# Digital House - Data Science a Distancia

## Proyecto Final Integrador

Autores: Daniel Borrino, Ivan Mongi, Jessica Polakoff, Julio Tentor

## Pre-entrega (Julio 2022)

### Tema de investigación

**Predecir el número de comentarios de una publicación**

#### Fundamentación

Realizar predicciones en el ámbito de las redes sociales es una tarea importante en el análisis de comportamiento de las personas.

El contexto en el que se desarrolla y entrena uno o más modelos de aprendizaje automático permiten predecir ciertos valores y su interpretación depende justamente del contexto.

Este proyecto final integrador pretende demostrar que los autores manejamos conceptos y desarrollamos habilidades para obtener un modelo o modelos de aprendizaje automático que predigan razonablemente la comunicación entre personas en este entorno.

### Antecedentes sobre el tema

En 2012 Krisztian Buza publica "Feedback Prediction for Blogs"[1]. En su conclusión manifiesta que "... existe *margen de mejora mientras se desarrollan nuevos modelos de aprendizaje …*" y que "*... el problema no es trivial …*".

En Kaggle, plataforma de competencias en temas relacionados con ciencia de datos, se encuentra una competencia cerrada desde hace cinco años de la que no fue posible hallar más información.

En GitHub es posible hallar repositorios con excelentes propuestas que permiten contrastar modelos y métricas establecidas en ellos.

En el documento publicado, Buza presenta las métricas "Hits@10" y "AUC@10"; esta última permitirá comparar el valor de "AUC" con otros modelos de aprendizaje automático. De esta forma podremos validar la mejora en los nuevos modelos desarrollados.

---

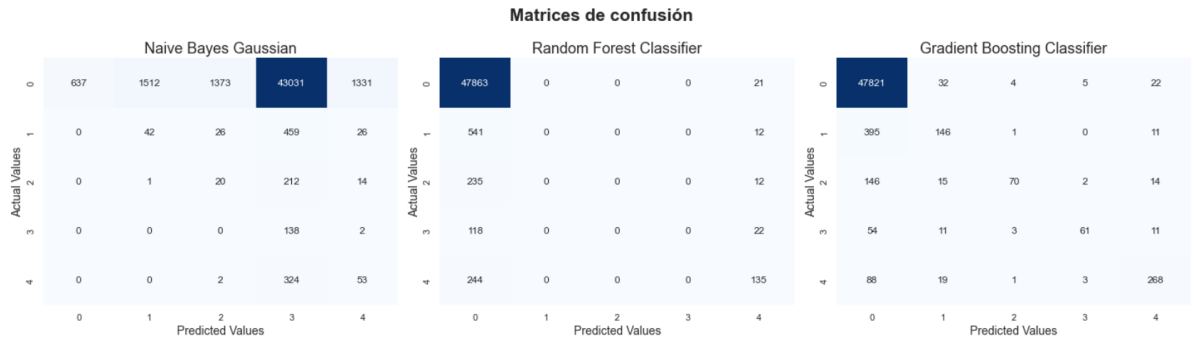
[1] Budapest University of Technology and Economics; Department of Computer Science and Information Theory http://cs.bme.hu/~buza/ https://www.linkedin.com/in/krisztian-buza-07b10a8/

En la competencia de Kaggle se puede visualizar los mejores puntajes de RMSE entre 19 y 24 mientras que en la primera aproximación al problema los autores obtuvimos valores entre 21 y 30 para la misma métrica.

## Aporte esperado

Se pretende mostrar el desarrollo de uno o más modelos de aprendizaje automático que realicen una predicción razonable del número de comentarios en las publicaciones; el trabajo final servirá como caso de estudio para analizar datos hallados en páginas web y/o social networking apps.

Con un enfoque de clasificación, estableciendo -arbitrariamente- cinco categorías se obtienen matrices de confusión que permiten suponer una tarea al menos desafiante.




Finalmente, el proyecto se concentrará en la determinación de parámetros de los modelos y detalle de características más significativas en la clasificación.

## Disponibilidad de datos e infraestructura

El conjunto de datos se encuentra disponible en [Machine Learning Repository].

Se trata de un conjunto de datos con 280 características para más de 49.000 observaciones útiles para entrenar y más de 7.000 observaciones para testear o validar.

En principio la infraestructura necesaria sólo implica acceso a internet y una computadora (real o virtual) estándar.

## Plan de trabajo y cronograma tentativo

1. Comprender el problema, Conocer el dataset - 1ra y 2da semana
2. Establecer métricas, Entrenar modelos - 3ra, y 4ta semana
3. Elaborar Informe y presentación - 5ta semana