Ejercicio 1.- En función de los resultados de la tabla adjunta de un clasificador con respecto al conjunto de test. Calcule la matriz de confusión y la puntuación de Briers ¿Qué sentido tiene esta puntuación?

	X1	...	Xk	C	$P(C_M=0\|x)$	$P(C_M=1\|x)$	Decisión
$(\mathbf{x_1}, c_1)$				1	0,15	0,85	E en 1
$(\mathbf{x_2}, c_2)$				0	0,20	0,80	F en 0
$(\mathbf{x_3}, c_3)$				0	0,60	0,40	E en 0
$(\mathbf{x_4}, c_4)$				1	0,55	0,45	F en 1
$(\mathbf{x_5}, c_5)$				0	0,62	0,38	E en 0
$(\mathbf{x_6}, c_6)$				0	0,12	0,88	F en 0
$(\mathbf{x_7}, c_7)$				1	0,83	0,17	F en 1
$(\mathbf{x_8}, c_8)$				1	0,30	0,70	E en 1
$(\mathbf{x_9}, c_9)$				1	0,70	0,30	F en 1
$(\mathbf{x_{10}}, c_{10})$				0	0,35	0,65	F en 0

La matriz de confusión es

C. predicha

C. real $\begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}$ El CCR=4/10, las sensibilidades son para la clase 1, 2/5 y para la clase 0, 2/5. La mínima sensibilidad es por tanto 2/5

La puntuación de Briers es

$$B = \frac{1}{N}\sum_{i=1}^{N}\sum_{C=0}^{1}\left[p(C_M = c \mid x_i) - \delta(C_i, C_{m,i})\right]^2$$

Y en este ejemplo concreto

$$B = \frac{1}{10}(0,15-0)^2 + (0,85-1)^2 + (0,20-0)^2 + (0,80-1)^2 + (0,60-0)^2 + (0,40-1)^2$$
$$+(0,55-0)^2 + (0,45-1)^2 + (0,62-0)^2 + (0,38-1)^2 + (0,12-0)^2 + (0,88-1)^2 +$$
$$(0,83-0)^2 + (0,17-1)^2 + (0,30-0)^2 + (0,70-1)^2 + (0,70-0)^2 + (0,30-1)^2 +$$
$$(0,35-0)^2 + (0,65-1)^2 = 0,5030$$

Como los valores de B están entre 0 y 2 es un error moderado

Ejercicio 2.- Dadas las siguientes matrices de confusión asociadas a otros tantos clasificadores, Justifique con que clasificador se quedaría atendiendo a los valores de CCR, F y mínima sensibilidad de las dos clases. A partir de estas matrices calcule el área de la curva ROC de cada clasificador. Este área es fiable o necesitaría tener más matrices asociadas a cada clasificador.

	Tabla 1	Tabla 2	Tabla 3
	C. pre	C. pre	C. pre
C. real	$\begin{pmatrix} 90 & 10 \\ 15 & 5 \end{pmatrix}$	$\begin{pmatrix} 80 & 20 \\ 10 & 10 \end{pmatrix}$	$\begin{pmatrix} 70 & 30 \\ 5 & 15 \end{pmatrix}$

Solución.- $\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \qquad \text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall=Sensibilidad} = \frac{TP}{TP+FN}$$

S_1= TP/(TP+FN) y S_2= TN/(TN+FP)

Para la primera matriz CCR=95/120=**0,79** TP=90, FN=10, FP=15, por lo que para la primera matriz la Precisión es 90/105= 0,86 y la Sensibilidad es 90/100=0,90 F=1,5428/1,7571=**0,88** y para MS S_1=90/100=0,90 y S_2=5/20= 0,25, luego MS= 0,25.

Para la segunda matriz CCR=90/120=0,75 TP=80, FN=20, FP=10, por lo que para la segunda matriz Precisión= 80/90= 0,89 y la Sensibilidad es 80/100=0,80 F=1,422/1,688=0,84 y para MS S_1=80/100=0,80 y S_2=10/20= 0,50, luego MS= 0,50.

