

TICS411- Prueba 2

MINERÍA DE DATOS

Universidad Adolfo Ibáñez

2024-1

Profesores: Claudio Díaz - Miguel Carrasco - Alfonso Tobar
Fecha: 11/06/24

Nombre: _____ Rut: _____ Sección: _____

*Esta prueba contiene **11 páginas** y **12 preguntas** totalizando 21 puntos. Además, en la parte final dispone de un **Formulario**. Recuerde responder con letra clara y legible.
¡Buena suerte!*

Parte I: Preguntas de Selección Múltiple (20 min)

- (1 punto) ¿Cuál de las siguientes alternativas es más correcta con respecto a KNN?
 - Los outliers tienen un impacto mínimo en el rendimiento de KNN debido a su enfoque basado en la proximidad de los vecinos más cercanos.
 - Los outliers pueden distorsionar los resultados de KNN, ya que afectan significativamente la proximidad de los vecinos más cercanos.
 - KNN es completamente inmune a la presencia de outliers, ya que ignora cualquier punto que se aleje demasiado de la mayoría de los datos.
 - KNN puede manejar outliers si se disminuye al mínimo el número de K.
- (1 punto) De acuerdo a la Figura 1 ¿Qué K escogería para que todos los puntos sean predichos como negativos de acuerdo a un Clasificador KNN?

A. 1

+

+

— —

B. 5

— —

C. 7

+

+

— —

D. 9

Figura 1: Clasificador KNN

3. (1 punto) Un modelo de **Árboles de Decisión** entregó un **F1-Score de 0.95** de entrenamiento, mientras que su puntaje en el set de prueba fue de un **F1-Score de 0.53**. ¿Qué puede decir con respecto al ajuste del modelo?
- A. El modelo está correctamente ajustado ya que el puntaje de entrenamiento es muy bueno. El puntaje de prueba sólo nos sirve para indicar que los hiperparámetros no fueron correctos.
 - B. El modelo presenta un claro underfitting. No fue capaz de aprender lo suficiente, por lo que eso se refleja en un bajo puntaje en el set de Validación.
 - C. El modelo presenta un claro overfitting. El modelo fue capaz de aprender demasiado en el set de entrenamiento pero no fue capaz de generalizar al set de validación.
 - D. El modelo obtuvo hiperparámetros óptimos, ya que el puntaje de entrenamiento es suficientemente alto.
4. (1 punto) Se tiene un modelo de clasificación, en la cual se debe detectar **potenciales fraudes** que se están realizando en el sistema público. Se sabe que el porcentaje de fraudes es bajo, por lo que esto no sería una tarea tan sencilla. Considerando que la clase positiva es fraude, ¿Qué **métrica** consideraría **más apropiada** para evaluar su modelo?
- A. Accuracy, ya que es la métrica más apropiada cuando existe un desbalance de clases.
 - B. Precision, ya que pone más énfasis en los Falsos Positivos.
 - C. Recall, ya que pone más énfasis en los Falsos Negativos.
 - D. Curva ROC, ya que no le importa el umbral de probabilidad escogido.
5. (1 punto) En el contexto de los **Árboles de Decisión**, ¿Cómo se relaciona la **Poda (pruning)** y la impureza para mejorar el rendimiento del modelo?
- A. La poda se utiliza para maximizar la impureza en las hojas del árbol, evitando nodos innecesarios.
 - B. La poda no penaliza la creación de nuevos nodos, aumentando la complejidad del árbol y mejorando su precisión.
 - C. La poda no está relacionado a la construcción ni a la optimización de árboles de decisión.
 - D. La poda elimina ramas del árbol que no contribuyen significativamente a la reducción de impureza, evitando sobreajuste (overfitting).

Hint: Pureza se refiere a nodos hojas que tengan **sólo una clase**.

