THE BRIDGE

Feature selection

Relación entre variables

- Existen diversas técnicas para conocer la dependencia o independencia entre variables

	CATEGÓRICAS	NUMÉRICAS
CATEGÓRICAS	**CHI-CUADRADO**	ANOVA
NUMÉRICAS	ANOVA	CORRELACIÓN

Test chi-cuadrado χ^2

- Permite saber si existe relación entre variables de tipo cualitativo
- Si al final del estudio concluimos que las variables no están relacionadas, podremos decir con un determinado nivel de confianza, previamente fijado, que ambas son independientes
- Ejemplo: Relación entre el sexo de una persona y su pasión por el fútbol (BAJA, MEDIA, ALTA)
- Los resultados se suelen presentar a modo de tablas de doble entrada que reciben el nombre de tablas de contingencia

	ALTA	MEDIA	BAJA	
HOMBRE	**5**	**3**	**8**	16
MUJER	**1**	**10**	**3**	14
	6	13	11	

Test chi-cuadrado χ^2

- Para el cómputo de χ^2 es necesario calcular las frecuencias esperadas (aquellas que deberían haberse observado si la hipótesis de independencia fuese cierta), y compararlas con las frecuencias observadas en la realidad.

- El valor del estadístico se calcula según la siguiente ecuación:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- El valor esperado se calcula como el producto de los totales marginales dividido por el número total de casos

Expected = (16x6)/30 = 3,2

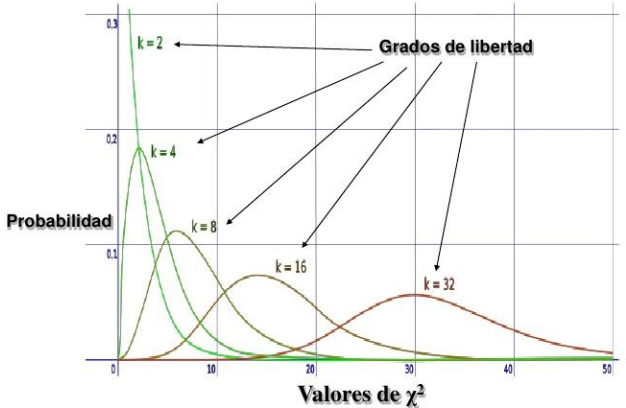
	ALTA	MEDIA	BAJA	
HOMBRE	5 3,2	3 6,9	8 5,9	16
MUJER	1 2,8	10 6,1	3 5,1	14
	6	13	11	TOTAL: 30

Test chi-cuadrado χ^2

- De esta forma, el valor χ2 mide la diferencia entre el valor que debiera resultar si las dos variables fuesen independientes y el que se ha observado en la realidad. Cuanto mayor sea esa diferencia (y, por lo tanto, el valor de χ2), mayor será la relación entre ambas variables.

- En el caso de no existir dependencia estadística, los valores de χ2 se distribuyen según una distribución denominada ji-cuadrado, que depende de un parámetro llamado “grados de libertad”. Para el caso de una tabla de contingencia de r filas y k columnas, los grados de libertad se calculan como (r-1) x (k-1)

- Si las variables no tienen relación, el valor de χ2 obtenido estará dentro del rango de mayor probabilidad según esta distribución



Test chi-cuadrado χ^2

DISTRIBUCION DE χ^2

Grados de libertad	Probabilidad										
	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,01	0,001
1	0,004	0,02	0,06	0,15	0,46	1,07	1,64	2,71	3,84	6,64	10,83
2	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	9,21	13,82
3	0,35	0,58	1,01	1,42	2,37	3,66	4,64	6,25	7,82	11,34	16,27
4	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	13,28	18,47
5	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,07	15,09	20,52
6	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,64	12,59	16,81	22,46
7	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,02	14,07	18,48	24,32
8	2,73	3,49	4,59	5.53	7,34	9,52	11,03	13,36	15,51	20,09	26,12
9	3,32	4,17	5,38	6,39	8,34	10,66	12,24	14,68	16,92	21,67	27,88
10	3,94	4,86	6,18	7,27	9,34	11,78	13,44	15,99	18,31	23,21	29,59
	No significativo								Significativo		

Test chi-cuadrado χ^2

```
        Alta Media Baja
Hombre     5     3    8
Mujer      1    10    3
> chisq.test(datos)

        Pearson's Chi-squared test

data:  datos
X-squared = 8.6136, df = 2, p-value = 0.01348
```

Relación entre variables

- Existen diversas técnicas para conocer la dependencia o independencia entre variables

	CATEGÓRICAS	NUMÉRICAS
CATEGÓRICAS	CHI-CUADRADO	**ANOVA**
NUMÉRICAS	**ANOVA**	CORRELACIÓN

ANOVA

- El análisis de varianza (ANOVA) es un método de prueba de igualdad de tres o más medias poblacionales, por medio del análisis de las varianzas muestrales