Para la tercera matriz CCR=85/120=0,71 TP=70, FN=15, FP= 5, por lo que para la tercera matriz Precisión= 70/75= 0,93 y la Sensibilidad es 70/100=0,70 F= 1,307/1,633=0,80 y para MS S_1=70/100=0,70 y S_2=15/20= 0,75, luego **MS= 0,70**.

Luego atendiendo a CCR el mejor es el primero, en cuanto a F este también es el mejor; pero en cuanto a MS el mejor es el tercero.

Los tres puntos de cada una de las curvas ROC son para Y=TPR, 90/100=0,90 versus X=FPR, 15/20=0,75 para la primera cuya área es (0,75*0,9)/2+(1-0,75)*0,9+(1-0,75)*(1-0.9)/2=0,58.

Para la segunda es Y=TPR, 80/100=0,8 versus X=FPR, 10/20=0,5 cuya área es (0,5*0,8)/2+(1-0,5)*0,8+(1-0,5)*(1-0.8)/2=0,65.

Para la tercera es Y=TPR, 70/100=0,7 versus X=FPR, 5/20=0,25 cuya área es (0,25*0,7)/2+(1-0,25)*0,7+(1-0,25)*(1-0.7)/2=0,73.

Luego el mejor en AUC es el tercero; aunque es necesario contar con muchos más puntos para poder tener una curva ROC suficientemente robusta.

Ejercicio 3.- En función de los resultados de dos clasificadores (Ver la Tablas 1 y 2). Calcule los valores de las métricas, F1, Accuracy y Precisión para analizar la bondad de los mismos. ¿Qué sentido tiene cada métrica? Con que clasificador se quedaría?

Tabla 1

	Clase predicha	
Clase Real	20	80
	0	10

Tabla 2

	90	10
	10	0

Tabla 3

	50	50
	5	5

Tabla 4

	Clase predicha	
Clase Real	70	30
	8	2

Si en vez de dos clasificadores fuera un sólo clasificador con cuatro matrices de confusión (Ver la Tablas 1, 2, 3 y 4) para diferentes umbrales de decisión. Calcule el Área bajo la curva ROC e interprete el resultado

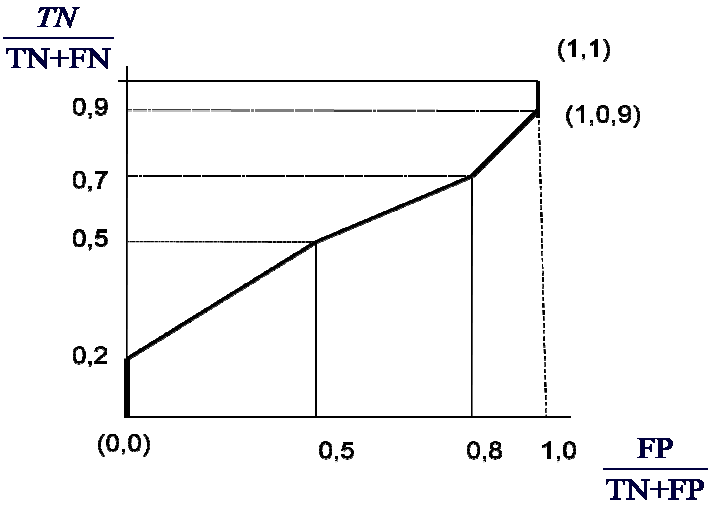
Solución.-

$$\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix} \quad Precision=\frac{TP}{TP+FP} \quad Recall=Sensibilidad=\frac{TP}{TP+FN} \quad F_1=\frac{2\times Precision\times Recall}{Precision+Recall}$$

