PROMiDAT Iberoamericano
Machine Learning con Python III
Métodos Supervisados

- Las tareas tienen fecha de entrega una semana después a la clase y deben ser entregadas antes del inicio de la clase siguiente.
- Cada día de atraso en implicará una pérdida de 10 puntos.
- Las tareas son estrictamente de carácter individual, tareas iguales se les asignará cero puntos.
- En nombre del archivo debe tener el siguiente formato: `Tarea1_nombre_apellido.pdf`. Por ejemplo, si el nombre del estudiante es Luis Pérez: `Tarea1_luis_perez.pdf`. Para la tarea número 2 sería: `Tarea2_luis_perez.pdf`, y así sucesivamente.
- Esta tarea tiene un valor de un 25 % respecto a la nota total del curso.

Tarea Número 4

- **Pregunta 1:** [35 puntos] En este ejercicio usaremos los datos (`voces.csv`). Se trata de un problema de reconocimiento de género mediante el análisis de la voz y el habla. Esta base de datos fue creada para identificar una voz como masculina o femenina, basándose en las propiedades acústicas de la voz y el habla. El conjunto de datos consta de 3.168 muestras de voz grabadas, recogidas de hablantes masculinos y femeninos.

  El conjunto de datos tiene las siguientes propiedades acústicas (variables) de cada voz:

  - `meanfreq`: frecuencia media (en kHz).
  - `sd`: desviación estándar de frecuencia.
  - `median`: frecuencia mediana (en kHz).
  - `Q25`: primer cuantil (en kHz).
  - `Q75`: tercer cuantil (en kHz).
  - `IQR`: rango intercuantile (en kHz).
  - `skew`: sesgo (ver nota en la descripción de specprop).
  - `kurt`: kurtosis (ver nota en la descripción de specprop).
  - `sp.ent`: entropía espectral.
  - `sfm`: planitud espectral.
  - `mode`: modo frecuencia.
  - `centroide`: centroide de frecuencia (ver specprop).
  - `peakf`: frecuencia de pico (frecuencia con mayor energía).
  - `meanfun`: promedio de la frecuencia fundamental medida a través de la señal acústica.
  - `minfun`: frecuencia mínima fundamental medida a través de la señal acústica.
  - `maxfun`: máxima frecuencia fundamental medida a través de la señal acústica.
  - `meandom`: promedio de la frecuencia dominante medida a través de la señal acústica.

- `mindom`: mínimo de la frecuencia dominante medida a través de la señal acústica.
- `maxdom`: máximo de la frecuencia dominante medida a través de la señal acústica.
- `dfrange`: rango de frecuencia dominante medido a través de la señal acústica.
- `modindx`: índice de modulación. Calculado como la diferencia absoluta acumulada entre las mediciones adyacentes de las frecuencias fundamentales dividida por la gama de frecuencias.
- `género`: Masculino o Femenino (variable a predecir).

Realice lo siguiente:

1. Cargue la tabla de datos `voces.csv` en **Python**.
2. Use Máquinas de Soporte Vectorial en **Python** (con los parámetros por defecto) para generar un modelo predictivo para la tabla `voces.csv` usando el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías. ¿Son buenos los resultados? Explique.
3. Usando la función programada en el ejercicio 1 de la tarea anterior, los datos `voces.csv` y los modelos generados arriba construya un `DataFrame` de manera que en cada una de las filas aparezca un modelo predictivo y en las columnas aparezcan los índices *Precisión Global, Error Global Precisión Positiva (PP), Precisión Negativa (PN), Falsos Positivos (FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN)*. Compare con todos los medelos generados en las tareas anteriores ¿Cuál de los modelos es mejor para estos datos?
4. Repita los ejercicios 1-3, pero esta vez use otro núcleo (Kernel). ¿Mejora la predicción?.
5. Repita los ejercicios 1-4, pero esta vez use 2 combinaciones diferentes de selección de 6 variables predictoras. ¿Mejora la predicción?.

- **Ejercicio 2:** [35 puntos] Esta pregunta utiliza los datos (`tumores.csv`). Se trata de un conjunto de datos de características del tumor cerebral que incluye cinco variables de primer orden y ocho de textura y cuatro parámetros de evaluación de la calidad con el nivel objetivo. La variables son: Media, Varianza, Desviación estándar, Asimetría, Kurtosis, Contraste, Energía, ASM (segundo momento angular), Entropía, Homogeneidad, Disimilitud, Correlación, Grosor, PSNR (Pico de la relación señal-ruido), SSIM (Índice de Similitud Estructurada), MSE (Mean Square Error), DC (Coeficiente de Dados) y la variable a predecir `tipo` (1 = Tumor, 0 = No-Tumor).

