



Universidad de Alcalá

Escuela Politécnica Superior

Universidad de Alcalá

Fundamentos de Ciencia de Datos

Juan J. Cuadrado Gallego

Martes 10:00 – 12:00

Grado en Ingeniería Informática – Curso 2019/2020

Marcos Barranquero Fernández – 51129104N

INFORMACIÓN BÁSICA

- Prof. Juan J. Cuadrado Gallego – jjcg@uah.es
- No sube diapositivas.
- Revisar probabilidad y estadística.
- No hay nota mínima en las PECs ni PECLs.
- **Comienzo de laboratorio** el 24/09
- En el laboratorio se deben extraer datos de otros sitios y crear nuestras propias prácticas.
- **PEC1** – 29/10. 20%. T1, T2, T3. Ejercicios de tomar muestra y devolver asociación, etc. A mano.
- **PEC2** – 17/12. 20%. Temas restantes.
- **PECL1...6** – cada una un 10%.
- Temario: Datos, Asociación, Clasificación supervisada, clasificación no supervisada, datos anómalos, visualización.
- ~~Inscripción de laboratorios el martes 17/09 a las 8h en CC – UAH.~~

TEMA 1 – DEFINICIONES

CONCEPTOS BASE

INTRODUCCIÓN

La ciencia de los datos busca obtener conocimiento de los datos. Estudiamos lo que observamos. Muchas de sus ramas están basadas en la estadística. Debemos tener en cuenta las siguientes definiciones básicas:

- **Característica:** que otorga carácter o sirve para distinguir algo de sus semejantes. Pueden ser:
 - **Cualitativas:** expresamos con texto o nombre. (Ej. *Localidad*).
 - **Cuantitativas:** son reflejadas con números y admiten operaciones aritméticas. (Ej. *Número pi*, contraej. El DNI no es cuantitativo).
 - **Binarias:** aquellas que tienen dos valores. (Ej. *Tener carnet*).
- **Propiedad:** atributo o cualidad esencial de algo.
- **Cualidad:** elemento distintivo de la naturaleza de algo.
- **Atributo:** propiedad o característica de algo.
- **Variable:** cualidad medible de algo que puede variar o adquirir distintos valores.
- **Dato:** información dispuesta de manera adecuada para su tratamiento. Es el valor obtenido de una característica en una observación.

PROPIEDADES DEL DATO

Un dato puede ser:

- **Cualitativo:** que otorgan distinción para nombrar la característica.
 - Nominal: que da información para nombrar o describir la característica.
Operaciones admitidas: =, ≠.
Ej. verde.
 - Ordinal: proporciona suficiente información para ordenar las observaciones.
Operaciones admitidas: =, ≠, >, <.
Ej. tercero.
- **Cuantitativo:** describen características numéricas y aritméticas.
 - Discreto: son valores fijos sobre los Naturales.
Operaciones admitidas: =, ≠, >, <, *, %.
Ej. número de alumnos de una clase.
 - Continuo: son valores reales sobre los Reales.
Operaciones admitidas: =, ≠, >, <, *, %.
Ej. nota de un examen.

Estos datos cuantitativos también admiten **otra clasificación**:

- De intervalo: pertenece a una escala de intervalo en la cual no existe el 0, o que el 0 no indica ausencia. Permite saber la diferencia entre dos datos.
Operaciones admitidas: =, ≠.
Ej. Temperatura.
- Razón: pertenecen a una escala donde si existe el 0 o ausencia.
Operaciones admitidas: =, ≠, >, <.
Ej. número de veces que he ido a la universidad esta semana.
- Lógico o binario: que adquiere valor verdadero o falso.
Ej. tener carnet de conducir.

CIENCIA DE LOS DATOS: OBJETIVOS Y DEFINICIÓN

Anteriormente, para obtener conocimiento se seguía el método científico, que consiste en:

1. Definimos el problema.
2. Se formula una hipótesis.
3. Recoges datos experimentales.
4. Analizamos esos datos.
5. Sacamos conclusiones.

La ciencia de los datos amplía esto. Permite extraer conocimiento de manera automatizada, permitiendo realizar múltiples veces los pasos 3 y 4, es decir, recoger datos y analizarlos. Tampoco es necesario formular una hipótesis.

La ciencia de los datos **se apoya en los conocimientos de**:

- Bases de datos.
- Estadística.
- Inteligencia Artificial.
- Programación.

La CCDD tiene 3 **áreas de conocimiento**:

- **Data warehousing**: cómo trabajar, preparar y almacenar los datos.
- **Data mining**: cómo y qué técnicas empleamos para analizar los datos.
- **Visualización**: como visualizamos los resultados del análisis.

Respecto al **big data**, la ciencia de los datos proporciona conocimiento para realizar la actividad de **extracción de información**.

ANÁLISIS DE DATOS

CLASES DE EQUIVALENCIA

Una clase de equivalencia es un subconjunto del conjunto de datos. Permite agrupar datos en conjuntos más grandes. Estas clases de equivalencia tienen propiedades **representativas** de los datos que contiene. Las principales son:

- **Rango:** diferencia del valor menor al valor mayor.
- **Límites:** el número menor admitido y el mayor admitido.
- **Amplitud:** Diferencia de límites.
- **Marca:** media de los posibles elementos de la marca de clase.

Sean v_{inf} el **valor** superior y v_{sup} el **valor** superior de la clase. Sean l_{inf} el **límite** inferior y l_{sup} el **límite** superior de la clase de equivalencia.

Rango	Límite	Amplitud	Marca
$v_{sup} - v_{inf}$	<ul style="list-style-type: none">• Superior: l_{sup}• Inferior: l_{inf}	$l_{sup} - l_{inf}$	$(l_{sup} + l_{inf})/2$

MEDIDAS DE FRECUENCIA

Una vez tenemos los datos agrupados, disponemos de varias herramientas para analizarlos.

- **Frecuencia:** mide la cantidad de veces que aparece el mismo dato.
 - Absoluta (f_i): número de **veces que aparece un mismo dato** en un conjunto.
 - Relativa (f_{ri}): **frecuencia absoluta en relación al total** de elementos.
 - Acumulada (absoluta - (f_{ar}) o relativa - (f_{ari})): frecuencia de un dato concreto y **todos los anteriores o menores** a ese dato.

