Aprendizaje Automático y Análisis de Datos
Taller Calificable 2, Segundo Parcial

Instrucciones:
Este taller, que corresponde al segundo parcial del curso, debe ser resuelto en grupos de dos personas y enviado por el botón de entrega que está al lado de estas instrucciones en la página de moodle del curso a más tardar el lunes 3 de mayo a las 11:59pm. Las sustentaciones serán el martes 4 de mayo. Cada grupo deberá apuntarse en el horario que le sirva usando el archivo excel que se encuentra también al lado de estas instrucciones.

Conceptos que se deben aplicar:

- Entendimiento y preparación de datos
- Métodos hold-out y cross validation para evaluación de sistemas de aprendizaje automático
- Análisis de resultados
- Entrenamiento y prueba de modelos
- Mejora de resultados

Actividades a realizar:

- Hacer tres iteraciones de trabajo sobre los datos de partida, con el propósito de incrementar tanto como sea posible el indicador F1-score.
    - Primera iteración:
        - Entendimiento y preparación de los datos
        - Entrenamiento de tres modelos de aprendizaje supervisado. Uno de ellos debe ser el perceptrón multicapa, los otros los pueden escoger teniendo en cuenta que se busca alcanzar el indicador F1-más alto posible.
        - Estimación de parámetros usando el método de grilla y las funciones disponibles en la librería sklearn.
        - Evaluación del desempeño utilizando holdout de 10 iteraciones (por cada técnica).
        - Visualización de los resultados del holdout (media y varianza por cada métrica a considerar. Mínimo accuracy, precision, recall y f1-score.)
        - Análisis de los resultados y selección de la técnica que produce mejor desempeño.
    - Segunda iteración:
        - Retomar el mejor modelo y hacerle ajustes que logren mejorar aún más su desempeño. Algunas ideas sobre qué cambiar para lograr una mejora son:
            - Con respecto a los datos: reconsiderar los atributos de entrada, el balanceo de los datos, la codificación de los datos, la normalización de valores, entre otros.

- Con respecto a los modelos: incluir otros parámetros adicionales en la estimación, cubrir rangos más amplios o más finos en el proceso de estimación.
- Otras variaciones que a ustedes se les ocurran.

■ En la segunda iteración deberán realizarse las siguientes actividades:
- Realización de los ajustes que cada equipo seleccione.
- Reentrenamiento del modelo manteniendo los valores de los parámetros estimados en la iteración uno y que no sean sujeto de modificación (es decir, no es necesario reestimar todos los parámetros sino sólo aquellos que ustedes decidan reconsiderar, si acaso seleccionan este aspecto como una mejora)
- Análisis de los resultados obtenidos y comparación con los de la primera iteración. Si no hay una mejora, deberán buscar otros cambios a realizar hasta lograr una mejora.

○ Tercera iteración:
Se debe hacer una nueva mejora al modelo obtenido en la segunda iteración, para lo cual se deben modificar variables diferentes a las modificadas en la segunda iteración. Por ejemplo, si para pasar de la primera a la segunda iteración se cambió el balanceo de los datos, para pasar de la segunda a la tercera no podrá usarse nuevamente esa opción y se deberá cambiar otra cosa diferente.

El conjunto de datos sobre el que van a trabajar está en el repositorio de UCI que ya conocen, el conjunto se llama Chronic Kidney Disease data set, y lo encuentran siguiendo este enlace: https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease.

Criterios de evaluación:

- Conocimiento y preprocesamiento de los datos: 10%
- Primera iteración: 20%
  - Estimación de parámetros
  - Entrenamiento
  - Obtención de resultados
- Segunda iteración: 20%
  - Modificaciones realizadas (es importante explicar porqué se decidió explorar esas modificaciones en particular)
  - Análisis de los resultados obtenidos
  - Mejora lograda
- Tercera iteración: 20%
  - Modificaciones realizadas (es importante explicar porqué se decidió explorar esas modificaciones en particular)
  - Análisis de los resultados obtenidos
  - Mejora lograda
- Análisis comparativo de los resultados obtenidos en las tres iteraciones y conclusiones del proceso: 20%
- Sustentación grupal: 10%