Tratamiento Inteligente de Datos

Autores

Alejandro Núñez Pérez
María Matilde Cabrera González

Escuela Técnica Superior de
Ingenierías Informática y de Telecomunicación

—

Granada, Enero de 2020

Trabajo: Estudio de un Caso Práctico y Análisis de un Algoritmo.

Matilde Cabrera González

Alejandro Núñez Pérez

Palabras clave: Tratamiento Inteligente de Datos, Minería de Datos, Información, Clasificación, Random Forest

Resumen:

En este trabajo nos centramos en estudios realizados para obtener una mejora social, concretamente en el caso de la educación. El estudio a analizar, de nombre completo Identifying Factors Driving School Dropout and Improving the Impact of Social Programs in El Salvador, ha sido realizado como proyecto del programa Data Science for Social Good por la Universidad de Chicago con la intención de disminuir el abandono escolar en El Salvador, un país de Centro América.

Llevado a cabo por Dhany Tjiptarto, Yago Baltasar del Valle-Inclán Redondo y Ana Valdivia, el estudio consistió en la elaboración de un modelo de clasificación binario que permitiese predecir si un alumno está en riesgo de abandono para el siguiente año escolar, basándose en datos recopilados en los años anteriores, de fuentes como el estado de los colegios, el estado familiar del alumno y los programas sociales en vigencia.

Vamos a ver cómo se seleccionan los datos, el análisis de las propiedades y la transformación de los datos, así como la interpretación y evaluación de los mismos.

Por último vamos a estudiar la técnica Random Forest y el algoritmo Extra-Trees con más detalle pues de esta parte el estudio descrito

Project Title: Estudio de un Caso Práctico y Análisis de un Algoritmo.

Alejandro Núñez Pérez

Matilde Cabrera González

Keywords: TODO

Abstract

In this work we focus on studies carried out to obtain a social improvement, specifically in the case of education. The study to be analyzed, fully named Identifying Factors Driving School Dropout and Improving the Impact of Social Programs in El Salvador, has been carried out as a project of the Data Science for Social Good program by the University of Chicago with the intention of reducing school dropout in El Salvador, a Central American country.

Carried out by Dhany Tjiptarto, Yago Baltasar del Valle-Inclán Redondo and Ana Valdivia, the study consisted in the development of a binary classification model to predict whether a student is at risk of dropping out for the next school year, based on data collected in previous years from sources such as the state of the schools, the student's family status and current social programs.

We will look at how the data are selected, the analysis of the properties and transformation of the data, and the interpretation and evaluation of the data.

Finally, we will study the Random Forest technique and the Extra-Trees algorithm in more detail, because from this part the study described.

Índice:

Introducción

En la actualidad, en gran parte gracias al auge de Internet, la transformación digital de servicios tradicionales en servicios Web y la proliferación de dispositivos móviles inteligentes, existe un grandísimo volumen de datos generados a diario en redes sociales, negocios en línea, entre otras.

Dicha ingente cantidad de datos a menudo no se encuentra estructurada de una forma que nos permita utilizarla directamente. Esto lleva a muchos grupos, interesados en la optimización de procesos, a intentar extraer información a partir de datos desestructurados con intención de predecir un resultado, una salida a partir de unos entradas. Ejemplos de esto incluye la búsqueda patrones de crecimiento y decrecimiento en bolsa, o la clasificación de comentarios en redes sociales en categorías.

Las técnicas de extracción de conocimiento de grandes volúmenes de datos se aplican con intención de obtener información relevante a un dominio, la cual puede después usarse en diferentes ámbitos, como optimización a un proceso, como estudio de mercado, entre otros.

Este proceso de extraer información de datos potencialmente desestructurados se conoce como Minería de Datos, aunque este nombre frecuentemente se utiliza concretamente sólo para la extracción bruta, o bajo la nomenclatura de Tratamiento Inteligente de Datos.

Tratamiento Inteligente de Datos

Un proceso completo de Minería de Datos consta de diferentes etapas, de forma ordenada:

1. Obtención de los Datos
2. Análisis Exploratorio
3. Preprocesamiento y Limpieza de Datos
4. Selección y Aplicación de la Técnica de Extracción
5. Construcción del Modelo y Extracción de Conocimiento
6. Interpretación y Evaluación de Resultados

Agrupación (Clustering)

Esta técnica consiste en la creación de *clusters*, conjuntos de objetos que se asemejen entre ellos. Un *cluster* puede considerarse compuesto por más de un *subcluster*, dando un raciocinio jerárquico a la estructura de la información.

Entre esas agrupaciones, podemos distinguir la pertenencia de dos formas: certera, donde un elemento puede pertenecer a un cluster o no, y difusa (fuzzy), donde un elemento tiene un grado de pertenencia a un conjunto.