- Ejemplo: Estamos interesados en conocer si hay colores más atractivos para los insectos. Para ello se diseñaron trampas con los siguientes colores: amarillo, azul, blanco y verde. Se cuantifica el número de insectos que quedaban atrapados

- Hipótesis: todas las medias son iguales (las variables no están relacionadas)

ANOVA

	AMARILLO	AZUL	BLANCO	VERDE
Dia 1	10	11	12	10
Dia 2	12	14	13	11
Dia 3	18	19	17	16
Dia 4	24	23	25	23
Dia 5	36	38	37	36

- Conclusión: En este caso son los días los que marcan la diferencia, no el color de las trampas

ANOVA

	AMARILLO	AZUL	BLANCO	VERDE
Dia 1	29	17	10	1
Dia 2	29	18	11	3
Dia 3	30	19	12	2
Dia 4	31	19	12	4
Dia 5	31	20	13	2

- No hay diferencias grandes dentro de cada grupo, pero sí entre los grupos
- En este caso, sí hay relación entre el color de la trampa y los insectos

Relación entre variables

- Existen diversas técnicas para conocer la dependencia o independencia entre variables

	CATEGÓRICAS	NUMÉRICAS
CATEGÓRICAS	CHI-CUADRADO	ANOVA
NUMÉRICAS	ANOVA	**CORRELACIÓN**

Relación entre variables numéricas

- Ejemplo: Se dispone de los siguientes datos acerca de las horas de sueño y el peso de una serie de personas