F. Absoluta (f_i)	F. Relativa (f_{ri})	F. Abs/Rel Acum.
N° Apariciones elemento	$\frac{N^{\circ} \text{ apariciones elemento}}{\text{Elementos totales}}$	$\sum_{i=0}^{\text{índice elem.}} f_i/f_{ri}(\text{elemento}_i)$

- **Media:** busca representar el conjunto mediante un valor resultado de la media de valores de ese conjunto. Se puede calcular de varias formas:
 - Aritmética (\bar{x}_a): suma de todos los valores dividido entre el número de valores.
 - Media geométrica (\bar{x}_g): suma de multiplicadores.
 - Media armónica ($(\bar{x}_a)^{-1}$): inversa que no sé pa que sirve.

Para datos agrupados (y **clases de equivalencia**) se cambia el dato por la marca de clase.

Aritmética ($\overline{x_a}$)	Geométrica ($\overline{x_g}$)	Armónica ($(\overline{x_a})^{-1}$):
$\frac{\sum_{i=0}^n x_i}{n}$	$\overline{x_g} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}}$	$\frac{\sum_{i=0}^n \frac{1}{x_i}}{n}$

EJERCICIO

Obtener rango de datos. Agrupar en 5 clases de equivalencia según decenas. Calcular límites, amplitud y marca de clase de cada una. Calcular también frecuencia absoluta y relativa. También media total.

16.5, 34.8, 20.7, 6.2, 4.4, 3.4, 24, 24, 32, 30, 33, 27, 15, 9.4, 2.1, 34, 24, 12, 4.4, 28, 31.4, 21.6, 3.1, 4.5, 5.1, 4.3, 2.25, 4.5, 20, 34, 12, 12, 12, 12, 5, 19, 30, 5.5, 38, 25, 3.7, 9, 30, 13, 30, 30, 26, 30, 30, 1, 26, 22, 10, 9.7, 11, 24.1, 33, 17.2, 27, 24, 27, 21, 28, 30, 4, 46, 29, 3.7, 2.7, 8.1, 19, 16.

Solución:

- C1: 1, 2.1, 2.7, 3.1, 3.2, 3.4, 3.7, 3.7, 4, 4, 4.4, 4.4, 4.5, 4.5, 5, 5.1, 5.5, 6.2, 8.1, 9, 9.4, 9.7.
- C2: 11, 12, 12, 12, 12, 12, 13, 15, 16, 16.5, 17.2, 19, 19.
- C3: 20.7, 21, 21.6, 22, 24, 24, 24, 24, 24.1, 25, 25, 26, 26, 27, 27, 27, 28, 28, 29.
- C4: 31.4, 32, 33, 33, 34, 34, 34.8, 38.
- C5: 46.

C.Eq.	Límite	Ampl.	Marca	Freq. Abs.	Freq. Rel.	F. Abs. Ac.	F. Rel. Ac.
C1	0-10	10	5	22	22/73	22	0.3
C2	10-20	10	15	14	14/73	36	0.49
C3	20-30	10	25	20	20/73	56	0.76
C4	30-40	10	35	16	16/73	72	0.98
C5	40-50	10	45	1	1/73	73	1

- Media usando todos los números: $(\frac{\sum x_i}{n}) = 18.53424657534247$.
- Media usando marcas de clase: $(\sum f_{ri} \cdot \text{marca de clase}) = 19.52054794520548$.
- Rango: $46 - 1 = 45$

ANÁLISIS DE LA MEDIA – COEF. DE VARIACIÓN DE PEARSON

El Coef. De variación indica lo representativa que es la media respecto a los datos, utilizando la desviación típica:

$$cv_{\bar{x}} = \frac{\sigma}{\bar{x}}$$

Conforme más cerca se encuentre el coef. De variación $cv_{\bar{x}}$ de 0, más representativa es la media.

MODA

La **moda** es el **dato que aparece más veces** en un conjunto.

MEDIDAS DE DISPERSIÓN

La dispersión indica lo mucho que se aleja el dato representativo de la realidad.

Distinguimos entre:

- Dispersión **absoluta**: resultado de desviación, varianza y rango.
- Dispersión **relativa**: relación entre la media y la dispersión absoluta. Coeficiente de variación de Poisson. (?).

Para calcularla, hacemos uso de diferentes herramientas:

Desviación típica o estándar (σ)	Desviación media (s)	Varianza (σ^2)
$\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$	$\frac{\sum_{i=1}^n x_i - \bar{x} }{\text{numero de elem.}}$	$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$

Teorema de Tchebychev: Para ver lo acertado que es el valor representativo respecto a la desviación, podemos utilizar la siguiente fórmula.

Pertenencia	Amplitud	Intervalo
$\begin{aligned} \text{pertenencia} &= 1 - \frac{1}{k^2}; k \\ &= \sqrt{\frac{1}{(1-p)}} \end{aligned}$	$\text{amplitud} = k \cdot \sigma$	$[\bar{x} - \text{amplitud}, \bar{x} + \text{amplitud}]$

El porcentaje de datos *pertenencia* se encontrará representado en el *intervalo*.

EJERCICIO

Dados los siguientes 12 números, calcular media, desviación típica o estándar, varianza, e intervalo para incluir el 75% de los datos.

13,15,16,20,20,22,27,29,30,33,34,42.

- Media aritmética (\bar{x}_a): 25'08
- Desviación típica o estándar (σ): $\sqrt{\frac{862'91}{12}} = 8'48$
- Para tener el 75% de los datos:
$$0'75 = 1 - \frac{1}{k^2}; k = 2; \text{La amplitud del intervalo debe de ser} = k \cdot \sigma$$
$$= 2 \cdot 8.48 = 16.96$$
- Varianza (σ^2): $8'48^2 = 71.91$
- Mediana = $\frac{22+27}{2} = 24.5$

MEDIDAS DE ORDENACIÓN

Dado un conjunto de datos ordenados, permite dividirlo en subconjuntos y extraer datos sobre estos.

- **Mediana:** indica el dato que se encuentra en el medio del conjunto de datos. Sea n el número de elementos en el conjunto:
 - Si n es par: $\bar{x} = \frac{x_{n/2} + x_{n/2+1}}{2}$
 - Si n es impar: $\bar{x} = x_{n/2}$
- **Cuantiles:** un cuantil indica el valor en una posición aproximada del conjunto de datos. Si n_t indica el número de elementos en el conjunto, n_c indica el índice del valor en la posición del cuantil Q_i .