CCR1=30/110=0,272, CCR2= 90/110=0,818, P1=20/20=1, P2=90/100=0,9, R1= 20/100=0,2, R2= 90/100=0,9,

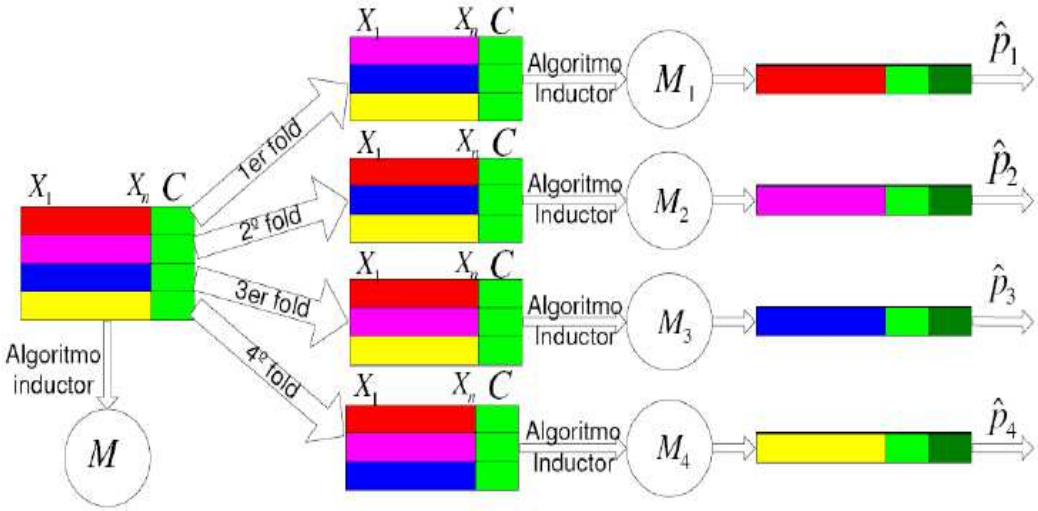
F11=(2*1*0,2)/(1,2)=0,33 y F12=(2*0,9*0,9)/(1,8)=0,9
Por CCR es mejor el clasificador 2 y también por F1 al tener valores mayores que el clasificador 1

b) **Curva ROC del clasificador**
Punto 1, Sensitivity = 20/100=0,2, 1-Specificity= 0
Punto 2, Sensitivity = 50/100=0,5, 1-Specificity= 5/10=0,5
Punto 3, Sensitivity = 70/100=0,7, 1-Specificity= 8/10=0,8
Punto 4, Sensitivity = 90/100=0,9, 1-Specificity= 10/10=1



AUC=0,2*0,5+(0,5*0,3)/2+0,3*0,5+(0,3*0,2)/2+0,2*0,7+(0,2*0,2)/2=0.515

Ejercicio 4.-
En el siguiente grafico ¿Qué tipo de diseño se realiza para validar la calidad de un clasificador? Explique cómo se construye un estimador de la precisión y analice el sentido de la Esperanza Matemática y la Varianza del estimador



$$E(\hat{p}_M)=\frac{1}{k}\sum_{i=1}^{k}E(\hat{p}_i);\quad V(\hat{p}_M)=\frac{1}{k^2}(\sum_{i=1}^{k}V(\hat{p}_i)+2\sum_{i=1<j}^{k}Cov(\hat{p}_i,\hat{p}_j))$$

Solución.-

El diseño es un 4-fold cross-validation. Esto es consideramos cuatro modelos de clasificación, de forma tal que el estimador sea la media de las probabilidades de pertenencia a la clase C1 o positiva.

Esto es $$\hat{p}_M = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_i$$

Este estimador es insesgado dado que la Esperanza Matemática del estimador es el parámetro a estimar. $E(\hat{p}_M) = \frac{1}{4}\sum_{i=1}^{4}E(\hat{p}_i)$

Por otra parte si los estimadores de la probabilidad de pertenencia a la clase positiva $\hat{p}_i, \hat{p}_j$ estan incorrelados, la varianza del estimador será

$$V(\hat{p}_M) = \frac{\sum_{i=1}^{4}V(\hat{p}_i)}{4^2}$$

siendo este un valor inferior a la varianza de cada estimador individual. Pero si las variables están correladas; entonces la varianza aumenta según la segunda sumatoria de la parte derecha de la ecuación

$$V(\hat{p}_M) = \frac{1}{4^2}(\sum_{i=1}^{4}V(\hat{p}_i) + 2\sum_{i=1<j}^{4}Cov(\hat{p}_i, \hat{p}_j))$$

Ejercicio 5.- Dadas las siguientes matrices de confusión asociadas a otros tantos clasificadores, Justifique con que clasificador se quedaría atendiendo a los valores de CCR, mínima sensibilidad y media geométrica.