Además, podemos también definir la granularidad del *clustering*, es decir, si todo objeto pertenece a un único *cluster* o a varios, si debe o no pertenecer al menos a uno, y si existe jerarquía donde un objeto pertenece a un *cluster* hijo de un *supercluster*.

Aunque es difícil definir concretamente qué es el clustering, principalmente por la gran cantidad de modelos que existen y los algoritmos que se pueden emplear con estos modelos. Entre los modelos más comunes, podemos encontrar…

- Modelos de Conectividad, como por ejemplo el Clustering Jerárquico, que construye modelos basados en la distancia, como pueda ser la euclídea o la manhattan.
- Modelos de Centróide, como por ejemplo el algoritmo de las K-medias, donde cada cluster se representa por una de las líneas divisorias.
- Modelos basados en Neuronas, como pueden verse las redes neuronales que se ajustan de manera automática para encasillar cada objeto en un grupo.

Regresión

La regresión consiste en la aproximación de una función que aproximen entradas similares a salidas similares. Generalmente usado con salidas numéricas, puede usarse para considerar el factor de pertenencia a un conjunto y, por tanto, se asemeja al *clustering* donde existe una función por grupo a definir.

Se utiliza de manera usual para aproximar resultados de datos de los que no disponemos, pero que se asemejan a otros que sí que podemos utilizar.

Clasificación

La clasificación es la técnica consistente en la asignación de etiquetas a cada elemento de un conjunto de datos. Esta etiqueta, también llamada variable dependiente, debe estar relacionada con el resto de características, a su vez llamadas independientes.

Se considera binaria cuando el número de etiquetas es de dos, o múltiple cuando es superior a dos. Aunque parezca una distinción trivial, implica multitud de cambios,

como por ejemplo a la similitud de técnicas utilizadas, como la regresión en la clasificación binaria, o el clustering en clasificación múltiple.

Es una técnica de uso frecuente, por virtud de la utilidad que proporciona en las empresas, como pueda ser la exploración de mercado o la toma de decisiones.

Caso de Estudio: *Dropouts in El Salvador*

El Tratamiento Inteligente de Datos también entra de forma directa en cuestiones de moralidad, al ser, quizás, los datos de origen de carácter íntimo y personal, o al aplicarlo con una finalidad de dudosa moralidad.

En este trabajo nos centramos en estudios realizados para obtener una mejora social, concretamente en el caso de la educación. El estudio a analizar, de nombre completo *Identifying Factors Driving School Dropout and Improving the Impact of Social Programs in El Salvador*, ha sido realizado como proyecto del programa *Data Science for Social Good* por la Universidad de Chicago con la intención de disminuir el abandono escolar en El Salvador, un país de Centro América.

Llevado a cabo por Dhany Tjiptarto, Yago Baltasar del Valle-Inclán Redondo y Ana Valdivia, el estudio consistió en la elaboración de un modelo de clasificación binario que permitiese predecir si un alumno está en riesgo de abandono para el siguiente año escolar, basándose en datos recopilados en los años anteriores, de fuentes como el estado de los colegios, el estado familiar del alumno y los programas sociales en vigencia.

Como previamente se ha descrito en la sección Tratamiento Inteligente de Datos, un proceso de Minería de Datos consiste en varias etapas, y este no es ninguno diferente. Aunque la fuente de datos ha sido entregada a los investigadores, estos han tenido que estudiar qué relaciones podrían ser más válidas a la hora de reducir el conjunto de datos a usar.

Selección del conjunto de datos

Al equipo se le proporcionó un gran conjunto de datos con el que realizar el estudio, que previamente había usado por las organizaciones del estado de El Salvador.