Cuartil	Decil	Percentil	Centil
$n_c = \frac{1}{4}n_t$	$n_c = \frac{1}{10}n_t$	$n_c = \frac{1}{100}n_t$	$n_c = \frac{1}{n}n_t$

- Si n_c es un número natural:
 - Si n_t es un número impar: $Q_i = x_{n_c}$
 - Si n_t es un número par: $Q_i = \frac{x_{n_c\downarrow} + x_{n_c\uparrow}}{2}$
- Si n_c es un número real: $Q_i = \frac{x_{n_c\downarrow} + x_{n_c\uparrow}}{2}$

EJEMPLO

Calcular mediana y primer cuartil de los datos:

13,15,16,20,20,22,27,29,30,33,34,42.

- **Mediana:** Como $n = 12$; $\bar{x} = \frac{x_{n/2} + x_{n/2+1}}{2} = \frac{22+27}{2} = 24'5$
- **Primer cuartil:** : Como $n = 12$; El primer cuartil corresponde a $12 \cdot \frac{1}{4} = 3$. Como n es par, $\frac{x_3 + x_4}{2} = \frac{16+20}{2} = 18$.

PROBABILIDAD

La probabilidad estudia la posibilidad de que suceda o no un suceso. Comprende el siguiente vocabulario:

- **Experimento:** operación para explorar fenómenos o principios. Hay experimentos aleatorios cuyo resultado no está determinado.
- **Suceso:** subconjunto de resultados posibles en un experimento aleatorio.
- **Suceso elemental:** suceso simple.
- **Espacio muestral:** conjunto de todos los sucesos elementales en un experimento aleatorio.
- **Partes de E:** conjunto de posibles subconjuntos del espacio muestral.
- **Suceso complementario:** suceso opuesto a otro suceso, lo que ocurre cuando no ocurre un determinado suceso.
- **Suceso seguro:** suceso que siempre se cumple.
- **Suceso imposible:** suceso que nunca se cumple.

EJEMPLO

Determinar cada elemento del vocabulario para el caso de lanzar un dado:

- **Experimento:** lanzar el dado.
- **Suceso:** el resultado de tirarlo. Par, impar, que sea 3, distinto de 3, etc.
- **Suceso elemental:** que el resultado sea 1,2,3,4,5 o 6.
- **Espacio muestral:** $\{1,2,3,4,5,6\}$
- **Partes de E:**
 - Tirarlo 1 vez y que salga $\{1\}, \{3\}, \{5\}$, etc.
 - tirarlo 2 veces y que salga $\{2,3\}, \{3,4\}, \{1,5\}$, etc.
- **Suceso complementario:** Si $A = \{2\}$, $\text{complementario}(A) = \{1,3,4,5,6\}$
- **Suceso seguro:** que el resultado esté entre 1 y 6.
- **Suceso imposible:** que salga 7.

REGLAS DE PROBABILIDAD

- $0 \leq P(A) \leq 1$
- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Si $A \subset B \rightarrow P(A) \leq P(B)$
- $P(E) = 1$
- $P(\emptyset) = 0$

REGLAS DE CONJUNTOS

- $B \cup C = \{\text{elementos}(B) + \text{elementos}(C)\}$
- $B \cap C = \{\text{elementos compartidos}(B, C)\}$
- $B - C = \{\text{elementos}(B) - \text{elementos compartidos}(B, C)\}$
- $B \Delta C = \{\text{elementos}(B) + \text{elementos}(C) - \text{elementos compartidos}(B, C)\}$

EJEMPLO

Sea:

- $B = \{2, 4, 6\}$
- $C = \{3, 4, 5, 6\}$

Calcular todas las reglas de conjunto:

- $B \cup C = \{\text{elementos}(B) + \text{elementos}(C)\} = \{2, 3, 4, 5, 6\}$
- $B \cap C = \{\text{elementos compartidos}(B, C)\} = \{4, 6\}$
- $B - C = \{\text{elementos}(B) - \text{elementos compartidos}(B, C)\} = \{2\}$
- $B \Delta C = \{\text{elementos}(B) + \text{elementos}(C) - \text{elementos compartidos}(B, C)\} = \{2, 3, 5\}$

T2 – ASOCIACIÓN

INTRODUCCIÓN ASOCIACIÓN

Los estudios de asociación buscan encontrar patrones de aparición conjunta de sucesos.

El algoritmo Apriori busca encontrar sucesos que permitan sacar conclusiones y tomar decisiones, en base a un espacio muestral, utilizando las herramientas del soporte y la confianza.

SOPORTE

El soporte es la probabilidad de aparición de unos sucesos determinados en el espacio de muestras total.

Soporte
$S(C) = \frac{\text{numero de apariciones}(C)}{\text{tamaño del espacio muestral}}$

CONFIANZA

Probabilidad de que, dados dos conjuntos de sucesos, si se cumple uno se cumpla el otro.

Confianza
$C(C_1 \rightarrow C_2) = \frac{\text{numero de apariciones}(C_1 \cup C_2)}{\text{numero de apariciones}(C_1)}$

ALGORITMO APRIORI

El algoritmo Apriori busca encontrar asociaciones que superen un umbral de confianza y soporte, a partir de un espacio muestral.

Tiene los siguientes pasos:

1. Se calcula soporte de los sucesos elementales de ese espacio muestral, y se descartan sucesos elementales que no superen el umbral de soporte.
2. Se generan sucesos candidatos empleando Apriori-Gen. Partiendo de los sucesos elementales, para generar casos de tamaño $k+1$, se toman sucesos de tamaño k que compartan $k-1$ elementos.
3. Se calcula soporte para los sucesos candidatos generados, y se descartan aquellos que no superen el umbral de soporte.
4. Se generan todas las asociaciones (permutaciones) posibles partiendo de los sucesos candidatos. Para cada asociación, se calcula la confianza y se descartan las asociaciones que no superen el umbral de confianza.
5. Nos quedamos con los casos no descartados.

OPTIMIZACIONES

El algoritmo se puede optimizar en los siguientes pasos:

PASO 1

Ya está optimizado, haciendo uso de la propiedad:

$$\text{soporte}(\text{caso}) \geq \text{soporte}(\text{subcaso que compone el caso})$$

EJEMPLO

$$\text{soporte}(PAL) > \text{soporte}(PA), \text{soporte}(PL), \text{soporte}(AL)$$

PASO 3

Utilizar un hash tree con función modulo y leer las ramas.