6. (1 punto) Indique cuál de las siguientes afirmaciones **son verdaderas** con respecto al algoritmo de Clasificación KNN.
- I. El algoritmo KNN utiliza la distancia Euclidiana ya que este pondera la covarianza entre las clases.
 - II. A medida que aumenta el número de dimensiones (más características), se pierde la noción de distancia ya que aumenta la distancia entre los puntos.
 - III. El algoritmo KNN no requiere un entrenamiento previo ya que la relación se genera a partir de los datos.
 - IV. El algoritmo KNN utiliza la matriz de covarianza para realizar un cambio de escala de las variables y así ponderar correctamente cada una de las características.
- A. Solo I
 - B. Solo III
 - C. Solo II y IV
 - D. Solo II y III
 - E. Todas son correctas.
7. (1 punto) Indique cuál de las siguientes afirmaciones **son verdaderas** con respecto a la Validación Cruzada (K-Fold).
- I. La validación cruzada con parámetro K , es una técnica que permite determinar el mejor rendimiento del clasificador en una de sus K versiones.
 - II. La validación cruzada con parámetro K es una técnica que permite conocer el rendimiento promedio del clasificador.
 - III. Mientras mayor sea el parámetro K de la validación cruzada, menor será el conjunto de entrenamiento del modelo.
 - IV. Mientras menor sea el parámetro K de la validación cruzada, mejor será la estimación de los parámetros del clasificador.
- A. Solo I
 - B. Solo II
 - C. Solo II y III
 - D. Solo II y IV
 - E. Todas son correctas

Parte II: Preguntas de Desarrollo (30 min)

8. Se tiene la siguiente curva de validación (Ver Figura 2).
- (a) (1 punto) Indique en el gráfico, ¿Cuál sería la complejidad de un modelo óptimo y en qué regiones el modelo sufriría de Overfitting y Underfitting?

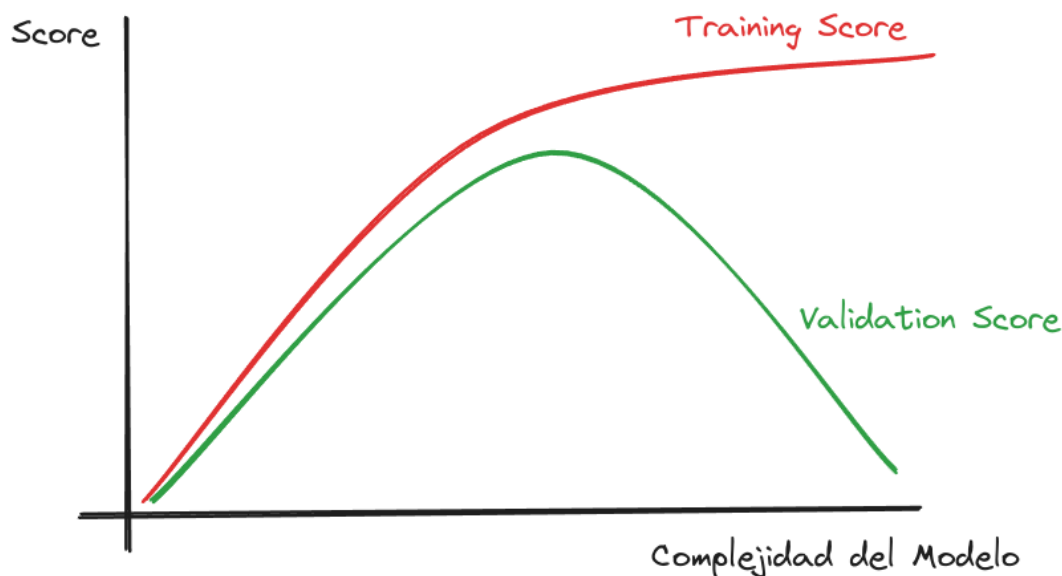


Figura 2: Curva de Validación de un Modelo

- (b) (1 punto) ¿Qué caracteriza el Underfitting?

.....

.....

.....

.....

.....

.....

.....

.....

(c) (1 punto) ¿Qué caracteriza el Overfitting?

.....

.....

.....

.....

.....

.....

.....

.....

.....

9. (1 punto) ¿Por qué el Modelo Naive Bayes se considera “**Ingenuo**” o “**Inexperto**”?

.....

.....

.....

.....

.....