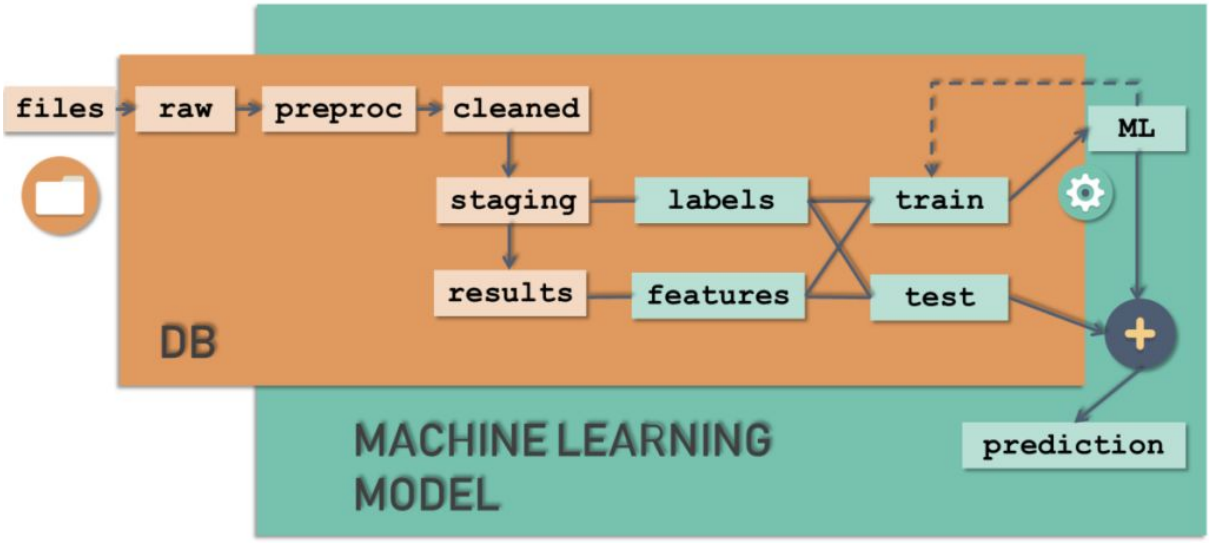
Este conjunto de datos no estaba en el mejor estado para ser usado en tareas de *Data Mining*, por el hecho de que éste no estaba muy bien estructurado, compuesto de encuestas cuyas preguntas cambiaban entre diferentes años, o datos que no tenían relevancia para el estudio.

Para discriminar qué datos se utilizarían, se debía partir desde un análisis exploratorio de los datos existentes, se diseñó un conjunto de códigos que automáticamente lee una base de datos, la estructura y la sube una nube, se realizan diferentes estrategias para solucionar las diferentes inconsistencias. Por ejemplo, para solucionar el problema de varios encabezados, creamos un código que parte un conjunto de datos en el encabezado (header) y el cuerpo (body), detectando la primera

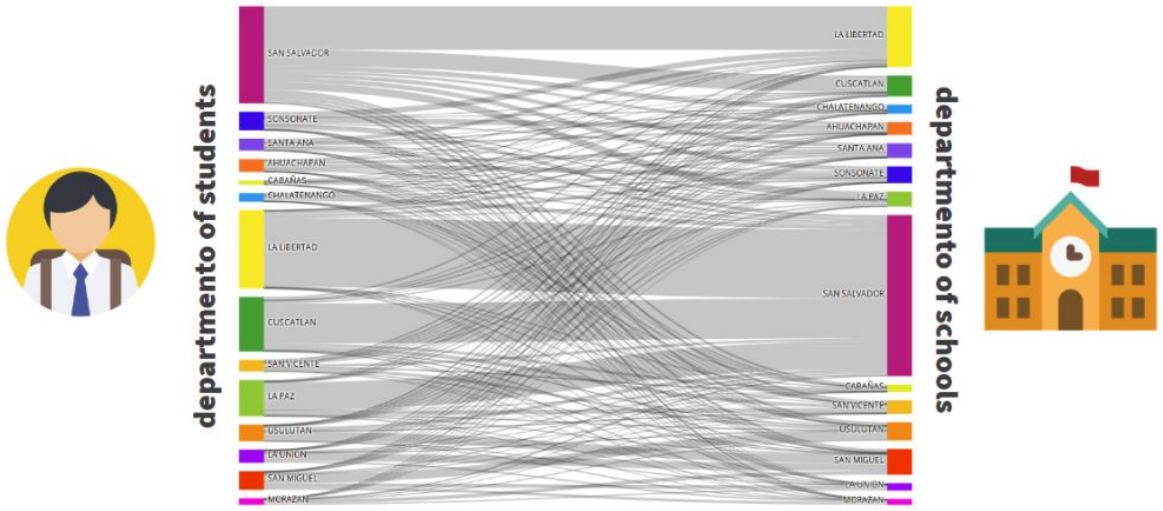
fila donde todos los campos están completos – la cual corresponde al encabezado -. Para solucionar el problema del nombre de las columnas, creamos una etiqueta única para cada una, la cuál fuera consistente en todos los datos que teníamos.

Análisis de las propiedades y transformación de los datos

Para una primera versión, se centraron en los datos de las escuelas y de los estudiantes, para ello se creó otro conjunto de códigos, el cual leía un archivo de configuración y lanzaba todos los modelos con los diferentes parámetros y variables que se encontraban en ese archivo. Una vez los modelos eran entrenados y validados, se iban almacenando todos los resultados en una base de datos, para posteriormente analizarlos.



Una vez los datos fueron limpiados, realizamos análisis descriptivos como la distribución de estudiantes que matriculados en una escuela de un Departamento diferente al de su domicilio. Visualización que muestra el porcentaje de estudiantes que no pertenecen a la comunidad de su centro escolar.



