Detección de Riesgo de Abandono Escolar

Descripción de un Estudio y Análisis del Algoritmo

Matilde Cabrera González
Alejandro Nuñez Perez

Bibliografía y Agradecimientos

Programa “*Data Science for Social Good*”, Universidad de Chicago: http://www.dssgfellowship.org/

THE UNIVERSITY OF CHICAGO

Entrada del blog de Ana Valdivia donde se comenta el caso:
https://valdilab.wordpress.com/2018/10/03/datos-para-el-bien-comun/

Descripción del Proyecto:
http://www.dssgfellowship.org/project/identifying-factors-driving-school-dropout-and-improving-the-impact-of-social-programs-in-el-salvador/

Código fuente del Proyecto:
https://github.com/dssg/el-salvador-mined-public

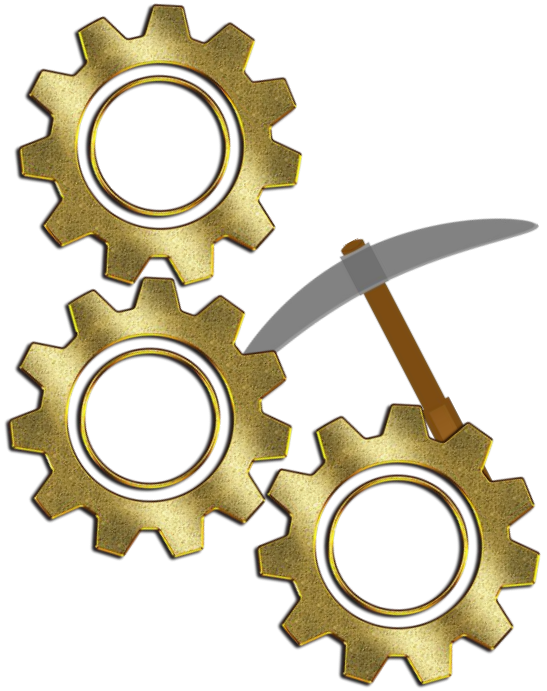
Descripción del Estudio: *Dropouts in El Salvador*

Conjunto de Datos de Orígen

- Datos gubernamentales:
  - Encuestas
  - Matrículas
- Diferentes conjuntos:
  - Colegios
  - Profesores
  - Alumnos
  - Programas Sociales

Preprocesamiento de los Datos

- Los datos de partida presentaban muchos defectos:
  - Repartidos en múltiples archivos
  - Formato irregular entre años
- Requirió un preprocesamiento extenso:
  - Combinación de diferentes fuentes
  - Etiquetado manual de columnas



Desarrollo de los Modelos

- Diferentes clasificadores:
  - Árboles de decisión
  - Regresión logística
  - Máquinas de soporte vectorial
- Características usadas:
  - Personales
  - Eventuales
  - Colegiales

