Detalles sobre el Informe a Presentar como Entrega del TP de Minería de Datos

Fecha límite de entrega:

- Antes de la primera fecha de examen de Minería de Datos (no importa si no rinden en esa fecha).
- La competencia de Kaggle finaliza el 19 de mayo de 2023.

Modalidad de entrega:

- Enviar por mail el informe en formato pdf a datamining.mim@gmail.com. En el correo deben estar copiados todos los integrantes del grupo y se debe identificar qué nombre le pusieron a su equipo en Kaggle.
- **IMPORTANTE:** junto al informe, se pide también que se entregue el código de R (o el lenguaje que hayan usado) que sustente que efectivamente lo que reportan fue hecho.

Longitud máxima:

- 6 carillas (pudiendo ser menor, siempre y cuando cubra lo que se pide).
- El código debe estar comentado y debe ser entendible para quien corrija el TP.

Contenido general y objetivo del informe:

El informe debe llevar adelante una descripción de la estrategia que utilizaron para generar predicciones. Idealmente el informe debe hacer referencia a secciones de dicho código que muestren cómo fue que llevaron adelante lo que reportan. A lo largo del informe pueden ir reportando los desafíos a los que se enfrentaron y cómo los superaron (por ej.: "*los datos eran muy voluminosos y no se podían cargar en su totalidad en memoria, de modo que optamos por usar una muestra elegida de … manera*", "*entrenar un modelo demora un tiempo considerable, de modo que no realizamos una búsqueda exhaustiva de hiperparámetros, entonces elegimos los hiperparámetros de acuerdo al criterio …*").

Secciones a incorporar:

A continuación se listan la estructura que **debe** seguir el informe:

1. Análisis exploratorio de datos. No debe ser hiper exhaustivo, pero se pide que sí se incorporen los siguientes puntos:
    a. Debería contener al menos dos figuras que muestren algún/algunos insights adquiridos y el informe los debe detallar.
    b. Debe mencionar cualquier característica de los datos que les haya llamado particularmente la atención (a modo de ejemplo, podría mencionar algunos de los siguientes puntos: valores missings en los predictores, predictores con poca varianza).

2. Selección de variables / Ingeniería de atributos probada. Detalles:
    a. Además de mencionar qué decisiones tomaron (por ej: no considerar la variable X), mencionen brevemente los motivos por el cual tomaron dicha decisión (por ej: "*optamos por no considerar la variable X por no poseer prácticamente variabilidad*").

   b. Esa sección puede presentar pruebas no hayan sido usadas en el modelo final, en cuyo caso se debe explicar por qué no se las usó y el motivo por el que creen que no funcionó. A modo de ejemplo, algunas opciones que podrían probar son:
      i. Crear atributos de fecha (ej.: hora, minuto, día de la semana, etc.).
      ii. Discretizar atributos continuos.
      iii. Hacer transformaciones de atributos numéricos (por ej. transformaciones logarítmicas).
      iv. Hacer *bin-counting* de variables categóricas que crean que pueden tener poder predictivo, pero no puedan manipular por la alta cantidad de valores distintos que poseen.

3. Proponer, justificar y utilizar un sistema de validación de modelos.
   a. Justifiquen brevemente la decisión tomada.
   b. Analicen si los valores de performance obtenidos en validación se condicen con los obtenidos en la plataforma de Kaggle sobre el conjunto de evaluación. Si no dieran los mismos valores, propongan brevemente una hipótesis del motivo por el cual se da esto.

4. Detallar el/los algoritmo/s de aprendizaje usado/s y los hiperparámetros seleccionados.
   a. Mencionar los distintos algoritmos probados.
   b. Detallar brevemente la estrategia que utilizaron para encontrar buenos hiperparámetros para el algoritmo final seleccionado.
   c. En caso de ensamblar distintos modelos, mencionar cómo fue hecho esto y si mejoró o no el resultado.

5. Detallar del total de tiempo que asignaron al trabajo práctico, cuánto dedicaron a cada uno de los ítems anteriores.