Generan varias características capturando dimensiones del estudiante y de la escuela que podrían ayudar a predecir el riesgo de que un estudiante abandone la escuela. Los grupos de características creadas (con características ilustrativas) se dan a continuación:

Promedio: si el estudiante es mayor que la edad promedio de su grado.

Repetición: cuántas veces el estudiante ha repetido un grado

Rural: si el estudiante asiste a la escuela en una zona rural

Eventos: número de diferentes escuelas a las que un estudiante asistió en un año determinado (basado en los eventos de inscripción)

Abandonos: número de veces que un estudiante ha abandonado la escuela.

Método de transporte utilizado por el estudiante para ir a la escuela

Illness: si el estudiante tenía una enfermedad registrada

Familia: número de miembros de la familia, si el padre está presente, etc.

Violencia: deserción por presencia de pandillas, deserción por trabajo, factores de drogas, explotación sexual, etc.

Escuela: número de aulas, número de computadoras de los estudiantes, fuente de electricidad, etc.

Departamento/Municipio: si el estudiante está ubicado en cada Municipio y Departamento de El Salvador

Cada una de las características mencionadas se agregan en diferentes períodos de tiempo (por ejemplo, en el último año, los últimos 3 años, los últimos 5 años, etc.) utilizando diferentes agregaciones (por ejemplo, promedio, valor máximo, suma). No

incluimos las características de violencia, escuela y departamento/municipio en el conjunto inicial de modelos entrenados en el archivo de configuración que figura a continuación debido a limitaciones de tiempo y memoria. Recomendamos incluir estas características en futuros análisis.

Selección y aplicación de la técnica de minería de datos

Uno de los aspectos más importantes relacionados con este proyecto se basan en analizar la alta probabilidad de fracaso escolar. Las organizaciones sólo pueden intervenir a un k% del total de la población, debido a un ajustado presupuesto para este tipo de problemas sociales. Es decir, solo pueden intervenir en k % estudiantes. Por ello, es muy importante que nuestro modelo sea eficiente en detectar aquellas personas que sean más vulnerables en ese k%. Así, las intervenciones podrán ser más eficaces. Como bien sabemos, las métricas clásicas (accuracy, precision, recall, AUC, etc.) evalúan el comportamiento de nuestro modelo en el 100% de la población. Es aquí donde aparecen precision@k y recall@k.

Estas medidas se basan en la misma idea que precision y recall, pero para diferentes k‘s. De esta manera, precision@k mide cuán eficaz es nuestro modelo en el k% de la población con más riesgo, y recall@k cuántas instancias estamos capturando en esa misma k. Se calcula ordenando la población en orden descendente según el score obtenido por nuestro modelo. Así, todo lo que quede por encima del corte en el k% se clasifica como positivo (abandona la escuela al año siguiente), y por debajo como negativo (no abandona la escuela al año siguiente). Este tipo de análisis son muy útiles ya que podemos analizar la eficacia de nuestro modelo en el cualquier porcentaje de la población con más riesgo. Además también, podemos evaluar el número de personas que vamos a intervenir según variaciones en el presupuesto.

Se proponen diferentes clasificadores:

Árboles de decisión son una técnica multivariada de clasificación de datos basada en diagramas de flujo, que permite organizar los casos en grupos o pronosticar valores de una variable dependiente, de escala o categórica, a partir de variables independientes o predictoras, que pueden ser a su vez de diferente naturaleza.

Regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras.

Máquinas de soporte vectorial es un método de clasificación supervisada que dado un conjunto de puntos, subconjunto de un conjunto mayor (espacio), en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo (cuya categoría desconocemos)

pertenece a una categoría o a la otra.

Vemos la importancia de las características para el mejor modelo:

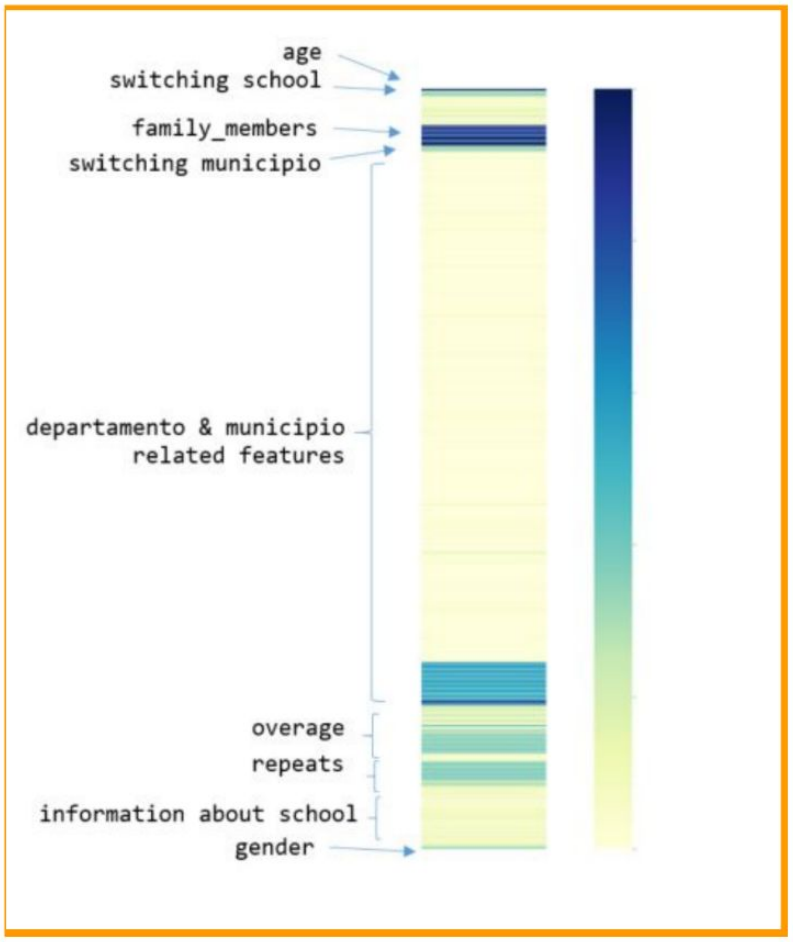


***Figure 5**: Feature importance for the best model.*

Interpretación y evaluación de datos

Seleccionan el mejor bosque aleatorio, árbol de decisión y modelos de regresión logística a escala² de todos los modelos entrenados en base a las siguientes dos métricas:

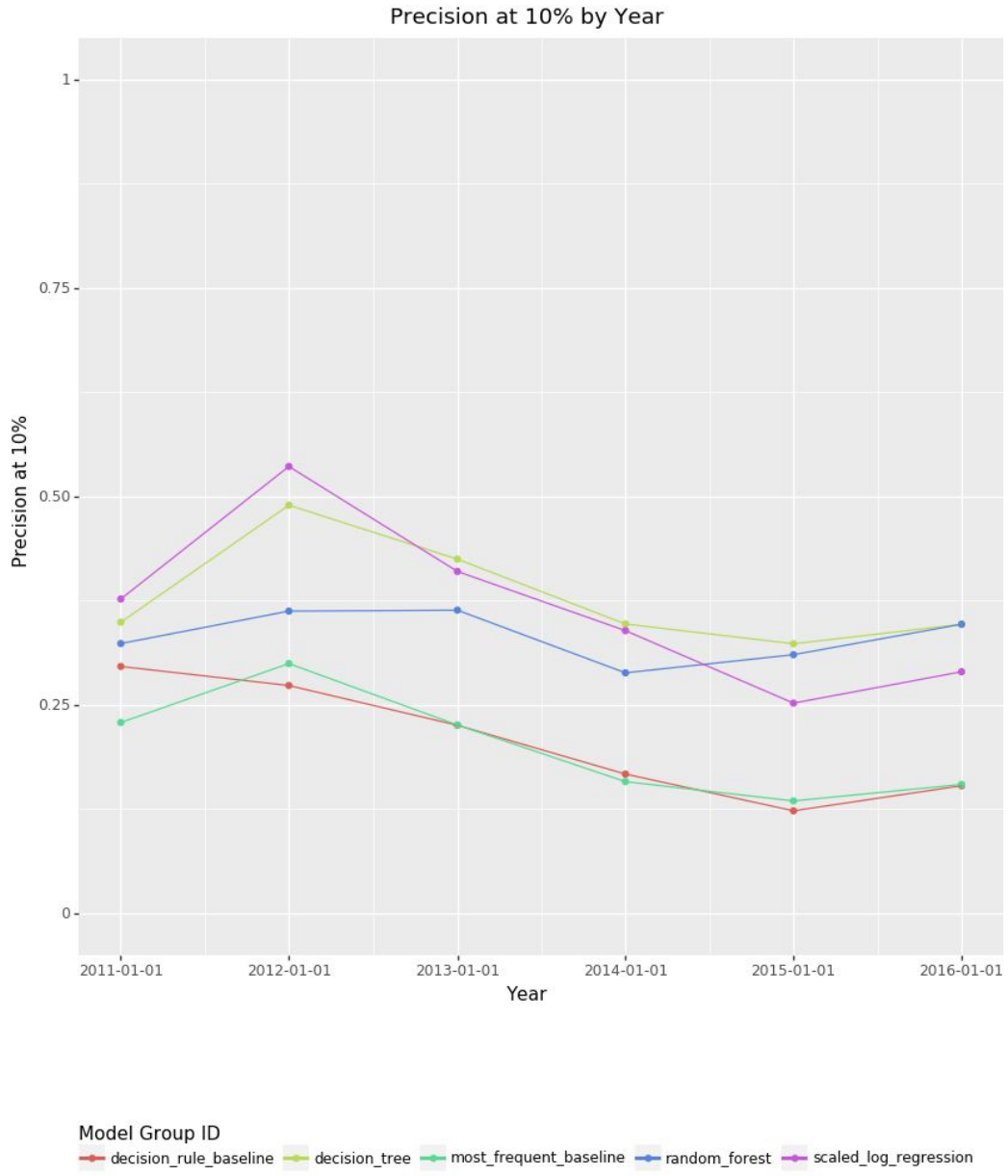
- La mejor precisión promedio al 10% en todos los años
- El mejor promedio de recuerdo en el 10% de todos los años