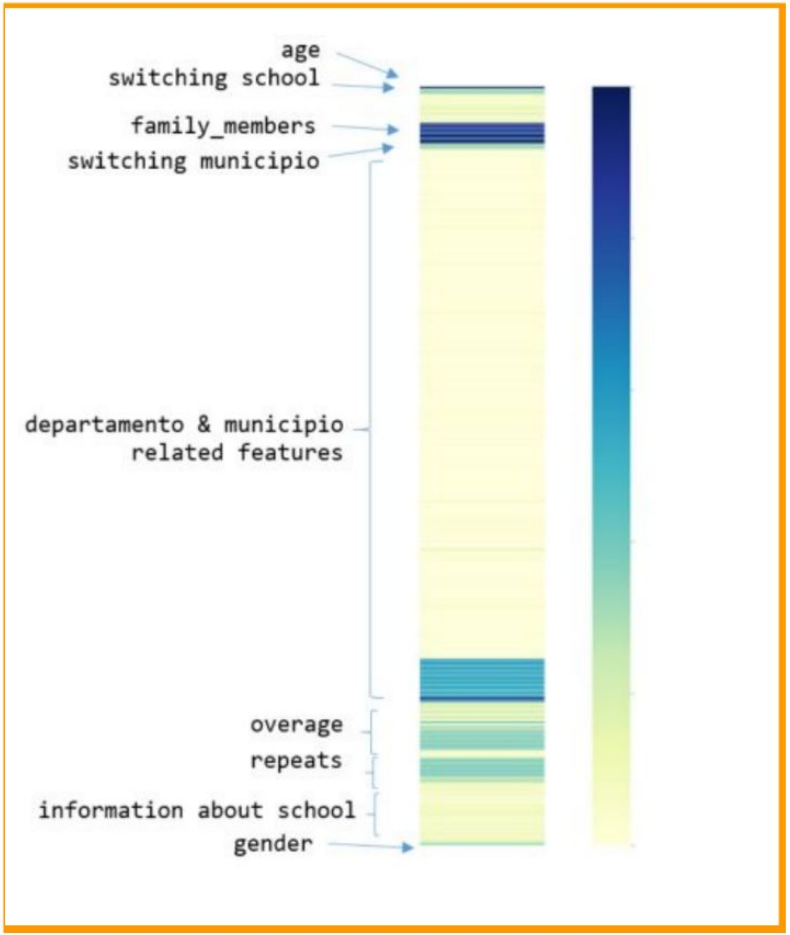
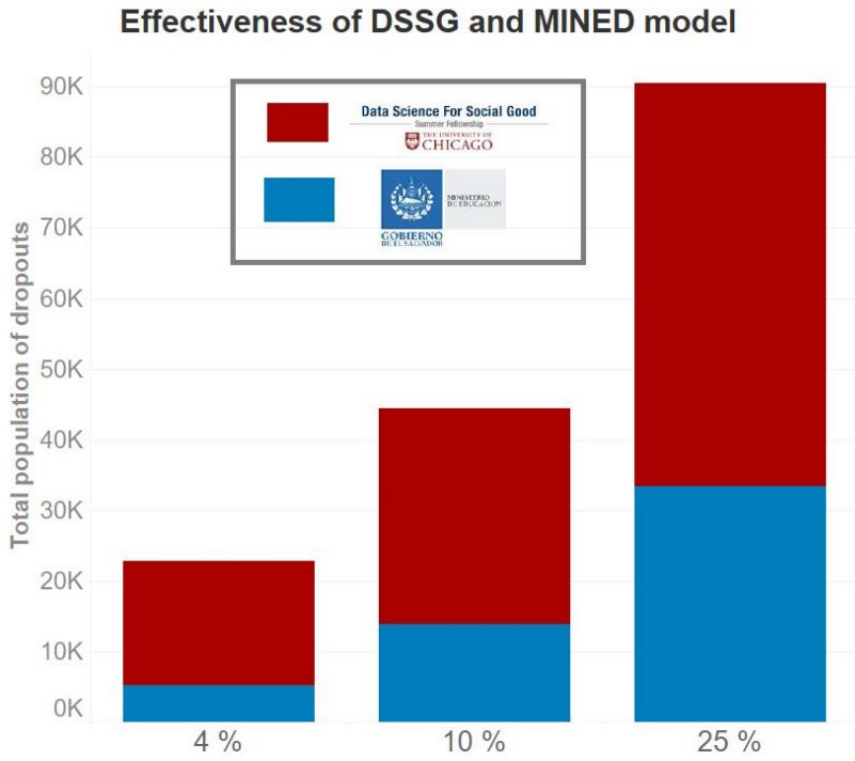


***Figure 5**: Feature importance for the best model.*

Resultados del Estudio

Con tan solo un **25%** de la población total y datos comprendidos entre los años **2011 y 2015**, se logró identificar, a casi **el triple** de lo que ya lograba identificar el gobierno de *El Salvador*.

Esto es una diferencia significante al modelo anterior, un buen resultado.