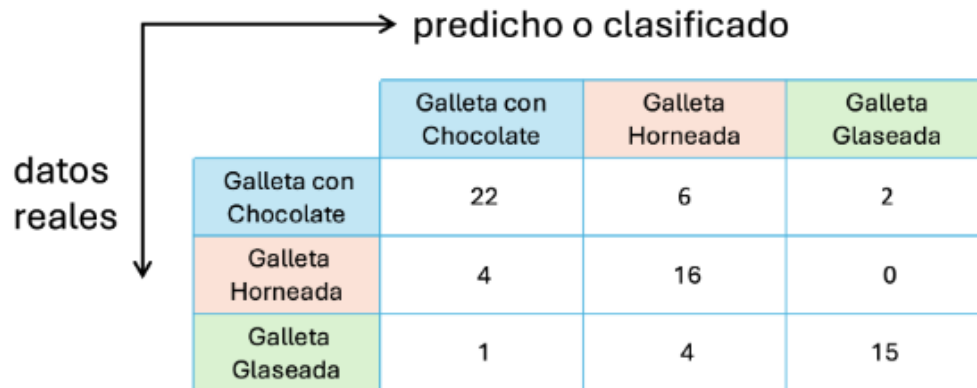
.....

.....

.....

.....

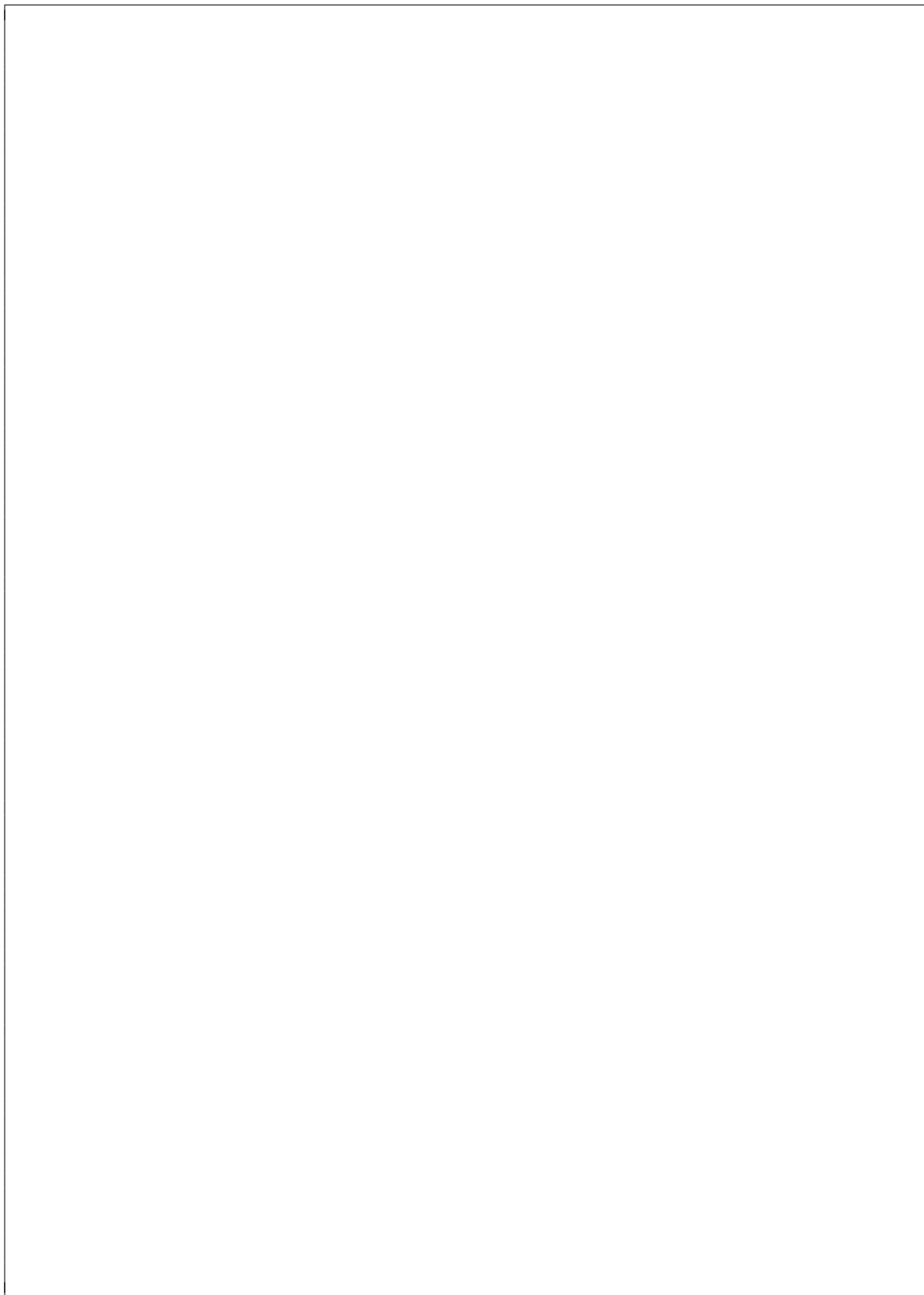
10. Una empresa de alimentos se encuentra diseñando un sistema de inspección automático muy avanzado que es capaz de determinar el tipo de galleta a partir de un algoritmo de clasificación “*ultra secreto*” que utiliza visión por computador. Se le ha pedido a usted determinar algunos indicadores y a partir de ello definir si el rendimiento del sistema supera un umbral específico. A partir de la Matriz de Confusión (ver Figura 4) responda las siguientes preguntas:



	predicho o clasificado		
	Galleta con Chocolate	Galleta Horneada	Galleta Glaseada
Galleta con Chocolate	22	6	2
Galleta Horneada	4	16	0
Galleta Glaseada	1	4	15

Figura 4: Matriz de Confusión

- ($\frac{1}{2}$ punto) ¿Cuál es el Accuracy obtenido de la Matriz de Confusión?
- (1 punto) Determine el F1-Score de cada clase y el F1-Score promedio.
- ($\frac{1}{2}$ punto) ¿Cuál es la clase que posee mejor clasificación a partir de los resultados obtenidos en la parte b)?
- (1 punto) Si la empresa que diseñó el algoritmo de visión por computador ha modificado el algoritmo de clasificación obteniendo una mejora de 10 % del Accuracy respecto al rendimiento original, ¿es correcto afirmar que todas las clases mejoraron en un 10 %?



Parte III: Ejercicios Prácticos (60 min)

11. Considere el conjunto de cinco puntos de entrenamiento mostrado en la Tabla 1:

	X	Y	Clase
0	7	4	1
1	5	7	1
2	3	5	1
3	5	7	-1
4	2	3	-1

Tabla 1: Dataset KNN

- (a) (3 puntos) Clasifique el punto (**X=4, Y=6**) usando un clasificador KNN con **k = 3** usando **distancia Manhattan**. Su respuesta debe ser +1 o -1. Debe acompañar su respuesta con los cálculos que realizó.

12. Un experto en analítica de datos se encuentra estudiando si se realizan o no torneos de tenis a partir únicamente de la humedad relativa del aire. Para ello ha tomado registros históricos de la última temporada y ha indicado en una última columna si se ha jugado o no el torneo.

	Humedad	Se juega?
1	85	No
2	90	No
3	86	Sí
4	96	Sí
5	80	Sí
6	65	Sí
7	70	Sí
8	70	No
9	95	No
10	80	Sí
11	91	No
12	70	Sí
13	90	Sí
14	75	Sí