Comparan los resultados de sus tres modelos con dos líneas de base: 1) línea de base de la regla de decisión y 2) línea de base más frecuente. Para la línea de base de la regla de decisión, utilizan una regla de decisión que predecía que los estudiantes abandonarían la escuela al año siguiente si se cumplían las tres condiciones siguientes:

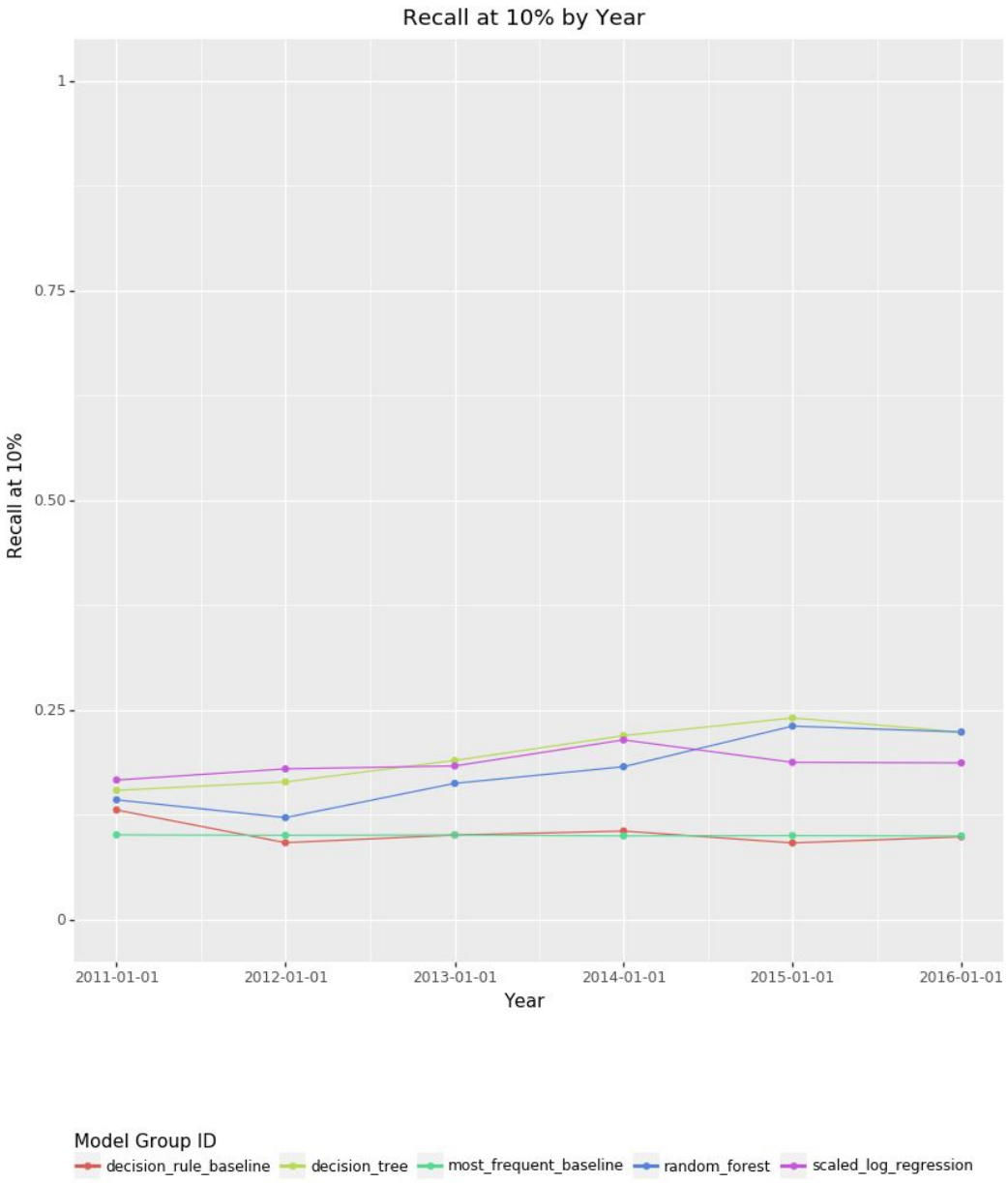
- Edad excesiva: la edad del estudiante es 2 desviaciones estándar por encima de la edad promedio en su grado