  Realice lo siguiente:

  1. Use Máquinas de Soporte Vectorial en **Python** para generar un modelo predictivo para la tabla `tumores.csv` usando el 70 % de los datos para la tabla aprendizaje y un 30 % para la tabla testing.
  2. Usando la función programada en el ejercicio 1 de la tarea anterior, los datos `tumores.csv` y los modelos generados arriba construya un `DataFrame` de manera que en cada una de las filas aparezca un modelo predictivo y en las columnas aparezcan los índices *Precisión Global, Error Global Precisión Positiva (PP), Precisión Negativa (PN), Falsos Positivos*

(FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN). Compare los resultados con todos los modelos generados en las tareas antariores ¿Cuál de los modelos es mejor para estos datos?

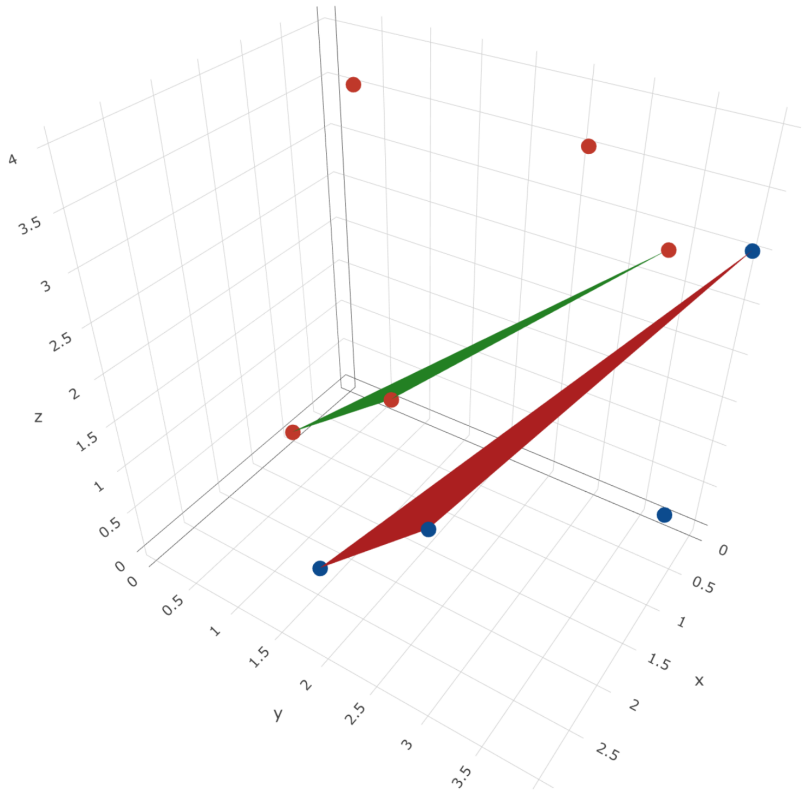
3. Repita los ejercicios 1-2, vez use otro núcleo (Kernel). ¿Mejora la predicción?

- **Pregunta 3:** [30 puntos] Suponga que se tiene la siguiente tabla de datos:

X	Y	Z	Clase
1	0	1	Rojo
1	0	2	Rojo
1	1	2	Rojo
3	1	4	Rojo
1	1	3	Rojo
3	2	3	Azul
1	2	1	Azul
3	2	1	Azul
1	1	0	Azul

Realice lo siguiente:

1. Investigue sobre paquete en Python que permiten realizar graficación en 3D. Escoja en que mejor considere para resolver los siguientes ejercicios.
2. Dibuje con colores los puntos de ambas clases en $\mathbb{R}^3$. Debería verse algo similar a lo siguiente:



3. Dibuje el hiperplano óptimo de separación e indique la ecuación de dicho hiperplano de la forma $ax+by+cz+d=0$[1]. Nota: Se debe observar con detenimiento los puntos de ambas clases para encontrar los vectores de soporte de cada margen y trazar con estos puntos los hiperplanos de los márgenes luego trazar el hiperplano de soporte justo en el centro.
4. Escriba la regla de clasificación para el clasificador con margen máximo. Debe ser algo como lo siguiente: $w=(w_1,w_2,w_3)$ se clasifica como *Rojo* si $ax+by+cz+d>0$ y otro caso se clasifica como *Azul*.
5. Indique el margen para el hiperplano óptimo y los vectores de soporte.
6. Explique por qué un ligero movimiento de la octava observación no afectaría el hiperplano de margen máximo.
7. Dibuje un hiperplano que no es el hiperplano óptimo de separación y proporcione la ecuación para este hiperplano.
8. Dibuje un hiperplano de separación pero que no es el hiperplano óptimo de separación, y escriba la ecuación correspondiente.
9. Dibuje una observación adicional de manera que las dos clases ya no sean separables por un hiperplano.

[1]En el Aula Virtual puede encontrar una presentación sobre planos en $\mathbb{R}^3$.