PASO 5

Utilizamos la propiedad:

$$\begin{aligned} \text{Si } C(A \rightarrow (B - A)) \leq \text{umbral de confianza, entonces} \\ C(A' \rightarrow (B - A')) \leq \text{umbral de confianza} \\ \text{donde } A' \subseteq A \end{aligned}$$

EJEMPLO

Si tenemos un umbral de confianza de 0'8, y $A = \{P, A\}$ y $B = \{P, A, L\}$

- $C(P, A \rightarrow L) = 0'75 < 0'8$.
- Entonces
 - $C(P \rightarrow A, L) = 0'6 < 0'8$
 - $C(A \rightarrow P, L) = 0'75 < 0'8$

EJEMPLO COMPLETO DE APRIORI

Sea el siguiente espacio muestral:

$$E = \{PALN, PACL, PAL, PCL, PA, L\}$$

Se desea aplicar el algoritmo A priori teniendo en cuenta:

- Umbral de soporte $\geq 50\%$
- Umbral de confianza $\geq 80\%$

PASO 1

Calculamos soporte de los sucesos elementales:

- $S(P) = \frac{\text{numero de apariciones}(P)}{\text{tamaño del espacio muestral}} = \frac{5}{6} = 83\%$
- $S(A) = \frac{\text{numero de apariciones}(A)}{\text{tamaño del espacio muestral}} = \frac{4}{6} = 66\%$
- $S(L) = \frac{\text{numero de apariciones}(L)}{\text{tamaño del espacio muestral}} = \frac{5}{6} = 83\%$
- $S(C) = \frac{\text{numero de apariciones}(C)}{\text{tamaño del espacio muestral}} = \frac{2}{6} = 33\%$
- $S(N) = \frac{\text{numero de apariciones}(N)}{\text{tamaño del espacio muestral}} = \frac{1}{6} = 16\%$

Observamos que para los casos elementales de C y N, el soporte se encuentra por debajo del umbral. Por tanto, los descartamos para el paso siguiente.

PASO 2

Identifico sucesos candidatos al estudio aplicando Apriori-Gen.

Para generar casos de tamaño $k+1$, tomo sucesos de tamaño k que comparten $k-1$ elementos.

Para $k = 1$					
C1	C2	Comparten	De tamaño	Genera	De tamaño
P	A	\emptyset	0	PA	2
A	L	\emptyset	0	AL	2
P	L	\emptyset	0	PL	2
Para $k = 2$					
C1	C2	Comparten	De tamaño	Genera	De tamaño
PA	AL	A	1	PAL	3

PASO 3

Calculamos soporte de sucesos candidatos generados:

- $S(PA) = \frac{\text{numero de apariciones}(P)}{\text{tamaño del espacio muestral}} = \frac{4}{6} = 66\%$
- $S(AL) = \frac{\text{numero de apariciones}(A)}{\text{tamaño del espacio muestral}} = \frac{3}{6} = 50\%$
- $S(PL) = \frac{\text{numero de apariciones}(L)}{\text{tamaño del espacio muestral}} = \frac{4}{6} = 50\%$
- $S(PAL) = \frac{\text{numero de apariciones}(C)}{\text{tamaño del espacio muestral}} = \frac{3}{6} = 66\%$

Como todos los casos superan el umbral de soporte establecido, no debemos descartar ninguno.

PASO 4

Estudiamos la confianza para todos los sucesos generados.

Para cada suceso, dependiendo de la dimensión, pueden generarse varios casos.

- *numero de casos para dim = 2* $\rightarrow 2^2 - 2$ (por cjto. vacíos) = 2
- *numero de casos para dim = 3* $\rightarrow 2^3 - 2$ (por cjto. vacíos) = 6

Como tenemos 3 casos de dimensión 2 y 1 caso de dimensión 3, estudiaremos:

$$\left[3 \text{ sucesos dimension 2} * 2 \frac{\text{casos}}{\text{suceso}} \right] + \left[1 \text{ suceso dimension 3} * 6 \frac{\text{casos}}{\text{suceso}} \right] = 12 \text{ casos totales}$$

Los casos son los siguientes:

DIMENSIÓN 2 – (PA, AL, PL)

- $C(P \rightarrow A) = \frac{\text{numero de apariciones}(PA)}{\text{numero de apariciones}(P)} = \frac{4}{5} = 80\%$
- $C(A \rightarrow P) = \frac{\text{numero de apariciones}(PA)}{\text{numero de apariciones}(A)} = \frac{4}{4} = 100\%$
- $C(A \rightarrow L) = \frac{\text{numero de apariciones}(AL)}{\text{numero de apariciones}(A)} = \frac{3}{4} = 75\%$
- $C(L \rightarrow A) = \frac{\text{numero de apariciones}(AL)}{\text{numero de apariciones}(L)} = \frac{3}{5} = 60\%$
- $C(P \rightarrow L) = \frac{\text{numero de apariciones}(PL)}{\text{numero de apariciones}(P)} = \frac{4}{5} = 80\%$
- $C(L \rightarrow P) = \frac{\text{numero de apariciones}(PL)}{\text{numero de apariciones}(L)} = \frac{4}{5} = 80\%$

Descartamos $C(A \rightarrow L)$ y $C(L \rightarrow A)$ debido a que están por debajo del umbral de confianza.

DIMENSIÓN 3 – (PAL)

- $C(A \rightarrow PL) = \frac{\text{numero de apariciones}(PAL)}{\text{numero de apariciones}(A)} = \frac{3}{4} = 75\%$
- $C(P \rightarrow AL) = \frac{\text{numero de apariciones}(PAL)}{\text{numero de apariciones}(P)} = \frac{3}{5} = 60\%$
- $C(L \rightarrow PA) = \frac{\text{numero de apariciones}(PAL)}{\text{numero de apariciones}(L)} = \frac{3}{5} = 60\%$
- $C(AL \rightarrow P) = \frac{\text{numero de apariciones}(PAL)}{\text{numero de apariciones}(LA)} = \frac{3}{3} = 100\%$
- $C(PA \rightarrow L) = \frac{\text{numero de apariciones}(PAL)}{\text{numero de apariciones}(PA)} = \frac{3}{4} = 75\%$
- $C(PL \rightarrow A) = \frac{\text{numero de apariciones}(PAL)}{\text{numero de apariciones}(PL)} = \frac{3}{4} = 75\%$

Solamente tomamos $C(AL \rightarrow P)$ debido a que se encuentra por encima del umbral de confianza.