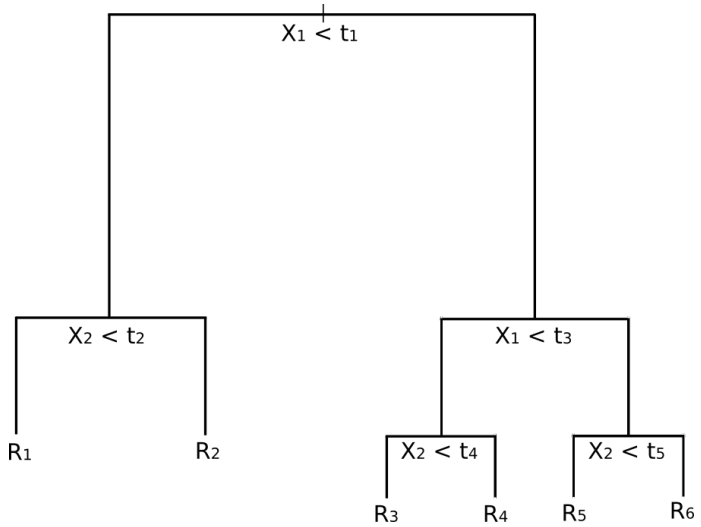


Number of students that were correctly labeled as dropouts, taking the 4%, 10% and 25% of total population.

Análisis del Algoritmo:
ExtraTrees Classifier

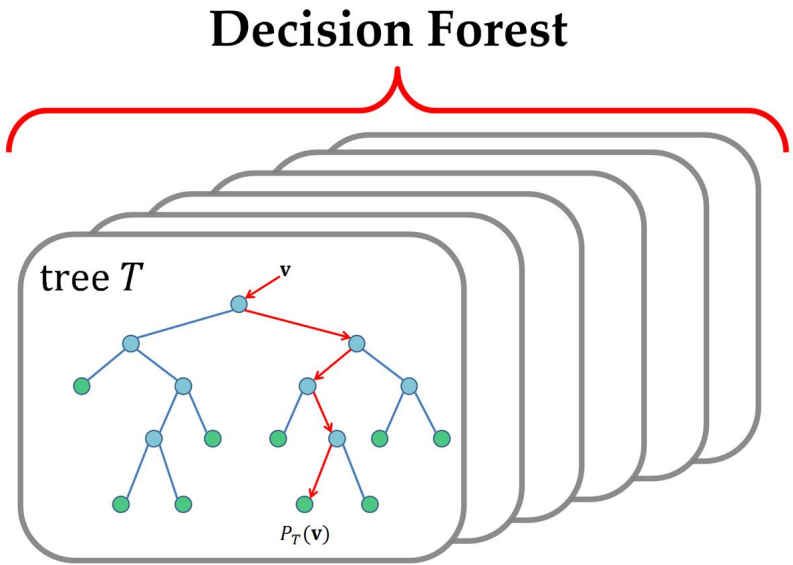
Classification And Regression Trees (CART)

- Versátil:
  - En clasificación: etiqueta moda
  - En regresión: media de los árboles
- Clasificación en regiones:
  - Condición
  - Rama
  - Etiqueta
- Muy susceptible a mal ajuste:
  - *Early stopping*
  - *Pruning*



Random Decision Forest

- Construye varios CART:
  - Conocimiento reducido
  - Elementos de *bootstrapping*
- Esquiva el mal ajuste de los CART:
  - Bajo *bias*
  - Baja *variance*
- Búsqueda del *split* óptimo:
  - Picos en la precisión
  - Alto coste computacional



