TAREA 4

1. Supongamos que (X, Y) son variables aleatorias discretas con la siguiente distribución conjunta:

	$X=1$	$X=2$	$X=3$	$X=4$
$Y=0$	0,1	0,05	0,05	0,15
$Y=1$	0,12	0,1	0,25	0,18

   Queremos predecir Y en base del valor observado para X.
   a) Calcula el clasificador Bayesiano Optimo si equivocarse de categoría tiene costo 1 y no equivocarse tiene costo 0. ¿Cuál es el costo (error) promedio para este clasificador?
   b) Calcula el clasificador Bayesiano Optimo si clasificar una observación mal cuando el verdadero valor es $Y = 1$ tiene un costo 3 y en el otro caso tiene costo 2.
2. Deriva el clasificador Bayesiano óptimo para el caso de tres clases y una función de costo simétrica cuando:

$$X|Y = 1 \sim \mathcal{N}(\mu_1, \Sigma) \quad X|Y = 2 \sim \mathcal{N}(\mu_2, \Sigma) \quad X|Y = 3 \sim \mathcal{N}(\mu_3, \Sigma)$$

   y

$$P(Y = 1) = 2P(Y = 2) = P(Y = 3).$$

3. (Seleccionando el valor de k en k-medias).
   a) Genere un conjunto de datos de entrenamiento de 200 puntos en $\mathbb{R}^3$ muestreando 100 puntos con coordendas independientes de una normal $\mathcal{N}(4, 1)$ y 100 puntos de una normal $\mathcal{N}(8, 1)$. Ejecute el algoritmo de agrupamiento de k-medias, para $k = 2, 3, \cdots, 15$, usando el conjunto de datos de entrenamiento. Para cada k use diez puntos iniciales aleatorios y solo guarde la solución que tenga el menor valor de la función objetivo de k-medias.

      Nota: Siéntase libre de usar una implementación existente del algoritmo, por ejemplo, puede usar **KMeans** de **sklearn.cluster**, o puede implementar su propia versión.

      Muestra en una gráfica el valor de la función objetivo de k-medias resultante sobre el conjunto de datos de entrenamiento como una función de k. Comenta lo que ves. Qué valor de k seleccionaría basándose solo en esta gráfica?
   b) Genere un conjunto de datos de validación del mismo tamaño que el conjunto de entrenamiento y de la misma manera. Para cada k, asigne cada punto o dato de validación a la media de clúster aprendida más cercana en la parte (a) y muestre en una gráfica el valor de la función objetivo resultante de k-medias usando los datos de validación como una función de k. Comenta lo que ves. Qué valor de k se seleccionaría si usara como criterio de selección el minimo valor de la funcion objetivo para los datos de validación?
   c) Para cada k, calcule y muestre en una gráfica el indice (score) **Calinski Harabasz** (CH) calculado para los datos de entrenamiento. Muestra CH como una función de k, y comente al respecto. Que valor de k maximiza este criterio?

```
from sklearn.metrics import calinski_harabasz_score
ch = calinski_harabasz_score(X, labels)
```

4. (Initializando k-medias). En este problema comparán varios esquemas de inicialización para al algoritmo de agrupamiento k-medias
   a) El esquema de inicialización aleratorio o 'random' .
   b) El esquema de inicialización 'k-means++'.
   c) Proponga su propio esquema de inicialización.
   Nota: el parámetro **init** de **KMeans** puede ser 'k-means++', 'random'. También puede ser el nombre de una función que recibe como parametros X, n_clusters y random_state y devuelve una initialización, ie, es decir devuelve un **ndarray** con n_clusters filas y el número de columnas es el número de características (o columnas) de X. Por ejemplo: 'inicializador tonto'

```
def my_init(X, n_clusters, random_state):
    return X[:n_clusters,:]

kmeans = KMeans(init=my_init,...)
```

Para $k \in \{2, 4, 8, 16, 32\}$ y para los 3 esquemas de inicialización, ejecute el algoritmo k-medias para la imagen a color Colorful-Flowers.jpg y obtenga la representación comprimida; considere cada pixel como un punto en 3 dimensiones.

Nota: Siéntase libre de usar una implementación existente del algoritmo, por ejemplo, puede usar **KMeans** de **sklearn.cluster**, o puede implementar su propia versión.

a) Ejecute cada inicialización 5 veces con diferentes semillas y muestre en una gráfica el mínimo y la media de la función obejtivo de k-medias como una función de k.

b) Para cada procedimiento y para cada k, visualice la imagen comprimida usando la mejor corrida aleatoria (Nota: No tiene que mostrar el resultado de las imágenes en el documento, pero el codigo que suba debe permitir mostrar las imagenes). Cual es el menor valor de k para el cual está satisfecho con el resultado obtenido?

5. Es importante que las compañías de tarjetas de crédito puedan reconocer transacciones fraudulentas con tarjetas de crédito para que a los clientes no se les cobren artículos que no han comprado. (Nota: Información tomada de Kaggle)

a) Descarga la base de datos 'creditcard.csv' de la pagina Kaggle, https://www.kaggle.com/mlg-ulb/creditcardfraud

Descripción de la base de datos:

El conjunto de datos contiene transacciones realizadas con tarjetas de crédito en septiembre de 2013 por titulares de tarjetas europeos. Este conjunto de datos presenta transacciones que ocurrieron en dos días, donde tenemos 492 fraudes en 284,807 transacciones. El conjunto de datos está muy desequilibrado, la clase positiva (fraudes) representa $0,172\,\%$ de todas las transacciones.

La base de datos solo contiene variables de entrada numéricas que son el resultado de una transformación PCA. Desafortunadamente, debido a problemas de confidencialidad, no se pueden proporcionar las características originales, ni más información general sobre los datos. Las características $V1, V2, \cdots, V28$ son los principales componentes obtenidos con PCA, las únicas características que no han sido transformadas con PCA son 'Tiempo' y 'Cantidad' ('Time' and 'Amount').

La función 'Tiempo' ('Time') contiene los segundos transcurridos entre cada transacción y la primera transacción en el conjunto de datos. La función 'Cantidad' ('Amount') es la cantidad de la transacción. Caracteristica 'Class' es la variable respuesta y toma valor 1 en caso de fraude y 0 en caso contrario.

Comentario: Dada la relación de desequilibrio de clase, se recomienda medir la precisión utilizando el 'Area Under the Precision-Recall Curve' (AUPRC). La precisión de la matriz de confusión no es adecuada en el caso de clasificación con datos no balanceados (unbalanced classification).

```
from sklearn.metrics import average_precision_score as aps
auprc=aps(true_labels,pred_labels)
```

b) Puede usar el siguiente codigo para leer y separar las variables dependiente e independientes:

```
# importing libraries/modules/packages
import pandas as pd

# Loading the data
df = pd.read_csv('creditcard.csv')

# Separating the dependent and independent variables
y=df['Class']
X=df.drop('Class',axis=1)
```

```
X.head()
```

c) Ejecuta 5 veces k-medias con diferentes semillas (random seeds) usando el esquema de inicialización 'k-means++'.

d) Muestra en una gráfica el mínimo de la función objetivo como una función de k, para $k \in \{2, 3, 4, 5, 6\}$

e) Calcula y muestra el índice **Fowlkes Mallows score** (FM) usando el conjunto de entrenamiento. Muestra en una gráfica FM como una función de k, y comenta al respecto. Qué valor de k maximiza este criterio?

```
from sklearn.metrics import fowlkes_mallows_score
fm = fowlkes_mallows_score(true_labels, pred_labels)
```