- Repetidor: el estudiante ha repetido un grado desde su última inscripción en la escuela[1]
- Rural: si el estudiante asiste a una escuela en una zona rural

La figura a continuación muestra la precisión de nuestros tres grupos de modelos principales en la métrica de precisión al 10%



La siguiente figura muestra el rendimiento en la métrica de recuerdo al 10% contra ambas líneas de base.



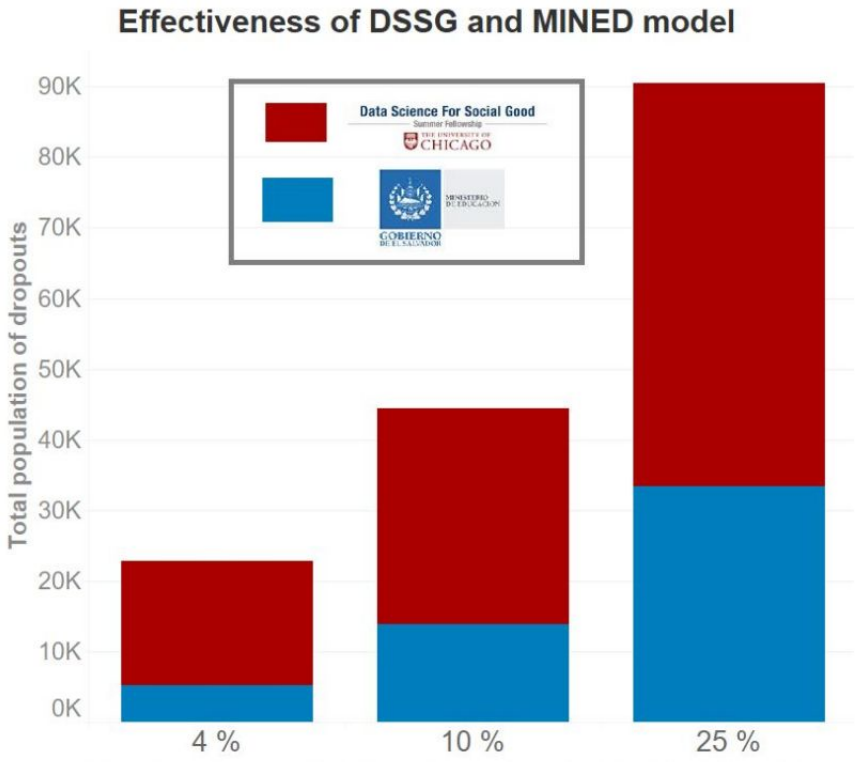
Estos gráficos iluminan algunos hallazgos clave:

En primer lugar, todos los modelos proporcionan considerables ganancias de predicción sobre la línea de base aleatoria y la línea de base de la regla de decisión. Vemos que la precisión al 10% está casi un 20% por encima de las líneas de base en cada año.

En segundo lugar, vemos que el rendimiento de todos los modelos es bastante consistente a lo largo del tiempo. El modelo de bosque aleatorio es el más consistente a lo largo del tiempo, mientras que la regresión logística a escala y el árbol de decisión alcanzan su máximo rendimiento en 2012, el año con la mayor tasa de abandono en nuestra serie temporal.

En conclusión al 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 significativa ante el modelo anterior, un buen resultado.

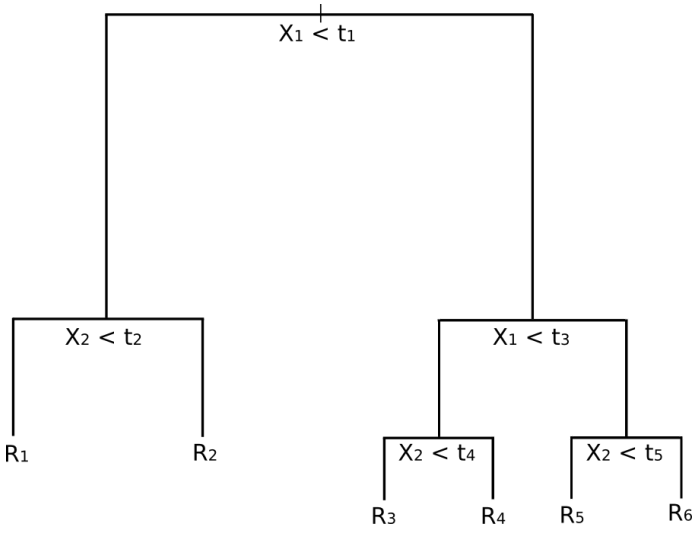


Number of students that were correctly labeled as dropouts, taking the 4%, 10% and 25% of total population.