Horas de sueño	Peso (kg)
7	74
4	50
12	89
11	84
8	65
6	60
11	70
5	52

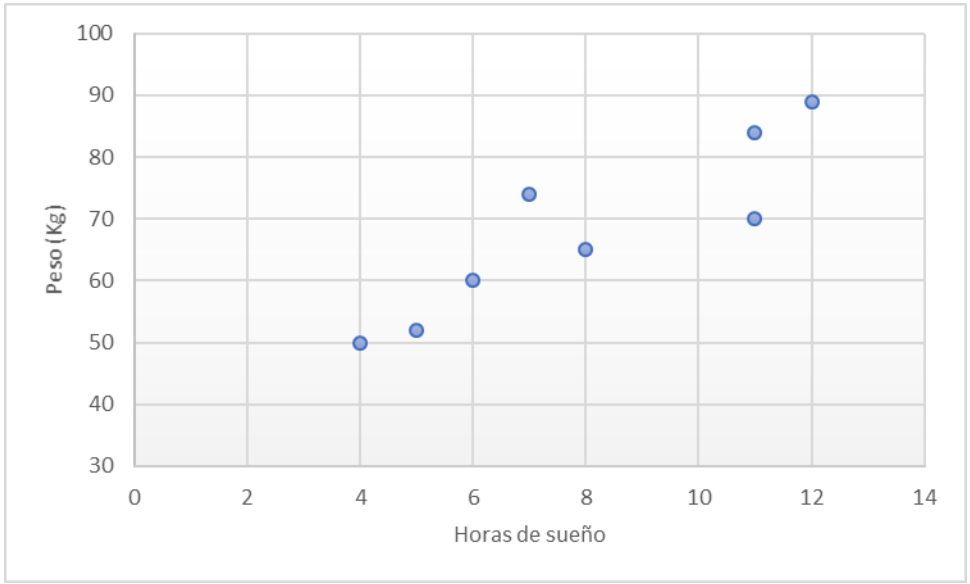


Gráfico de dispersión

Coeficiente de correlación

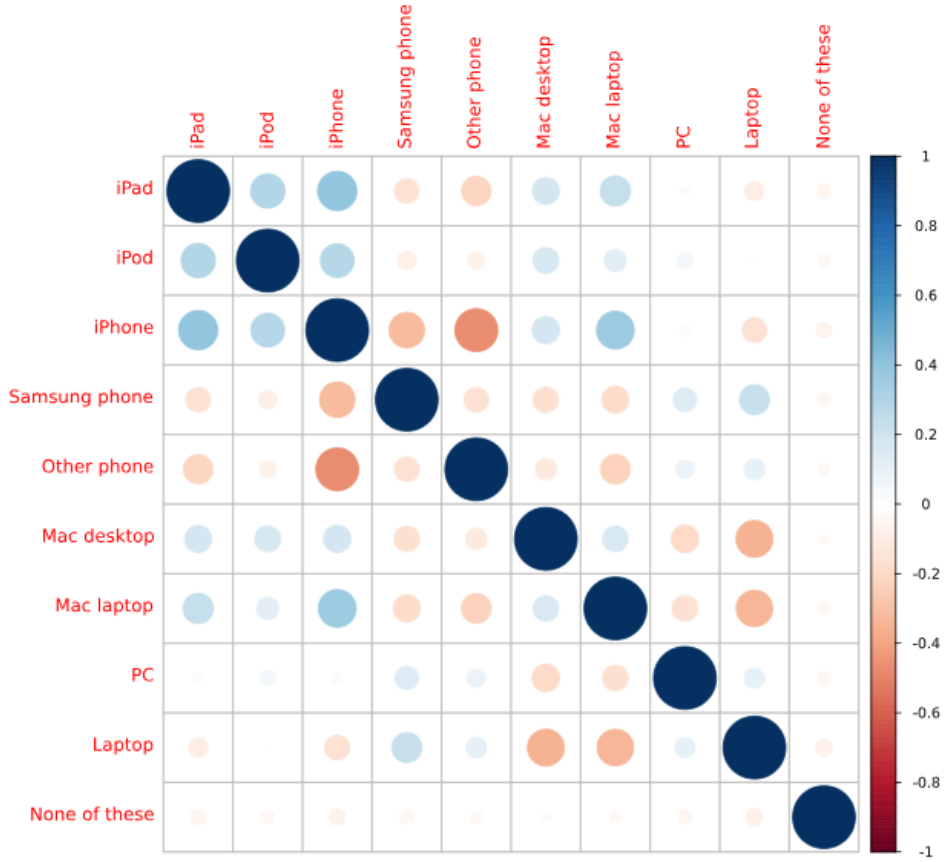
- Un coeficiente de correlación mide el grado en que dos variables tienden a cambiar al mismo tiempo. El coeficiente describe tanto la fuerza como la dirección de la relación

- La correlación de Pearson evalúa la relación lineal entre dos variables continuas. Una relación es lineal cuando un cambio en una variable se asocia con un cambio proporcional en la otra variable

$$\rho(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y - \bar{y})^2}}$$

- El valor está comprendido entre -1 y +1

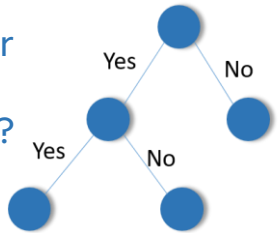
Matriz de correlación



Árboles de Decisión

EJEMPLO: Queremos clasificar los ingresos de los jugadores en dos categorías: ALTOS y BAJOS

¿Qué variable utilizar para segmentar en cada nodo del árbol?



Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Árboles de Decisión

- Hay que medir cómo de bien separan las variables candidatos a la variable objetivo
- Normalmente, ninguna de las variables consigue separar perfectamente a la variable objetivo (existe impureza)
- La métrica más común para medir impurezas se conoce como “**Gini**”

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

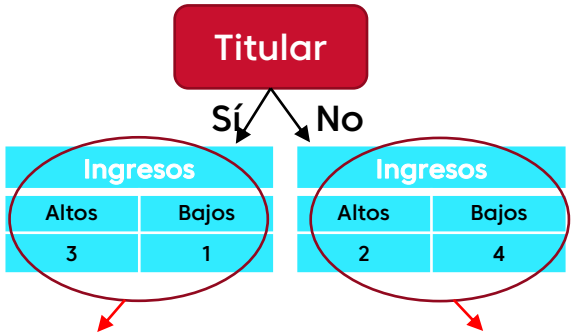
¿Qué variable tiene menos impureza?

Árboles de Decisión

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

- Impureza de Gini, para cada nodo:

1- (probabilidad de la clase 1)^2 – (probabilidad de la clase 2)^2



4 jugadores

$$1-\left(\frac{3}{3+1}\right)^2-\left(\frac{1}{3+1}\right)^2=0.375$$

6 jugadores

$$1-\left(\frac{2}{2+4}\right)^2-\left(\frac{4}{2+4}\right)^2=0.44$$

$$0.375\cdot\left(\frac{4}{10}\right)+0.44\cdot\left(\frac{6}{10}\right)=\mathbf{0.414}$$

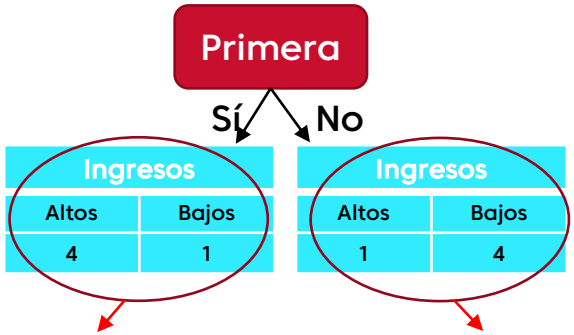
Impureza de Gini de la variable "**Titular**"

Árboles de Decisión

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

- Impureza de Gini, para cada nodo:

