

**Proyecto 2 – This dataset represents a set of possible advertisements on Internet pages.**

**Taller integrado de ciencias de datos**

Yorch Sepúlveda

*Julio de 2018*

1. **Construya la estructura de un proyecto de Ciencias de Datos y defina las tareas a realizar durante cada fase de un PCD.**

* **Identificar el problema:** En la actualidad, al momento navegar en las páginas web, se puede apreciar anuncios en estas. Ahora bien, existen funciones codifican la geometría de la imagen (si está disponible), así como las frases que aparecen en la URL, la URL de la imagen y el texto alternativo, el texto de anclaje y las palabras que se encuentran cerca del texto de anclaje. La tarea es predecir si una imagen es publicitaria ("publicitaria") o no ("nonad").
* **Fuentes de datos del instrumento:**

(a) Creador y donante: Nicholas Kushmerick <nick@ucd.ie>

(c) Generado: abril-julio de 1998

La información se obtuvo de N. Kushmerick (1999). Esta página hace referencia de "Aprender a eliminar publicidad en Internet", con el 3º Agentes autónomos de Int Conf. Disponible en [www.cs.ucd.ie/staff/nick/research/download/kushmerick-aa99.ps.gz](http://www.cs.ucd.ie/staff/nick/research/download/kushmerick-aa99.ps.gz). Este agente tiene una precisión mayor al 97% usando reglas c4.5 para predecir su una imagen es un anuncio.

El algoritmo C4.5 es una versión mejorada del ID3. Permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas en función de un umbral. Los arboles son menos frondosos debido a que cada hoja no cubre una clase en particular sino una distribución de clases.

Este algoritmo genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de profundidad-primero (Depth-first). Antes de cada partición de datos, el algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información. Para cada atributo discreto, se considera una prueba con n resultados, siendo n el numero de valores posibles que puede tomar el atributo. Para cada atributo continuo se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en los datos.

* **Recopila datos:** Ver [Kushmerick, 99] para detalles de los atributos; en formato "nombres":
* **altura: continua. | posiblemente falta**
* **ancho: continuo. | posiblemente falta**
* **aratio: continuo. | posiblemente falta**
* **local: 0,1.**

**| 457 características de términos url, cada una de las formas "url \* term1 + term2 ...";**

**| por ejemplo:**

**url \* imágenes + botones: 0,1.**

**...**

**| 495 características de términos origurl, en la misma forma; por ejemplo:**

**laberinto origurl \*: 0,1.**

**...**

**| 472 características de términos ancuros, en la misma forma; por ejemplo:**

**ancurl \* búsqueda + directa: 0,1.**

**...**

**| 111 características de términos alt, en la misma forma; por ejemplo:**

**alt \* tu: 0,1.**

**...**

**| 19 características de los términos de pie de foto**

**leyenda \* y: 0,1.**

**...**

* **Valores de atributo faltantes:** ¿cuántos por cada atributo?

Al 28% de las instancias les faltan algunos de los atributos continuos.

* **Preparar datos:** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco (N/A), datos inconsistentes o que están fuera de rango, obteniendo una estructura de datos adecuada para su posterior transformación.
  1. **Integrar**
  2. **Transformar**
  3. **Limpiar**: primero, se analizarán los datos en el estado que se encuentran en el documento, sin manipular los datos (en estado bruto), para ver como se encuentran los datos en el archivo .csv. Esto se realiza debido a que los datos NA son representados por “?” y el problema es que puede haber distintas formas de estar escrito en el doc. lo cual no hace fácil reemplazar todos estos por NA.

Una vez realizado esto, se procede a eliminar las filas que contengan valores NA y así poder tener una tabla consistente.

Debido a la magnitud de datos que existen es imposible ver caso a caso las filas que contienen NA, pero como la cantidad de datos con filas consistentes es mayor a la de filas con NA es suficiente para proceder al análisis de los datos sin estas.

* 1. **Filtrar**
  2. **agregar**
* **Modelo de construcción**
* **Evaluar el modelo**
* **Comunica los resultados**

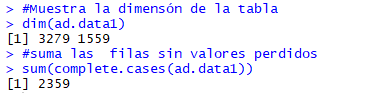
1. **Caracterice la(s) herramienta(s) de preprocesado de datos que posea el lenguaje escogido por Usted, identificando ventajas y desventajas.**
2. **Caracterice su correspondiente data set teniendo en cuenta: a. Número de instancias y de atributos, tipo de datos, valores perdidos, valores con ruido, outliers, otro que considere importante.** 
   1. **Numero de instancias y atributos: 3279 (2821 nonads, 458 anuncios)**



**Número de filas:** 3279

**Número de columnas:** 1559

* 1. **Tipos de datos:** Los tipos de datos fueron mencionados en el ítem 1 en la etapa de recopilación de datos.
  2. **Valores perdidos**



Este comando muestra todas las celdas que se encuentran vacías, pero es muy difícil de dimensionar la cantidad de datos perdidos ya que son muchos datos.



Este comando facilita la vida preguntando si existe alguna celda sin datos. Debió ser lo primero que se debería consultar al momento de analizar los datos.



Cuenta el total de celdas perdidas. Aunque aun así hay muchas perdidas de datos esas pueden ser filas completas o columnas completas. Aunque no da un indicio de que existe muchos datos perdidos y que ya debemos enfocarnos en un grupo de datos en donde se concentre la mayor cantidad de datos.

* 1. **Valores con ruido**
  2. **Outliers**
  3. **Otros**

1. **Teniendo en cuenta el objetivo del desafío planteado (Kaggle) y las características de sus datos realice la limpieza documentando cada uno de los pasos seguidos. Conserve en todo momento la estructura del proyecto creada durante la actividad**

<http://www.it.uc3m.es/jvillena/irc/practicas/03-04/18.mem.pdf>

<http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/lecturas-clasificacion/abstracts-resumir/kushmerick99learning.pdf>