgan-Keenan (O, B, A, F, G, K, M)

Clasificación tipo de estrella

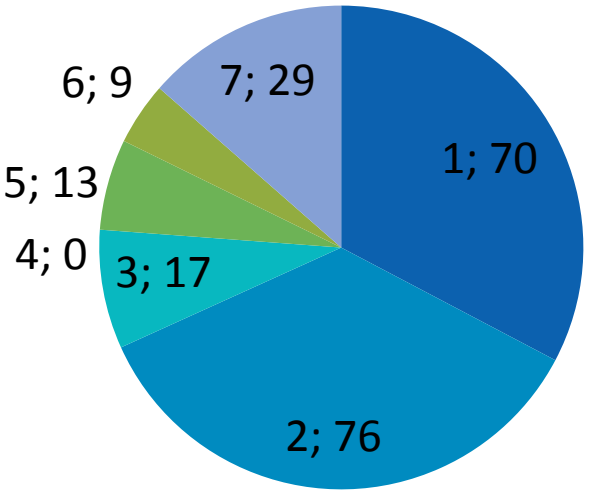
- **Ejemplos:** 240

	Temperature (K)	Luminosity(L/Lo)	Radius(R/Ro)	Absolute magnitude(Mv)	Star type	Spectral Class
0	3068	0.002400	0.1700	16.12	0	M
1	3042	0.000500	0.1542	16.60	0	M
2	2600	0.000300	0.1020	18.70	0	M
3	2800	0.000200	0.1600	16.65	0	M
4	1939	0.000138	0.1030	20.06	0	M
5	2840	0.000650	0.1100	16.98	0	M
6	2637	0.000730	0.1270	17.22	0	M
7	2600	0.000400	0.0960	17.40	0	M
8	2650	0.000690	0.1100	17.45	0	M
9	2700	0.000180	0.1300	16.05	0	M

Clasificación tipo de vidrio

- **Clases:** Tipo de vidrio
  - 7 tipos de vidrio diferentes

- **Atributos:**
  1. RI: Índice de refracción
  2. Na: Sodio (% en peso de óxido)
  3. Mg: Magnesio (% en peso de óxido)
  4. Al: Aluminio (% en peso de óxido)
  5. Si: Silicio (% en peso de óxido)
  6. K: Potasio (% en peso de óxido)
  7. Ca: Calcio (% en peso de óxido)
  8. Ba: Bario (% en peso de óxido)
  9. Fe: Hierro (% en peso de óxido)



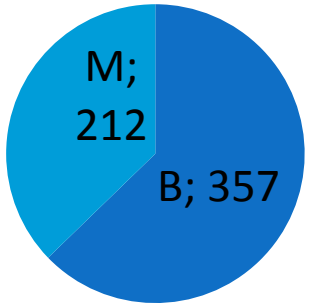
Clasificación tipo de vidrio

- **Ejemplos:** 214

	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Type
0	1.52101	13.64	4.49	1.10	71.78	0.06	8.75	0.0	0.00	1
1	1.51761	13.89	3.60	1.36	72.73	0.48	7.83	0.0	0.00	1
2	1.51618	13.53	3.55	1.54	72.99	0.39	7.78	0.0	0.00	1
3	1.51766	13.21	3.69	1.29	72.61	0.57	8.22	0.0	0.00	1
4	1.51742	13.27	3.62	1.24	73.08	0.55	8.07	0.0	0.00	1
5	1.51596	12.79	3.61	1.62	72.97	0.64	8.07	0.0	0.26	1
6	1.51743	13.30	3.60	1.14	73.09	0.58	8.17	0.0	0.00	1
7	1.51756	13.15	3.61	1.05	73.24	0.57	8.24	0.0	0.00	1
8	1.51918	14.04	3.58	1.37	72.08	0.56	8.30	0.0	0.00	1
9	1.51755	13.00	3.60	1.36	72.99	0.57	8.40	0.0	0.11	1

Diagnóstico cáncer de mama

- **Clases:** Diagnóstico
  - B: Benigno
  - M: Maligno



- **Atributos:**
  1. Radio: Distancia media del centro a los puntos del perímetro
  2. Textura: Desviación estándar en los valores de grises
  3. Perímetro: Tamaño del tumor
  4. Área: Área del tumor
  5. Tersura: Promedio de la variación local de las longitudes del radio
  6. Compactación: Perímetro^2 / áreas – 1
  7. Concavidad: Grado de las porciones cóncavas del perímetro
  8. Puntos cóncavos: Número de porciones cóncavas del contorno
  9. Simetría: Medida del grado de simetría
  10. Dimensión fractal: Estimación fractal de la dimensión

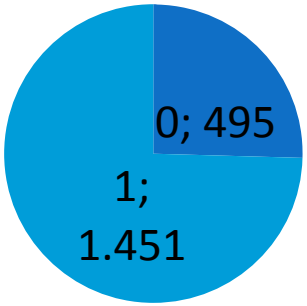
Diagnóstico cáncer de mama

- **Ejemplos:** 569

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean
0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419
1	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812
2	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069
3	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597
4	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809
5	M	12.45	15.70	82.57	477.1	0.12780	0.17000	0.15780	0.08089	0.2087
6	M	18.25	19.98	119.60	1040.0	0.09463	0.10900	0.11270	0.07400	0.1794
7	M	13.71	20.83	90.20	577.9	0.11890	0.16450	0.09366	0.05985	0.2196
8	M	13.00	21.82	87.50	519.8	0.12730	0.19320	0.18590	0.09353	0.2350
9	M	12.46	24.04	83.97	475.9	0.11860	0.23960	0.22730	0.08543	0.2030

Predicción vida de personaje

- **Clases:** Está vivo
  - 0: No
  - 1: Sí



- **Atributos:**
  1. Masculino: Booleano sobre el género
  2. Libro1: Booleano sobre presencia en Libro 1
  3. Libro2: Booleano sobre presencia en Libro 2
  4. Libro3: Booleano sobre presencia en Libro 3
  5. Libro4: Booleano sobre presencia en Libro 4
  6. Libro5: Booleano sobre presencia en Libro 5
  7. Matrimonio: Booleano sobre matrimonio
  8. Nobleza: Booleano sobre pertenencia a la nobleza
  9. Muertes relacionadas: Número de muertes con las que ha guardado relación
  10. Popularidad: Grado de popularidad del personaje

Predicción vida de personaje

- **Ejemplos:** 1946

	male	book1	book2	book3	book4	book5	isMarried	isNoble	numDeadRelations	popularity	isAlive
0	1	0	0	0	0	0	0	0	11	0.605351	0
1	1	1	1	1	1	1	1	1	1	0.896321	1
2	1	0	0	0	1	0	0	1	0	0.267559	1
3	0	0	0	0	0	0	1	1	0	0.183946	0
4	0	0	0	0	1	0	1	1	0	0.043478	1
5	1	0	0	0	0	0	0	0	5	1.000000	1
6	1	0	0	0	0	0	1	1	0	0.431438	0
7	1	0	0	0	0	0	0	0	5	0.678930	0
8	1	0	0	1	0	0	0	1	0	0.006689	0
9	1	0	0	0	0	0	0	1	0	0.020067	1