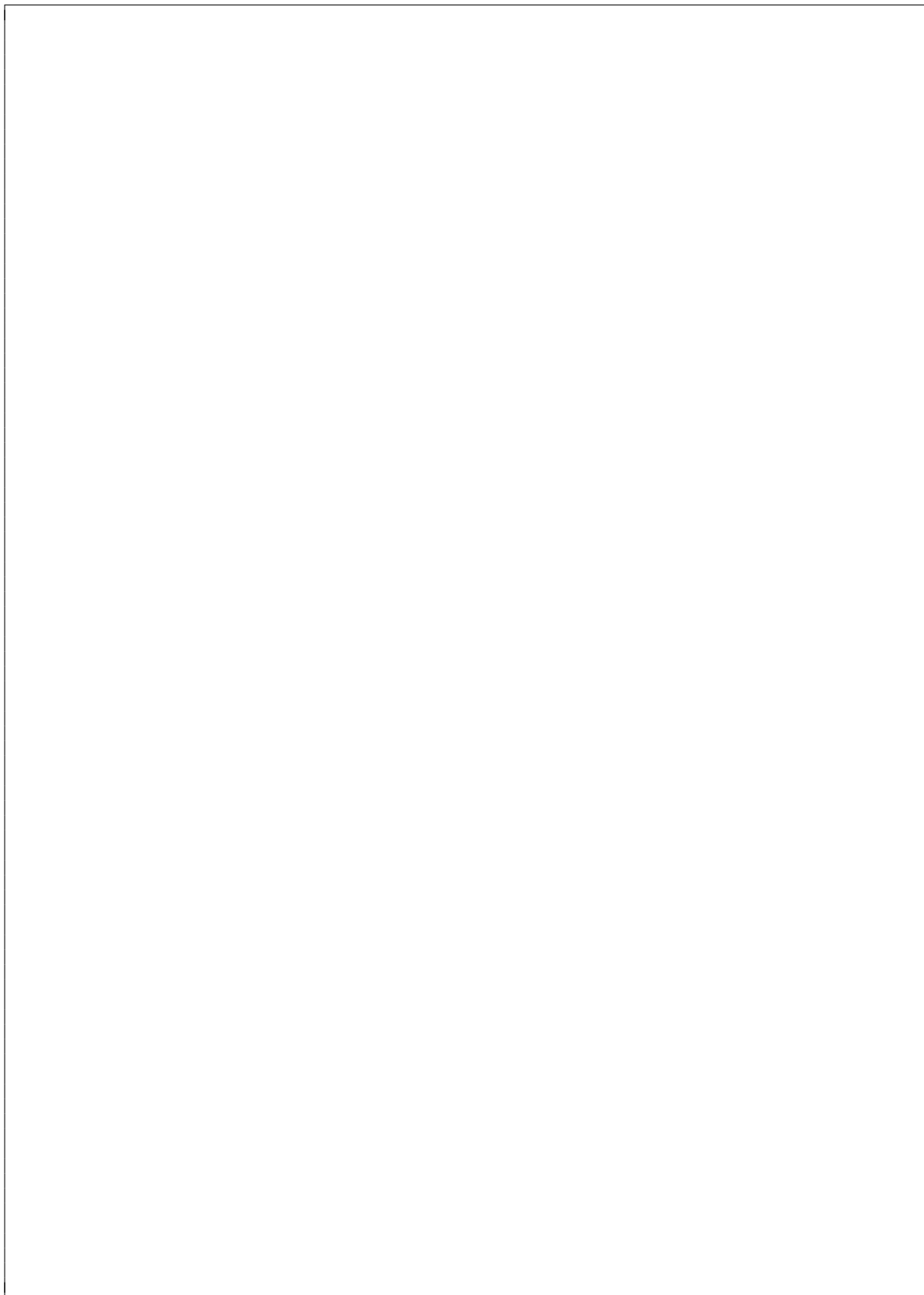
Tabla 2: Dataset Naive Bayes

Basado en la información disponible en la tabla, se le ha encargado a usted responder las siguientes preguntas:

- (a) (1 punto) ¿Cómo puede estimar la probabilidad para una variable que es continua?
- (b) (1 punto) Suponga que usted ha recibido información de un especialista para definir rangos de la variable humedad en forma categórica:
- Humedad Baja: $[0 - 50]$
 - Humedad Media : $[51 - 75]$
 - Humedad Alta: $[76 - 100]$

¿Cuáles serían las probabilidades para el evento (**Humedad = Media**) cuando la clase es Sí y cuando la clase es No?

- (c) (1 punto) Según el resultado anterior, determine si se juega o no el partido de acuerdo a un clasificador de Naive Bayes utilizando un nivel de (**Humedad = Media**).
- (d) (1 punto) Si empleamos los datos originales con variable continua, ¿Cuál sería la predicción en el caso que la (**Humedad = 65**)?



Formulario

$$\text{GeneralScoring}(\mathbf{M}) = \sum_{i=1}^N d[f(x_i); M], y(i)]$$

$$\text{ZeroOneLoss}(\mathbf{M}) = \frac{1}{N} \sum_{i=1}^N I[f(x_i); M], y(i)]$$

$$\text{SquaredLoss}(\mathbf{M}) = \frac{1}{N} \sum_{i=1}^N [f(x_i); M] - y(i)]^2$$

Confusion matrix		Predicted Class	
		No	Yes
Actual Class	No	True Negative	False Positive
	Yes	False Negative	True Positive

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Gini

$$G(X) = 1 - \sum_{x_i} p(x_i)^2$$

Entropía

$$H(X) = - \sum_{x_i} p(x_i) \log_2 p(x_i)$$

Naive Bayes

$$P(C|X_1, X_2, \dots, X_k) \propto \prod_{i=1}^k P(X_i|C)P(C)$$

Variables Continuas

$$P(X_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$