Ciencia de Datos

Parcial N°2

Problema 1: [2 pts]

a) Defina el clasificador de ensemble con voto ponderado.

b) Suponga que tiene tres clasificadores binarios que calculan las siguientes probabilidades de pertenencia a clase para el punto x

clase 0	clase1
$0,9$	$0,1$
$0,8$	$0,2$
$0,4$	$0,6$

dé la predicción para x del clasificador de ensemble con voto ponderado construido con los tres clasificadores anteriores y el vector de pesos [0,2, 0,2, 0,6].

c) Describa el algoritmo básico de bagging.

Problema 2: [4 pts] Descargar el Telecom Churn Dataset disponible en Kaggle y considerar el problema de prediccion de bajas (churn) de la suscripcion del servicio de la empresa de telefonos. Ignorar completamente las variables `['State', 'International plan', 'Voice mail plan']`.

a) Implementar una búsqueda en grid para optimizar las parámetros C y `gamma` de una *Support Vector Machine* con kernel RBF implementado en scikit-learn. Visualizar los resultados del proceso con el dataset pedido. Por razones de tiempo usar para entrenamiento solo el 30 % de los datos y para test el 20 %.

b) Entrenar `con todos los datos` el modelo ajustado en el punto anterior, predecir el churn y evaluar el resultado imprimiendo un `classification report`, mostrando la matriz de confusión y calculando el coeficiente κ de Cohen.

c) Graficar la curva ROC.

d) Discutir los resultados obtenidos en la evaluación.

Problema 3: [4 pts] Una empresa que desarrolla tecnología para la industria agrícola ha desarrollado un nuevo escáner óptico de alta velocidad que puede medir rápidamente las propiedades geométricas de granos de trigo individuales.

Se quiere utilizar este escáner para clasificar automáticamente los granos de trigo en tres variedades conocidas: Kama, Rosa y Canadiense. Determinar si un modelo de aprendizaje no supervisado puede encontrar estos tres grupos distintos, basándose únicamente en las mediciones del escáner.

La fuente de datos a utilizar es el archivo seeds_dataset.txt:

```
# agregar los nombres de las columnas en base a la documentacion de la página
names = ['area', 'perimeter', 'compactness', 'length', 'width', 'asymmetry', 'length of kernel groove', 'class']
df = pd.read_csv('/content/seeds_dataset.txt', names=names, sep='\s+')
X = df.drop('class', axis=1)
y = df['class']
```

a) Utilize el coeficiente silhouette (valor promedio y plot) para estudiar el resultado al usar k-means con inicialización `random_state=42` y $k = 2, 3, 4$ y 5 clusters. Interprete el resultado y exprese una conclusión.

b) Evalue el resultado de k-means con inicialización `random_state=42` con $k = 2, 3, 4$ y 5 clusters usando los siguientes scores: ajusted Rand index, ajusted mutual information, homogeneity, completeness y V measure. Genere una tabla con estos valores como columnas y el número de grupos como filas (etiquetando las columnas) y agregue los valores del coeficiente silhouette que calculó en [(a)]. Interprete el resultado y exprese una conclusión.

c) Visualice el resultado de k-medias con el k obtenido en el punto [a)] y con el k obtenido en el punto [b)] usando el embedding t-SNE. Compare con el gráfico de las etiquetas reales. Interprete el resultado y exprese una conclusión.