PASO 5

Nos quedamos con los casos no descartados.

Caso	Soporte	Confianza
$C(A \rightarrow P)$	66%	100%
$C(P \rightarrow A)$	66%	80%
$C(L \rightarrow P)$	50%	80%
$C(P \rightarrow L)$	50%	80%
$C(AL \rightarrow P)$	66%	100%

Vemos que es congruente con lo realizado por la librería aRules de R:

```
> inspect(asociaciones)
      lhs      rhs      support  confidence lift count
[1] {}      => {Leche} 0.8333333 0.8333333  1.00  5
[2] {}      => {Pan}   0.8333333 0.8333333  1.00  5
[3] {Agua}   => {Pan}   0.6666667 1.0000000  1.20  4
[4] {Pan}    => {Agua}   0.6666667 0.8000000  1.20  4
[5] {Leche}  => {Pan}   0.6666667 0.8000000  0.96  4
[6] {Pan}    => {Leche} 0.6666667 0.8000000  0.96  4
[7] {Agua,Leche} => {Pan} 0.5000000 1.0000000  1.20  3
> |
```

T3 – CLASIFICACIÓN SUPERVISADA

INTRODUCCIÓN CLASIFICACIÓN SUPERVISADA

Los estudios de clasificación buscan definir un modelo de clasificación que defina el valor de un suceso desconocido.

La clasificación supervisada requiere:

- Conocer los posibles valores de ese suceso.
- Disponer de una muestra con valores conocidos.

TÉCNICAS

- **Árboles de decisión (para datos cualitativos)**
- Reglas de decisión
- Random forest
- Máquinas vectoriales de soporte (JVM)
- Naive Bayes (probab. Bayesiana)
- Redes neuronales
- **Regresión (para datos cuantitativos)**

ÁRBOLES DE DECISIÓN

Para realizar un árbol de decisión, debemos de conocer:

- Espacio muestral (E).
- Elementos usados para la clasificación y su dominio.
- Elemento clasificador (el que queremos clasificar) y su dominio.

Buscamos una **función clasificadora $f(x)$** que, dados los elementos clasificadores como argumentos, devuelva el valor del elemento a clasificar.

Esta $f(x)$ será mejor cuantas menos variables necesite.

Para diseñar $f(x)$, estudiamos el espacio muestral utilizando el algoritmo de Hunt.

ALGORITMO DE HUNT

1. Obtener suceso elemental en nodo inicial (no puede ser el clasificador).
2. Clasificar sucesos en base a nodo elegido.
 - a. Si podemos clasificar todos, hemos finalizado.
 - b. Si no, obtener suceso elemental en nodo intermedio y ejecutar paso 2.

GANANCIA DE INFORMACIÓN

Para optimizar el árbol de decisión, podemos estudiar la ganancia de información en cada iteración para decidir sobre qué elemento del dominio elegimos estudiar como suceso elemental para la siguiente iteración.

El mejor elemento será aquel cuya **variación de impureza** sea mayor.

Cálculo de la impureza
$\Delta Impureza = I_{padre} \cdot freq_{padre} - I_{hijos} \cdot freq_{hijos}$
$I_{hijos} = \sum_{j=1}^k \frac{N_{sucesos}(Nodo_j)}{N_{total\ sucesos}} I(Nodo_j)$

La impureza se puede medir en:

Entropía	Error	Gini
$-\sum(freq_i(nodo_i) \cdot \log_2(freq_i(nodo_i)))$	$1 - \max(freq_i(nodo_i))$	$1 - \sum(freq_i(nodo_i))^2$

EJEMPLO

ENUNCIADO

Se desea realizar un modelo de clasificación que defina el valor de las notas finales del laboratorio, dando como entrada las notas obtenidas a lo largo del curso, es decir, Teoría, Práctica y Laboratorio.

Cada nota del curso puede tener los valores de A, B, C y D.

La nota final consistirá en Aprobado o Suspenso.

Se tiene el siguiente espacio muestral:

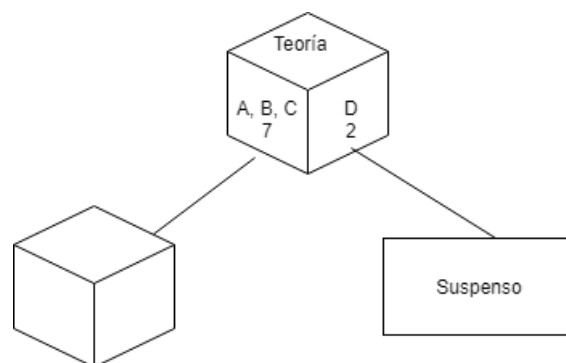
Suceso	Teoría	Laboratorio	Práctica	Nota final
1	A	A	B	Aprobado
2	A	B	D	Suspenso
3	D	C	C	Suspenso
4	D	D	A	Suspenso
5	B	C	B	Suspenso
6	C	B	B	Aprobado
7	B	B	A	Aprobado
8	C	D	C	Suspenso
9	B	A	C	Suspenso

ELEMENTOS

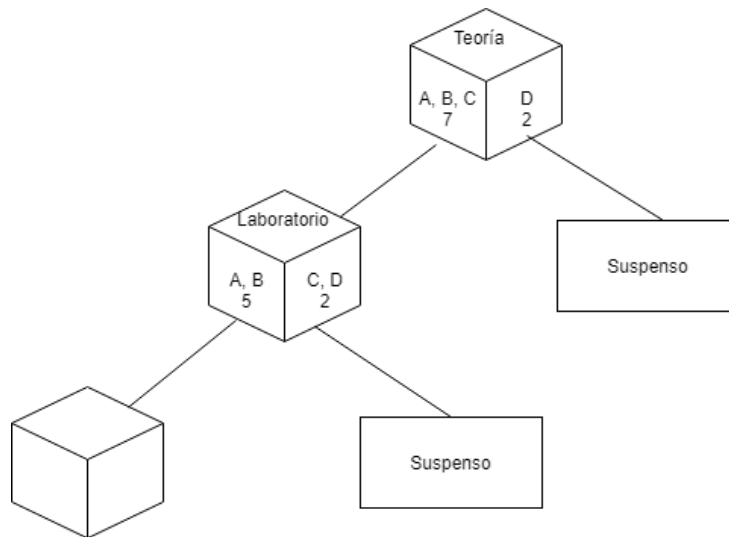
Espacio muestral	Teoría, Laboratorio, Práctica – {A,B,C,D}
Elem. Clasific.	Nota final – {Aprobado, Suspenso}
Función clasificadora	$f(\text{teoría, laboratorio, práctica}) \rightarrow \text{nota final}$