	Tabla 1			Tabla 2			Tabla 3		
	C. pre			C. pre			C. pre		
C. real	90	10	10	80	30	0	110	0	0
	10	80	10	0	80	20	0	100	0
	5	5	10	5	10	5	10	10	0

Solución.- CCR1=180/230=0,7826; CCR2=165/230=0,7173; CCR3= 210/230=0,9130
MS1={90/110,80/100,10/20}=0,5;MS2={80/110,80/100,5/20}=0,25;
MS3={110/110,100/100,0/20}=0

GM1= $\sqrt[3]{(90/110)*(80/100)*(10/20)} = 0,6894$

GM2= $\sqrt[3]{(80/110)*(80/100)*(5/20)} = 0,5275$

GM3= $\sqrt[3]{(110/110)*(100/100)*(0/20)} = 0$

Sin duda es preferible el clasificador primero porque aunque no tiene el mejor valor de CCR si lo tiene en MS y GM.

Ejercicio 6.- Se define la Entropía de Shannon (1948), H(X), de una variable aleatoria discreta X como la esperanza matemática de la cantidad de información I(X). Sea una variable aleatoria asociada al resultado del lanzamiento de un dado. Calcule la entropía de esta variable, en el caso de que el dado sea perfecto y en el caso en el que tenga dos seises y ningún cinco. ¿Cuál de las dos variables presenta mayor incertidumbre?

Solución.-

En el caso del dado perfecto, la entropía es

$$H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log_2 \frac{1}{6} = 2.58$$

y en el caso del dado sesgado incorrecto es

$$H(X) = -\sum_{i=1}^{4} \frac{1}{6} \log_2 \frac{1}{6} - \frac{2}{6} \log_2 \frac{2}{6} = 2.25$$

Luego es más incierto el resultado del dado primero que el del segundo dado, al ser mayor la entropía.

Ejercicio 7.- Definición de entropía de dos variables. Calcular la entropía condicional de la variable puntuación obtenida por la segunda tirada de un dado en función de la variable puntuación obtenida en la primera tirada del mismo dado

$$H(X \mid Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i \mid y_j)$$

Solución.-

Son dos variables aleatorias independientes por lo tanto

$$H(\mathrm{X} \mid Y) = H(X) = -\sum_{i=1}^{6} \frac{1}{6} \log_2 \frac{1}{6} = 2.58$$

Siendo X la variable aleatoria asociada al lanzamiento del segundo dado

Ejercicio 8.- ¿La siguiente ecuación que tipo de modelo representa? Explique lo que significan las funciones, variables y coeficientes que aparecen en el mismo. Construya el grafo del modelo ¿Cuál es su misión? Calcule la probabilidad de pertenencia a la clase positiva de un determinado patrón. A que clase pertenece el patrón de características (0,1)?

$$\log \frac{p\left(y=1 \middle| \mathbf{x}_i\right)}{1-p\left(y=1 \middle| \mathbf{x}_i\right)} = -1 + 3x_1 - x_2 + 3x_1^2 x_2^{-1}$$

Solución.-

Es un modelo de regresión logística no lineal, que veremos más adelante; pero hay una parte que si podemos hacer con los conocimientos que tenemos ahora.

Si despejamos $p\left(y=1 \middle| \mathbf{x}_i\right)$, a dicha probabilidad por simplicidad la llamaremos p, tenemos que

$$\log\frac{p}{1-p} = -1+3x_1 - x_2 + 3x_1^2 x_2^{-1}, \text{ luego}$$

$$\frac{p}{1-p} = e^{(-1+3x_1-x_2+3x_1^2x_2^{-1})}$$

y despejando p tenemos que

$$p = \frac{e^{(-1+3x_1-x_2+3x_1^2x_2^{-1})}}{1+e^{(-1+3x_1-x_2+3x_1^2x_2^{-1})}}$$

es la probabilidad de pertenencia de un patrón a la clase positiva.
Para un patrón de características (0,1) tenemos que

