UNIVERSIDAD DE GRANADA

/ UGR / decsai

Departamento de Ciencias de la Computación e Inteligencia Artificial

# Sistemas Inteligentes para la Gestión en la Empresa

## Práctica 1: Pre-procesamiento de datos y clasificación binaria

Curso 2020-2021

## Objetivos y evaluación

En esta primera práctica de la asignatura Sistemas Inteligentes para la Gestión en la Empresa estudiaremos cómo realizar diversas tareas de pre-procesamiento de datos, como paso previo e imprescindible para el aprendizaje automático.

La práctica consistirá en la resolución de un problema de pre-procesamiento y aprendizaje automático. **Junto a la solución del problema (código R), se entregará una memoria explicativa de las tareas realizadas.**

La práctica se desarrollará de forma individual. La calificación constituirá el 15% de la nota final de la asignatura (1.5 puntos). Se evaluará, en este orden: (1) la aproximación al problema y la evaluación de alternativas; (2) la calidad de la memoria presentada; (3) la precisión y la exactitud obtenida por el clasificador. Se valorará especialmente la claridad en la redacción y en la presentación del trabajo realizado.

La entrega se realizará a través de la plataforma docente de PRADO, en el enlace que se habilitará al efecto.

## Descripción del problema

En esta práctica se analizarán datos del experimento ATLAS del CERN-LHC, que perseguía la identificación experimental de la partícula bosón de Higgs.

El problema consiste en predecir si un registro de evento corresponde al decaimiento de un bosón de Higgs o se trata de ruido de fondo.

Se trabajará sobre el conjunto de datos ofrecido en la competición de Kaggle Higgs Boson Machine Learning Challenge: https://www.kaggle.com/c/higgs-boson/. El conjunto de datos se puede descargar directamente desde este enlace: http://sl.ugr.es/higgs_sige. Los eventos recogidos en este conjunto de datos han sido generados de forma sintética con un simulador.

La descripción de las variables se encuentra en la sección *Data* del desafío de Kaggle. Cada evento está caracterizado por un identificador, los valores de 30 variables y la etiqueta correspondiente ('b': ruido de fondo, 's': bosón).

La descripción detallada de las variables se encuentra en el siguiente enlace: https://higgsml.lal.in2p3.fr/files/2014/04/documentation_v1.8.pdf. No obstante, únicamente hay que tener en cuenta lo siguiente:

1. Todas las variables son reales, excepto `PRI_jet_num`, que es un entero.

2. Las variables con el prefijo PRI (PRImitivas) son valores del experimento, mientras que las variables DER (DERivadas) son calculadas por los investigadores del experimento a partir de las anteriores. Se asume que no se conoce cómo se realiza este cálculo.
3. Los valores perdidos o desconocidos se codifican como -999.0.

El conjunto de datos consta de varios ficheros:

- training.csv: conjunto de entrenamiento con 250.000 eventos. Cada evento incluye, además de las columnas mencionadas, una columna adicional *weight.* Aproximadamente, columna especifica la probabilidad de que el evento simulado ocurra en la realidad. Se utiliza para entrenar y validar los clasificadores.
- test.csv: conjunto de test con 550.000 eventos. Se utiliza para realizar un envío (*submission*) a la competición, según las instrucciones de la sección *Evaluation* del desafío.
- random_submission.csv: Fichero de ejemplo con un envío a Kaggle, según las instrucciones de la sección *Evaluation* del desafío.
- HiggsBosonCompetition_AMSMetric.py: Implementación en Python de la métrica utilizada en la competición para evaluar la calidad de las soluciones.

El ejercicio se abordará como un problema de clasificación binaria, con dos posibles salidas: {b, s}. La elección de los procedimientos de selección de datos, clasificación y (pre-)procesamiento queda a criterio del estudiante.

En esta práctica no es necesario realizar un envío a Kaggle para su evaluación. La calidad de las soluciones puede estudiarse utilizando las métricas habituales: *accuracy*, AUC, etc. No obstante, se valorará positivamente la incorporación de la métrica AMS en la resolución del problema.

## Material

Se ofrece código con un análisis exploratorio muy elemental en el repositorio GitHub de la asignatura: https://github.com/jgromero/sige2021/tree/main/pr%C3%A1cticas/p1.

## Contenido de la memoria

La memoria explicará qué **tareas de pre-procesamiento** se han llevado a cabo y con qué objetivo, así como los resultados obtenidos. Se enumeran a continuación de forma no exhaustiva algunas de las tareas que pueden realizarse:

- EDA y visualización
- Transformación y limpieza de valores numéricos
- Selección de variables
  - Identificación de variables "útiles" mediante correlación
  - Análisis de componentes principales
  - Variables importantes según técnicas de clasificación
- Detección de conflictos e inconsistencias en los datos
  - Identificación y tratamiento de 'outliers'
  - Identificación e imputación de valores perdidos
  - Tratamiento del ruido
- Normalización

- Discretización
- Reducción y ampliación de datos
    - Selección de ejemplos
    - Tratamiento de clases no balanceadas

La memoria explicará qué **técnicas de clasificación se han utilizado**. Al menos se **utilizarán dos técnicas diferentes**, cuya selección deberá justificarse. También se detallará el proceso de generación de los conjuntos de entrenamiento, validación y test. Se analizarán los resultados obtenidos con las técnicas utilizadas, haciendo especial énfasis en las medidas de precisión y exhaustividad. También se discutirá el impacto del pre-procesamiento en los resultados de clasificación.

Se recomienda apoyar las explicaciones con gráficos, diagramas, etc.

## Entrega

**Dónde:** A través del enlace de PRADO habilitado para cada grupo de prácticas

**Cuándo**: 6 de abril de 2020

**Qué**: Un fichero .zip, incluyendo:

- Código en R (.R, .Rmd)
- Memoria
    - Portada: nombre, título
    - Índice
    - Contenidos (no exhaustivamente)
        - Exploración
        - Pre-procesamiento
        - Clasificación
        - Discusión de resultados
        - Conclusiones
    - Bibliografía