RESOLUCIÓN SIN GANANCIA DE INFORMACIÓN

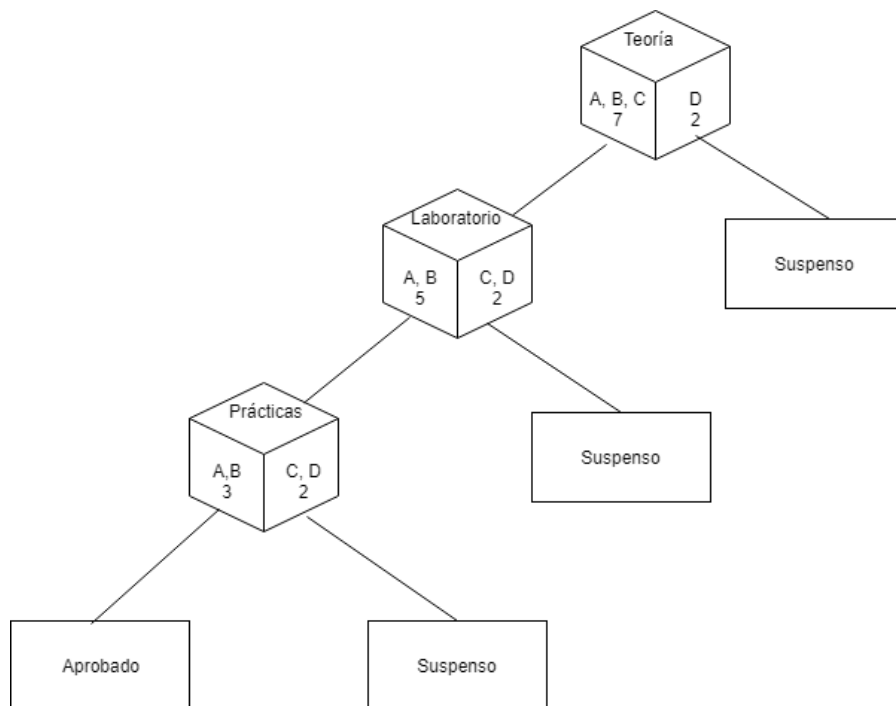
Partimos de la nota de teoría, que es la primera variable. Observando los sucesos del espacio muestral, observamos que los que tienen una nota de D en la teoría siempre tienen como nota final un suspenso. Lo representamos:



Tras esto, observamos la nota del laboratorio. Aquellos con una calificación de C o de D en el laboratorio tienen su nota final suspensa, mientras que los que tienen A o B pueden estar aprobados o suspensos. Lo representamos:



Finalmente, nos queda la nota de las prácticas. Observamos que aquellos con un A o B en las prácticas se encuentran aprobados, y el resto suspenso. Lo representamos:

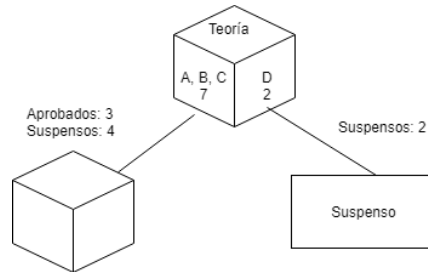


Y, al haber llegado a clasificar todos los sucesos, hemos llegado al final de nuestro árbol de decisión.

EJEMPLOS CON CÁLCULO DE GANANCIA DE INFORMACIÓN

Realizaremos el paso 1 utilizando la nota de teoría y calculando la ganancia de información de las 3 formas posibles.

Teniendo el paso 1 realizado, añadimos número de aprobados y suspensos por nodo:



IMPUREZA CON ENTROPÍA

$$E(\text{padre}) = -\sum f_i(p) \cdot \log_2(f_i(\text{padre})) = -\left(\frac{3}{9}\right) \cdot \log_2\left(\frac{3}{9}\right) - \left(\frac{6}{9}\right) \cdot \log_2\left(\frac{6}{9}\right) = 0'9$$

$$E(\text{hijo}_{izq.}) = -\sum f_i(\text{hijo}_{izq.}) \cdot \log_2(f_i(\text{hijo}_{izq.})) = -\left(\frac{3}{7}\right) \cdot \log_2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right) \cdot \log_2\left(\frac{4}{7}\right) = 0'98$$

$$E(\text{hijo}_{der.}) = -\sum f_i(\text{hijo}_{der.}) \cdot \log_2(f_i(\text{hijo}_{der.})) = -\left(\frac{0}{2}\right) \cdot \log_2\left(\frac{0}{2}\right) - \left(\frac{2}{2}\right) \cdot \log_2\left(\frac{2}{2}\right) = 0$$

$$\Delta \text{Impureza} = I_{\text{padre}} - I_{\text{hijos}} = 0.9 - \left(\left(\frac{7}{9}\right) \cdot 0.98 + \left(\frac{2}{9}\right) \cdot 0\right) = 0'9 - 0'76 = 0.14$$

IMPUREZA CON ERROR

$$E(\text{padre}) = 1 - \max(\text{freq}_i(\text{padre})) = 1 - \max\left(\frac{3}{9}, \frac{6}{9}\right) = 0'33$$

$$E(\text{hijo}_{izq.}) = 1 - \max(\text{freq}_i(\text{hijo}_{izq.})) = 1 - \max\left(\frac{3}{7}, \frac{4}{7}\right) = 0'42$$

$$E(\text{hijo}_{der.}) = 1 - \max(\text{freq}_i(\text{hijo}_{der.})) = 1 - \max\left(\frac{2}{2}, \frac{0}{2}\right) = 0$$

$$\Delta \text{Impureza} = I_{\text{padre}} - I_{\text{hijos}} = 0'33 - \left(\left(\frac{7}{9}\right) \cdot 0.42 + \left(\frac{2}{9}\right) \cdot 0\right) = 0'33 - 0'33 = 0$$