$p = \frac{e^{(-1-1)}}{1+e^{(-1-1)}} = \frac{e^{-2}}{1+e^{-2}} = 0{,}119$, luego pertenece a la clase negativa

Ejercicio 9.- Calcule la matriz de confusión y las puntuaciones de Brier obtenidas con el clasificador anterior para la base de datos

X_1	X_2	Clase
2	4	0
3	6	0
2	7	0
-1	3	0
8	8	1
6	9	1
5	7	1
1	9	1

Nota (Si no sabe calcular la probabilidad en el ejercicio anterior utilice un clasificador inventado)

Solución.-La probabilidad de pertenencia a la clase positiva es

$$p = \frac{e^{(-1+3x_1-x_2+3x_1^2x_2^{-1})}}{1+e^{(-1+3x_1-x_2+3x_1^2x_2^{-1})}}$$

luego construimos la tabla de probabilidades

X_1	X_2	Clase	P(y=1) Clase positiva	P(y=0) Clase negativa	decisión
2	4	0	0,9820	0,0180	F en 0
3	6	0	0,9985	0,0015	F en 0
2	7	0	0,4290	0,5710	E en 0
-1	3	0	0,0025	0,9975	E en 0
8	8	1	1,0000	0,0000	E en 1
6	9	1	1,0000	0,0000	E en 1
5	7	1	0,9878	0,0122	E en 1
1	9	1	0,0013	0,9987	F en 1

Para el primer patrón tenemos

$$p=\frac{e^{(-1+3*2-4+3*4*1/4)}}{1+e^{(-1+3*2-4+3*4*1/4)}}=\frac{e^{4}}{1+e^{4}}=0,9820$$

para el segundo patrón

$$p=\frac{e^{(-1+3*3-6+3*9*1/6)}}{1+e^{(-1+3*3-6+3*9*1/6)}}=\frac{e^{6,5}}{1+e^{6,5}}=0,9985$$

para el tercero

$$p=\frac{e^{(-1+3*2-7+3*4*1/7)}}{1+e^{(-1+3*2-7+3*4*1/7)}}=\frac{e^{-2/7}}{1+e^{-2/7}}=0,4290$$

para el cuarto

$$p=\frac{e^{(-1+3*(-1)-3+3*1*1/3)}}{1+e^{(-1+3*(-1)-3+3*1*1/3)}}=\frac{e^{-6}}{1+e^{-6}}=0,0025$$

para el quinto

$$p=\frac{e^{(-1+3*8-8+3*64*1/8)}}{1+e^{(-1+3*8-8+3*64*1/8)}}=\frac{e^{39}}{1+e^{39}}\approx 1$$

para el sexto

$$p=\frac{e^{(-1+3*6-9+3*36*1/9)}}{1+e^{(-1+3*6-9+3*36*1/9)}}=\frac{e^{20}}{1+e^{20}}\approx 1$$

para el séptimo

$$p=\frac{e^{(-1+3*5-7+3*25*1/7)}}{1+e^{(-1+3*2-7+3*4*1/7)}}=\frac{e^{124/7}}{1+e^{124/7}}=0,9878$$

para el octavo

$$p=\frac{e^{(-1+3*1-9+3*1*1/9)}}{1+e^{(-1+3*1-9+3*1*1/9)}}=\frac{e^{-20/3}}{1+e^{-20/3}}=0,0013$$

La matriz de confusión es

	C. predicha	
C. real	3	1
	2	2

El CCR=5/8, las sensibilidades son para la clase 1, 3/4 y para la clase 0, 2/4. La mínima sensibilidad es por tanto 2/4

La puntuación de Briers será por tanto

$$B=\frac{1}{8}\left(0,9820-1\right)^{2}+(0,0180-0)^{2}+(0,9985-1)^{2}+(0,0015-0)^{2}+(0,4290-1)^{2}+(0,5710-0)^{2}$$
$$+(0,0025-1)^{2}+(0,9975-0)^{2}+(1-1)^{2}+(0-0)^{2}+(1-1)^{2}+(0-0)^{2}+$$
$$(0,9878-1)^{2}+(0,0122-0)^{2}+(0,0013-1)^{2}+(0,9987-0)^{2}=0,5797$$