1- (probabilidad de la clase 1)^2 – (probabilidad de la clase 2)^2



5 jugadores

$$1-\left(\frac{4}{4+1}\right)^2-\left(\frac{1}{4+1}\right)^2=0.32$$

5 jugadores

$$1-\left(\frac{1}{1+4}\right)^2-\left(\frac{4}{1+4}\right)^2=0.32$$

$$0.32\cdot\left(\frac{5}{10}\right)+0.32\cdot\left(\frac{5}{10}\right)=\mathbf{0.32}$$

Impureza de Gini de la variable "**Primera**"

Árboles de Decisión

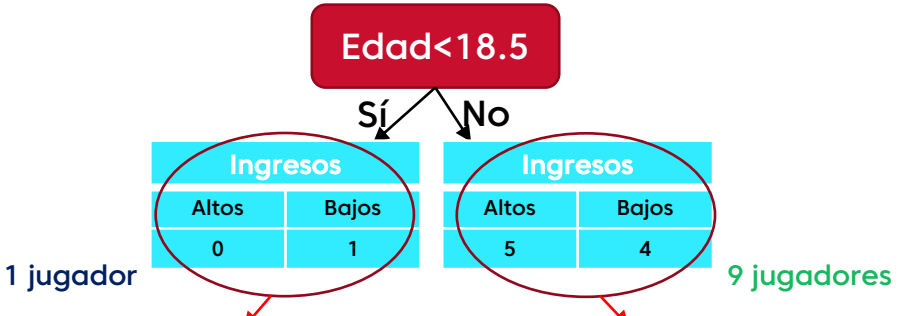
- En variables numéricas:

1. Ordenar de menor a mayor
2. Calcular la media para pares adyacentes
3. Calcular el índice Gini para cada media
4. Escoger el que tenga el menor Gini

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

Edad	INGRESOS
18	BAJOS
19	ALTOS
19	BAJOS
20	ALTOS
20	BAJOS
24	BAJOS
28	ALTOS
29	ALTOS
30	ALTOS
31	BAJOS

18.5
19
19.5
20
22
26
28.5
29.5
30.5



$$\text{Gini}(18.5) = 0\cdot\left(\frac{1}{10}\right)+0.49\cdot\left(\frac{9}{10}\right)=\mathbf{0.44}$$

Gini(18.5) = 0.44
Gini(19) = 0.44
Gini(19.5) = 0.48
Gini(20) = 0.48
Gini(22) = 0.48
Gini(26) = 0.42
Gini(28.5) = 0.48
Gini(29.5) = 0.5
Gini(30.4) = 0.44

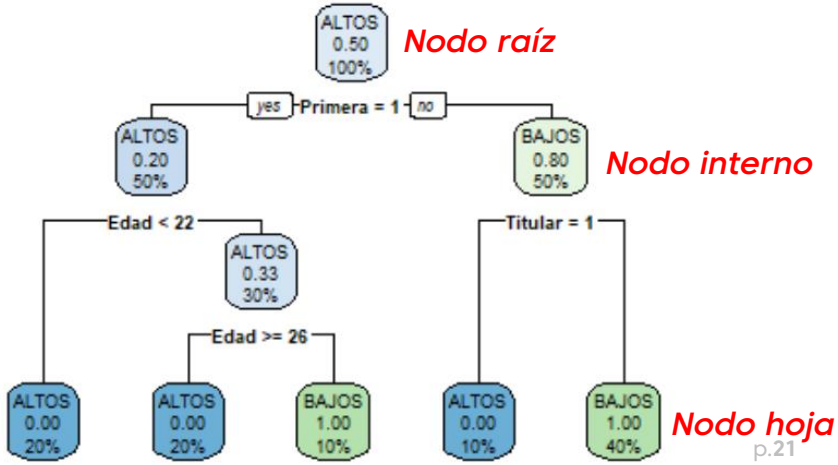
Árboles de Decisión

Jugador	Edad	Primera	Titular	INGRESOS
1	19	SI	SI	ALTOS
2	20	SI	SI	ALTOS
3	20	NO	NO	BAJOS
4	19	NO	NO	BAJOS
5	28	SI	NO	ALTOS
6	24	SI	SI	BAJOS
7	18	NO	NO	BAJOS
8	29	SI	NO	ALTOS
9	30	NO	SI	ALTOS
10	31	NO	NO	BAJOS

- Impureza Gini de la variable "**Edad**": 0.42
- Impureza Gini de la variable "**Primera**": **0.32**
- Impureza Gini de la variable "**Titular**": 0.41

⇒ La variable "Primera" tiene menos impureza, por lo que funciona mejor a la hora de separar la variable objetivo, utilizándose como nodo raíz

- Este proceso se repite en los nodos intermedios
- Un nodo se convierte en hoja cuando ninguna variable separa mejor el resultado de ese nodo



¡Gracias!