IMPUREZA CON GINI

$$E(\text{padre}) = 1 - \sum (\text{freq}_i(\text{padre}))^2 = 1 - \left[\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right] = 0'45$$

$$E(\text{hijo}_{izq.}) = 1 - \sum (\text{freq}_i(\text{hijo}_{izq.}))^2 = 1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0'48$$

$$E(\text{hijo}_{der.}) = 1 - \sum (\text{freq}_i(\text{hijo}_{der.}))^2 = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

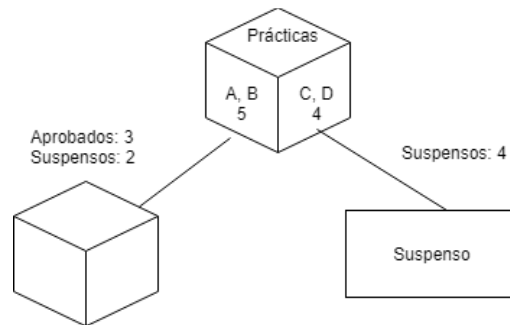
$$\Delta \text{Impureza} = I_{\text{padre}} - I_{\text{hijos}} = 0'45 - \left(\left(\frac{7}{9}\right) \cdot 0'48 + \left(\frac{2}{9}\right) \cdot 0\right) = 0'45 - 0'038 = 0'069$$

RESOLUCIÓN CON GINI

Calculo como antes Gini para Teoría, Laboratorio y Prácticas. Debo elegir empezar por la variable cuyo Gini calculado sea mayor.

	Laboratorio	Teoría	Prácticas
Gini – padre	0.45	0.45	0.45
Gini – hijos	0.27	0.38	0.27
Impureza	0.18	0.069	0.18

Resulta que el menor es el de teoría y prácticas, por tanto, empiezo por Prácticas.

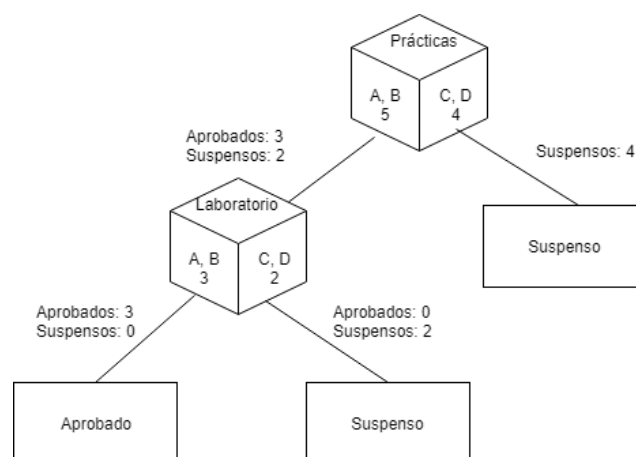


Ahora debo calcular de nuevo el Gini para las variables restantes desde este nodo, y elegir el que tenga mayor impureza de nuevo:

	Laboratorio	Teoría
Gini – padre	0.4	0.4
Gini – hijos	0.0	0.2
Impureza	0.4	0.2

Elegimos laboratorio ya que contiene el grado mayor de impureza.



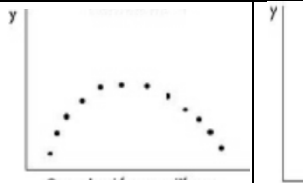
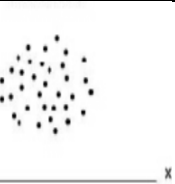
Observamos que ya hemos clasificado todos los sucesos, y en menos pasos que sin tener en cuenta la ganancia de información. Vemos que es más eficiente.



REGRESIÓN – ANÁLISIS PARA VARIABLES CUANTITATIVAS

Busca estudiar la relación entre 2 o más variables.

Se tienen 4 casos:

Rel. Lineal positiva	Rel. Lineal negativa	Rel. No lineal	No relación
			
Modelable.	Modelable.	Difícilmente modelable.	No modelable.

ANÁLISIS DE 2 CARACTERÍSTICAS

Para analizar un par de datos de **forma conjunta** disponemos de 2 características:

- Frecuencia de aparición absoluta de los datos
- Frecuencia de aparición relativa de los datos

Para estudiar estas características, podemos apoyarnos en una **tabla de doble entrada** para variables cuantitativas o **tabla de contingencias** para variables cualitativas.

Por otro lado, tenemos la **distribución marginal** de 1 variable, que es la frecuencia absoluta de los datos de la variable.

EJEMPLO

Realizar tabla de doble entrada para los siguientes datos:

$E = (2,5), (6,5), (6,5), (3,3)$.

Solución:

	1	2	3	4	5	6	Suma
1	-	-	-	-	-	-	0
2	-	-	-	-	1	-	1
3	-	-	1	-	-	-	1
4	-	-	-	-	-	-	0
5	-	-	-	-	-	-	0
6	-	-	-	-	2	-	2
Suma	0	0	1	0	3	0	4

MEDIDAS DE DEPENDENCIA

Las medidas más utilizadas son:

- **Covarianza:** mide lo relacionado que están dos variables. **No está normalizada.**
- **Correlación:** mide la relación entre las dos variables. **Si está normalizada.**

Covarianza	Correlación [0,1]
$S_{xy} = \frac{(\sum_{i=0}^n x_i \cdot y_i)}{n} - (\bar{x} \cdot \bar{y})$	$r_{xy} = \frac{S_{xy}}{\sigma_x \cdot \sigma_y}$
<ul style="list-style-type: none"> • $S_{xy} > 0 \rightarrow$ Dependencia directa positiva. • $S_{xy} = 0 \rightarrow$ No existe relación lineal. • $S_{xy} < 0 \rightarrow$ Dependencia inversa negativa. 	<ul style="list-style-type: none"> • $[0,1] \rightarrow$ relación ascendente. • $0 \rightarrow$ no hay correlación. • $[-1,0] \rightarrow$ relación descendente. <p>Se consideran fuertemente correlacionadas a partir de 0'8.</p>

