

Práctica 2 - Naive Bayes
Sistemas de Aprendizaje Automático - C.E. Inteligencia Artificial y Big Data.
I.E.S. Camas - Curso 2025-26

Realiza con Python un programa que cumpla con los requisitos que se muestran a continuación. Entrégalo en un cuaderno de Jupiter de Google Colab.

En este ejercicio vamos a entrenar un modelo basado en Naive Bayes para predecir el cáncer de mama con el *Breast Cancer Wisconsin (Diagnostic) Data Set*. Tienes toda la información del dataset en la siguiente URL:

<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?resource=download>

Realiza las siguientes tareas:

1. Tarea 1: Importación de librerías, carga del dataset en un data frame y muestra inicial.

- 1.1. Carga el dataset en un dataframe llamado *datos*.
- 1.2. Muestra el dataset.
- 1.3. Muestra un resumen estadístico del dataset de forma numérica.
- 1.4. Muestra al menos tres gráficos estadísticos sobre el dataset, elige los que creas más oportunos.

2. Tarea 2: Preparación de los datos:

En esta tarea debes crear un nuevo dataframe con los datos preparados, con el nombre *breastcancer*. El dataframe original *datos*, no debe modificarse.

Los requisitos para los valores de los distintos datos en el nuevo dataframe (tanto atributos como resultado) son los siguientes:

- No pueden estar vacíos.
 - El campo objetivo (*diagnosis*), debe contener “M” o “B”, siendo M=Maligno y B=Benigno.
 - Se considera que los siguientes datos no pueden ser igual a cero: *radius_mean*, *perimeter_mean*, *area_mean*, *radius_worst*, *perimeter_worst*, *area_worst*, *compactness_mean*, *concavity_mean*, *concave_points_mean*, *smoothness_mean*, *fractal_dimension_mean*.
 - No se realizará cálculo de outliers, ni de observaciones influyentes.
- 2.1. Comprueba que todos los datos cumplen los requisitos indicados. Si algún dato no lo cumple, elimina la fila completa.
 - 2.2. Modifica o mapea los tipos que creas oportuno.
 - 2.3. Escala los atributos por el método de estandarización que creas más oportuno. Ten en cuenta el clasificador bayesiano que vayas a usar después.

3. Tarea 3: Preparación de los conjuntos de entrenamiento y de test

Se creará un conjunto de entrenamiento con el 80% de las observaciones y un conjunto de test con el 20% de los valores. Utiliza una semilla aleatoria para obtener siempre los mismos resultados, eso te facilitará las pruebas. Haz el reparto de forma proporcional, activando *stratify*.

4. Tarea 4: Elección del clasificador bayesiano

Elige el clasificador bayesano que creas más apropiado. Puedes hacer pruebas si lo crees oportuno.

Explica justificadamente tu elección en un cuadro de texto del cuaderno.

5. Tarea 5: Muestra de los resultados del modelo seleccionado

Se ejecutará el modelo elegido con el conjunto de test, haciendo una predicción. Se mostrarán los resultados obtenidos, contrastando los valores de la etiqueta predicha con los reales, al menos de esta manera:

- 5.1. Se mostrarán los valores numéricos de eficacia
- 5.2. Se mostrará una matriz de confusión gráfica.
- 5.3. Se creará un dataframe llamado *predicciones* donde aparezcan los valores del conjunto de test con todas sus columnas y dos columnas más añadidas al final: La columna *prediccion*, con el valor predicho del resultado y la columna *acierto*, que contendrá un 0 si la predicción es distinta al resultado real y un 1 si es igual.

6. Tarea 6: Guarda la predicción en un fichero csv

- 6.1. Guarda el dataframe *predicciones* en tu disco local, con el nombre *predicciones_nombre_apellido1_apellido2.csv*.

Observa el siguiente ejemplo para exportar un dataframe a csv y descargarlo al disco local de tu PC:

```
from google.colab import files
mi_dataframe.to_csv('mi_csv.csv')
files.download('mi_csv.csv')
```

OBSERVACIONES A TENER EN CUENTA:

- ✓ El código debe estar comentado.
 - ✓ Se valorará la simplicidad y legibilidad del código.
 - ✓ Se valorará el añadido de bloques de tipo texto entre los bloques de código, en los que se explique el proceso y se analicen y valoren los resultados que se vayan obteniendo.
-

Práctica 2 - Naive Bayes
Sistemas de Aprendizaje Automático - C.E. Inteligencia Artificial y Big Data.
I.E.S. Camas - Curso 2025-26

ENTREGA DE LA PRÁCTICA:

En la tarea de la plataforma Moodle, debes entregar lo siguiente:

1. El código de la práctica, que debe ser un archivo de Google Colab, con el nombre *Práctica2_nombre_apellido1_apellido2.ipynb*.
 2. El fichero csv resultante, con el nombre *predicciones_nombre_apellido1_apellido2.csv*.
-

RÚBRICA DE CORRECCIÓN:

ÍTEMS EVALUABLES		PESO
ÍTEM 1	Tarea 1: Importación de librerías, carga del dataset en un data frame y muestra inicial	20%
ÍTEM 2	Tarea 2: Preparación de los datos	20%
ÍTEM 3	Tarea 3: Preparación de los conjuntos de entrenamiento y de test	5%
ÍTEM 4	Tarea 4: Elección del clasificador bayesiano	20%
ÍTEM 5	Tarea 5: Muestra de los resultados del modelo seleccionado	20%
ÍTEM 6	Tarea 6: Guarda la predicción en un fichero csv	5%
ÍTEM 7	Comentarios del código y explicación del proceso en bloques de texto.	10%