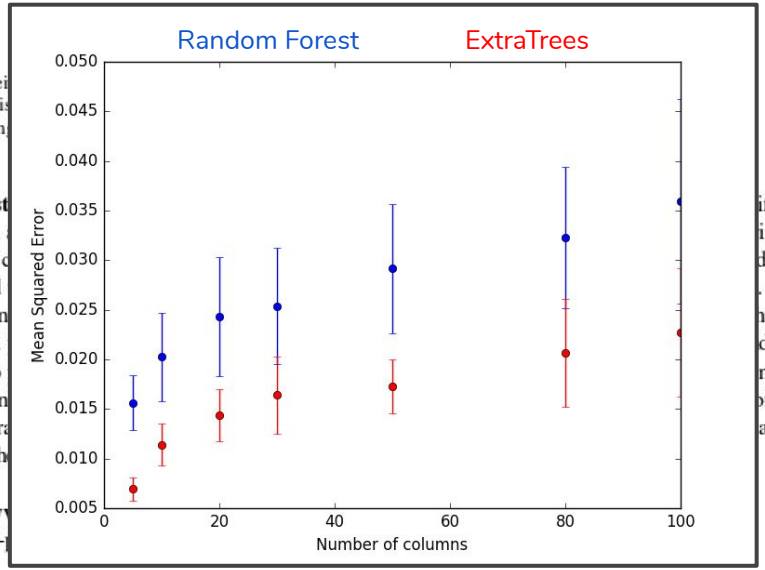
ExtraTrees

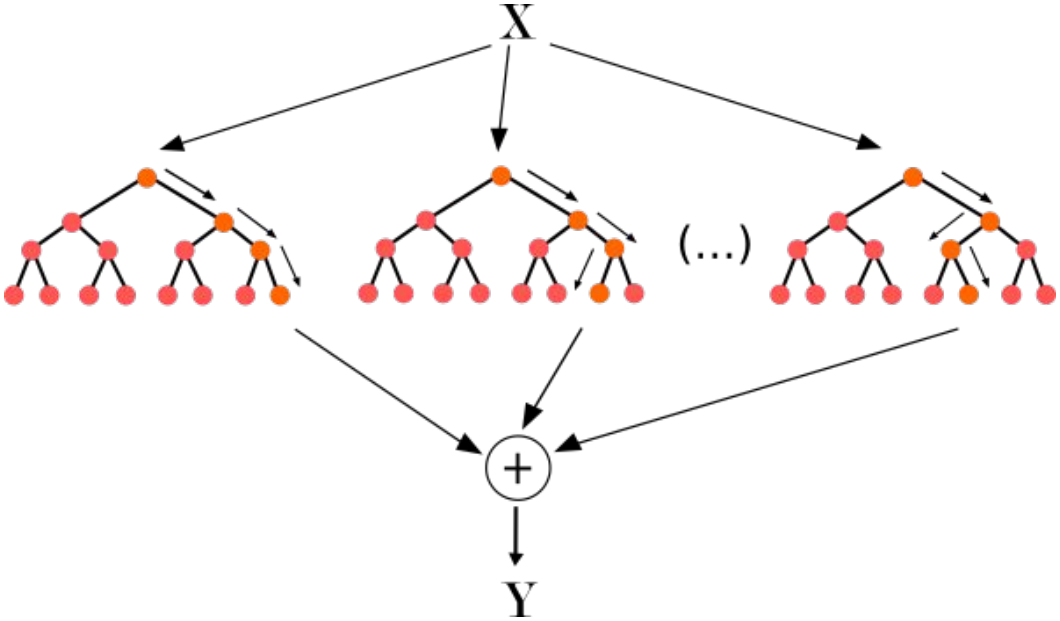
- Construye varios CART:
  - Conocimiento reducido
  - *Split* aleatorizado
- Construye sobre Random Forest:
  - Bajo *bias*
  - Bajo *variance*
- Mejores tiempos.
  - Split aleatorio, no óptimo
- Precisión suavizada.
  - Más aleatorizado

Mach Learn () :
DOI 10.1007/s10994-006-6226-1

Extremely randomized trees

Pierre Geurts · Damien Ernst · Louis Wehenkel





ExtraTrees (*Extremely Randomized Trees*)

Split_a_node(S)
Input: the local learning subset S corresponding to the node we want to split
Output: a split $[a < a_c]$ or nothing
– If **Stop_split**(S) is TRUE then return nothing.
– Otherwise select K attributes $\{a_1, \ldots, a_K\}$ among all non constant (in S) candidate attributes;
– Draw K splits $\{s_1, \ldots, s_K\}$, where $s_i = \textbf{Pick\_a\_random\_split}(S, a_i), \forall i = 1, \ldots, K$;
– Return a split s_* such that $\text{Score}(s_*, S) = \max_{i=1,\ldots,K} \text{Score}(s_i, S)$.

Pick_a_random_split(S,a)
Inputs: a subset S and an attribute a
Output: a split
– Let $a^S_{\max}$ and $a^S_{\min}$ denote the maximal and minimal value of a in S;
– Draw a random cut-point a_c uniformly in $[a^S_{\min}, a^S_{\max}]$;
– Return the split $[a < a_c]$.

Stop_split(S)
Input: a subset S
Output: a boolean
– If $|S| < n_{\min}$, then return TRUE;
– If all attributes are constant in S, then return TRUE;
– If the output is constant in S, then return TRUE;
– Otherwise, return FALSE.

ExtraTrees (*Extremely Randomized Trees*)

¿Preguntas?