Descripción

El trabajo final será un trabajo a realizar en grupos de tres personas. Para formar los grupos, debéis apuntaros en este libro Excel en el que se deben especificar:

- Los 3 componentes del grupo en la hoja llamada Grupos
    - Posteriormente, en este mismo documento se especificará el trabajo a realizar (link al problema en caso de que sea uno de los mencionados en la siguiente sección).
- En la hoja Alumnxs se debe poner el fondo en rojo de los alumnxs que formen el grupo (para que el resto de compañeras(os) sepa quién queda libre y puedan seguir formando grupos).

Problema a abordar

El trabajo final consistirá en abordar un problema de **clasificación no balanceada binario** con datos abiertos (el que más os motive).

- Si el problema que os gusta es multi-clase lo podemos convertir en un problema binario (una clase contra el resto o un conjunto de clases contra el resto).
- Si es un problema de regresión lo podemos convertir en un problema de clasificación no balanceado binario. Para ello, debéis establecer un umbral de tal forma que tras umbralizar queden dos grupos (clases) con una distribución no balanceada.

Existen multitud de fuentes donde pueden encontrarse este tipo de problemas, desde repositorios de conjuntos de datos hasta páginas web de competiciones de ciencia de datos:

- UCI Machine Learning repository
- KEEL dataset respository
- Kaggle Datasets
- ...

Para que se pueda analizar el efecto de las diferentes técnicas de preparación de datos, el dataset debe tener un **mínimo de 10 variables de entrada** (teniendo tanto variables numéricas como categóricas). El problema tendrá más juego si:

- Las variables numéricas tienen rangos diferentes.
- Las variables numéricas tienen distribuciones de datos diferentes.
- Las variables categóricas tienen diferente número de valores.
- El número de ejemplos no es muy pequeño.
- Hay valores perdidos.
- Hay variables con fechas y/o horas.

Evidentemente, es muy difícil que un problema tenga todas las características anteriores y no es necesario forzar esos problemas.

Una vez elegido el problema deberéis desarrollar la solución al mismo:

1. Se debe aplicar una métrica y una metodología de validación de modelos apropiada.
2. Se deben analizar las variables de entrada y su relación con la variable a predecir.
3. Se debe analizar el impacto de las diferentes técnicas de preparación de datos.

4. Se puede analizar el orden de aplicación de las técnicas de preparación de datos.
5. Se analizará el impacto de la técnica de predicción (KNN o árboles de decisión) así como la importancia de los umbrales de decisión elegidos.
6. Se aplicarán técnicas de interpretabilidad de modelos.

Entregables

El trabajo a entregar consistirá en

- **Notebook** ejecutable (**y ejecutado**) para reproducir todos los resultados:
    - El Notebook debe describir el trabajo desarrollado.
    - Debe permitir ejecutar todas las pruebas realizadas.
    - Debe de estar adecuadamente estructurado y ser fácil de seguir (comentarios o celdas con la explicación de qué se hace en cada parte del Notebook).
- Los **datos** necesarios para la ejecución del Notebook (si ocupan mucho un enlace a algún espacio compartido tipo Google Drive, OneDrive o Dropbox para descargarlos).
- **Memoria** donde describir el problema, así como la métrica de rendimiento y la metodología de validación de modelos aplicada. También se describirán los experimentos realizados, se mostrarán los resultados obtenidos y se expondrán las conclusiones obtenidas.
- **Presentación** del trabajo. Se creará una presentación en PowerPoint (o equivalente) que y será expuesta ante el resto de compañeras(os)
    - Las presentaciones se realizarán el **16 de mayo**.
    - Deben hablar todas las personas que conformen el grupo.
    - La duración de la presentación es de **8 minutos**.
    - Las presentaciones **deben contener** (no es necesario explicar todo el trabajo realizado)
        - La explicación necesaria para entender el problema abordado.
        - La explicación de la elección de la métrica y la metodología de validación de modelos elegidas.
        - La explicación de la solución obtenida al problema.
        - Los resultados obtenidos y las conclusiones extraídas a partir de los mismos.
            - Análisis del impacto de las técnicas aplicadas.
            - Evolución de los resultados.
    - **Se valorará**
        - La explicación del problema.
        - La explicación de la solución.
        - El análisis de los resultados obtenidos.
        - La organización de la exposición.
        - La calidad visual de la presentación.
        - El ajuste al tiempo de presentación.

Fechas

- El trabajo se realizará en las clases de prácticas desde el 14 de marzo.
    - Cuando se acabe de impartir toda la teoría, las sesiones de los lunes también serán utilizadas para realizar el trabajo final.
        - A partir del lunes 29 de abril las clases de los martes serán en el laboratorio de Arquitectura del edificio Los Pinos (mismo aula que para las prácticas).
- El trabajo tendrá dos hitos intermedios antes de la entrega final (y su presentación). En cada hito intermedio el profesor se pasará por cada grupo para realizar el seguimiento y valoración del trabajo (además de analizar posibles vías para seguir).
    - 28 de marzo (movida al **4 de abril** por la carpa): en este hito se debe tener el problema elegido, se debe haber determinado la métrica de rendimiento y la metodología de validación de modelos a utilizar. Además, se deberán haber analizado las variables de entrada y su relación con la variable a predecir. También se deberá haber creado una primera solución automatizada (ColumnTransformer, Pipeline) al problema (con KNN) que se irá mejorando conforme pasen las semanas.
    - 11 de abril (movida al **29 de abril**): en este hito se deberá haber analizado el comportamiento de las diferentes técnicas de preparación de datos que involucren tratamiento de variables (estandarización, transformación del tipo de variable, creación de nuevas variables, selección de variables) así como la detección de outliers y la imputación de valores perdidos (en caso de que haya).
- El trabajo final se entregará en MiAulario como máximo el 15 de mayo a las 23:55 y lo debe entregar solamente un componente del grupo.
    - En esta entrega, además de todo lo anterior, también se deberá haber analizado el efecto del uso de árboles de decisión, de las técnicas de detección de ruido y muestreo de datos para problemas no balanceados, así como los posibles cambios de umbrales de clasificación. También se deberá haber analizado la interpretabilidad del modelo final obtenido (el mejor modelo obtenido).

Evaluación

- Solución técnica y profundidad del trabajo: 65%
    - El 10% de esta parte se evaluará en el primer hito.
    - El 15% de esta parte se evaluará en el segundo hito.
- Contenido y claridad del Notebook / memoria: 10%
- Presentación: 25%