EJEMPLO COVARIANZA

Calcular covarianza para los datos:

$$E = (2,5), (6,5), (6,5), (3,3).$$

$$\begin{aligned}
 S_{xy} &= \frac{(\sum_{i=0}^n x_i \cdot y_i)}{n} - (\bar{x} \cdot \bar{y}) \\
 &= \frac{(2 \cdot 5) + (6 \cdot 5) + (6 \cdot 5) + (3 \cdot 3)}{4} - \left(\frac{(2 + 6 + 6 + 3)}{4} \right) \cdot \left(\frac{(5 + 5 + 5 + 3)}{4} \right) \\
 &= \frac{10 + 30 + 30 + 9}{4} - 4'25 \cdot 4'5 = 19'75 - 19'125 = 0'625
 \end{aligned}$$

EJEMPLO CORRELACIÓN

Calcular correlación para los datos:

$$E = (2'4, 5'4), (6'1, 5'2), (6'4, 5'5), (3'4, 3'9).$$

$$S_{xy} = \frac{(\sum_{i=0}^n x_i \cdot y_i)}{n} - (\bar{x} \cdot \bar{y}) = \frac{12'96 + 31'72 + 35'2 + 13'26}{4} - 4'575 \cdot 5 = 0'41$$

$$\sigma_x = \sqrt{\frac{(\sum_{i=0}^n (x_i - \bar{x})^2)}{n}} = \sqrt{\frac{(2'4 - 4'575)^2 + (6'1 - 4'575)^2 + (6'4 - 4'575)^2 + (3'4 - 4'575)^2}{4}} = 1'71$$

$$\sigma_y = \sqrt{\frac{(\sum_{i=0}^n (y_i - \bar{y})^2)}{n}} = \sqrt{\frac{(5'4 - 5)^2 + (5'2 - 5)^2 + (5'5 - 5)^2 + (3'9 - 5)^2}{4}} = 0'64$$

$$r_{xy} = \frac{S_{xy}}{\sigma_x \cdot \sigma_y} = \frac{0'41}{1'71 \cdot 0'64} = 0'374$$

FUNCIÓN DE REGRESIÓN

La función de regresión permite modelar la gráfica que muestra la relación entre los datos.

Se emplea el **ajuste por mínimos cuadrados**:

$y = a + bx$	<ul style="list-style-type: none">• $a = \bar{y} - b \cdot \bar{x}$• $b = \frac{S_{xy}}{(\sigma_x)^2}$
--------------	-----------------------------------------------------------------------------------------------------------------------------------------------------

EJEMPLO

Calcular recta de regresión para el conjunto de datos:

$$E = (2'4,5'4), (6'1,5'2), (6'4,5'5), (3'4,3'9).$$

- $b = \frac{S_{xy}}{(\sigma_x)^2} = \frac{0'41}{(1'71)^2} = 0'14$
- $a = \bar{y} - b \cdot \bar{x} = 5 - 0'14 \cdot 4'575 = 4'36$
- $y = a + bx = 4'36 + 0'14x$

ANÁLISIS DE LA RECTA DE REGRESIÓN

Para saber si la recta de regresión es adecuada, se tienen 3 herramientas de análisis:

ANOVA – SSR

$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})$	<ul style="list-style-type: none">• \hat{y}_i = valor calculado con la recta de regresión, tomando su x_i asociado.• \bar{y} = media de y.
--------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

DISPERSIÓN DE VALOR OBSERVADO – SSY

$SSy = \sum_{i=1}^n (y_i - \bar{y})$	<ul style="list-style-type: none">• y_i = valor original del dato.• \bar{y} = media de y.
--------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------

CORRELACIÓN CUADRÁTICA – R^2

$r^2 = \frac{SSR}{SSy}$	<ul style="list-style-type: none">• $0 \leq r \leq 1$; Conforme más se aproxime a 1, más correcta es la recta de regresión.
-------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------

EJEMPLO

Analizar la recta de regresión $y = 4'36 + 0'14x$ para los siguientes datos:

$$E = (2'4, 5'4), (6'1, 5'2), (6'4, 5'5), (3'4, 3'9).$$

Calculamos valores de \bar{y} con la recta de regresión:

$\bar{y}(2'4) = 4'696$	$\bar{y}(6'1) = 5'214$	$\bar{y}(6'4) = 5'246$	$\bar{y}(3'4) = 4'836$
------------------------	------------------------	------------------------	------------------------

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y}) = (4'696 - 5) + (5'214 - 5) + (5'246 - 5) + (4'836 - 5) = 0'23$$

$$SSy = \sum_{i=1}^n (y_i - \bar{y}) = (5'4 - 5) + (5'2 - 5) + (5'5 - 5) + (3'9 - 5) = 1'66$$

$$r^2 = \frac{SSR}{SSy} = \frac{0'23}{1'66} = 0'14; r = \sqrt{0'14} = 0'37$$

Como r se aproxima más a 0 que a 1, podemos concluir que la recta de regresión es una mala aproximación a la recta real.

ANÁLISIS DEL ERROR

El **error estándar** o **desviación típica residual** de la recta de regresión nos muestra lo desviada que se encuentra la recta de regresión respecto a la recta real.

$Sr = \sqrt{\frac{\sum_{i=1}^n ((y_i - \hat{y}_i)^2)}{n}}$	<ul style="list-style-type: none">• \hat{y}_i = valor calculado con la recta de regresión, tomando su x_i asociado.• y_i = valor original del dato.
------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Cuanto más cercano a 0 se encuentre, mejor.

EJEMPLO

Analizar error estándar para la recta de regresión $y = 4'36 + 0'14x$ para los sig. datos:

$$E = (2'4, 5'4), (6'1, 5'2), (6'4, 5'5), (3'4, 3'9).$$

Calculamos valores de \bar{y} con la recta de regresión:

$\bar{y}(2'4) = 4'696$	$\bar{y}(6'1) = 5'214$	$\bar{y}(6'4) = 5'246$	$\bar{y}(3'4) = 4'836$
------------------------	------------------------	------------------------	------------------------

$$Sr = \sqrt{\frac{\sum_{i=1}^n ((y_i - \hat{y}_i)^2)}{n}} = \sqrt{\frac{(5'4 - 4'696)^2 + (5'2 - 5'214)^2 + (5'5 - 5'246)^2 + (3'9 - 4'836)^2}{4}} = 0'598$$

Vemos que el error está alejado de 0, por tanto, es un gran error cometido. La recta de regresión se aleja bastante de la recta real.