Técnica a Estudiar: Random Forest

Random Forest, la técnica utilizada en el caso estudiado, es una técnica que mejora la precisión en la clasificación mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual.

Es capaz de realizar tanto tareas de regresión como de clasificación. También lleva a cabo métodos de reducción dimensional, trata valores perdidos, valores atípicos y otros pasos esenciales de exploración de datos. Es un tipo de método de aprendizaje por conjuntos, muy susceptible a un mal ajuste.

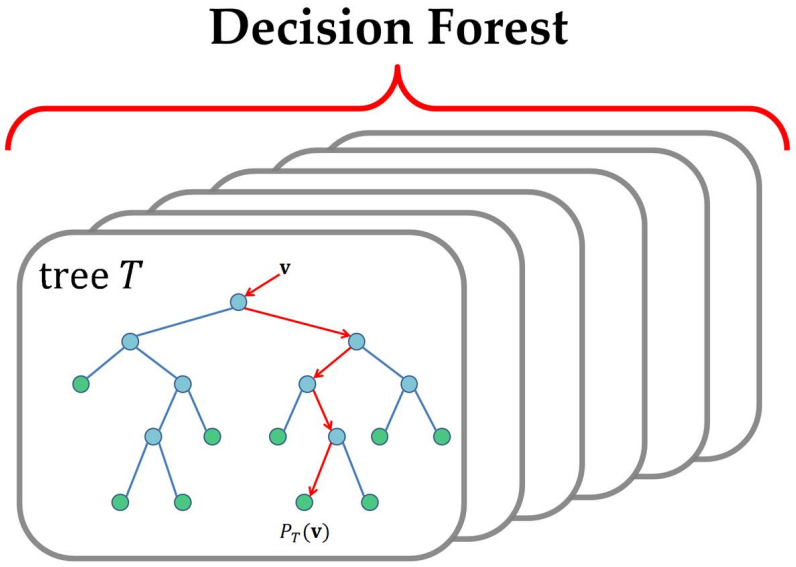


Random Forest está formado por árboles de decisión. En un árbol de decisión, se introduce una entrada en la parte superior del árbol (raíz), y va hacia abajo a medida que los datos recorren el árbol, los datos se acumulan en conjuntos más pequeños.

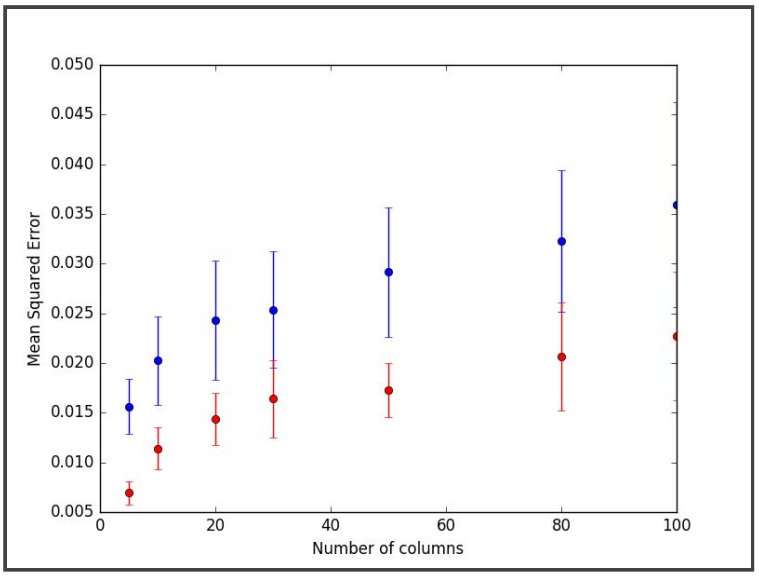
El algoritmo que forma este método funciona de la siguiente manera:

- Muestra N casos al azar con el reemplazo para crear un subconjunto de los datos. El subconjunto debe ser aproximadamente 66% del conjunto total.
- En cada nodo:
    - Para un número m, las variables predictoras m son seleccionados al azar entre todas las variables predictoras.
    - La variable de predicción que proporciona la mejor división, de acuerdo con una función objetiva, se utiliza para hacer una división binaria en ese nodo.

- En el siguiente nodo, elige otras m variables al azar entre todas las variables predictoras y hace lo mismo.



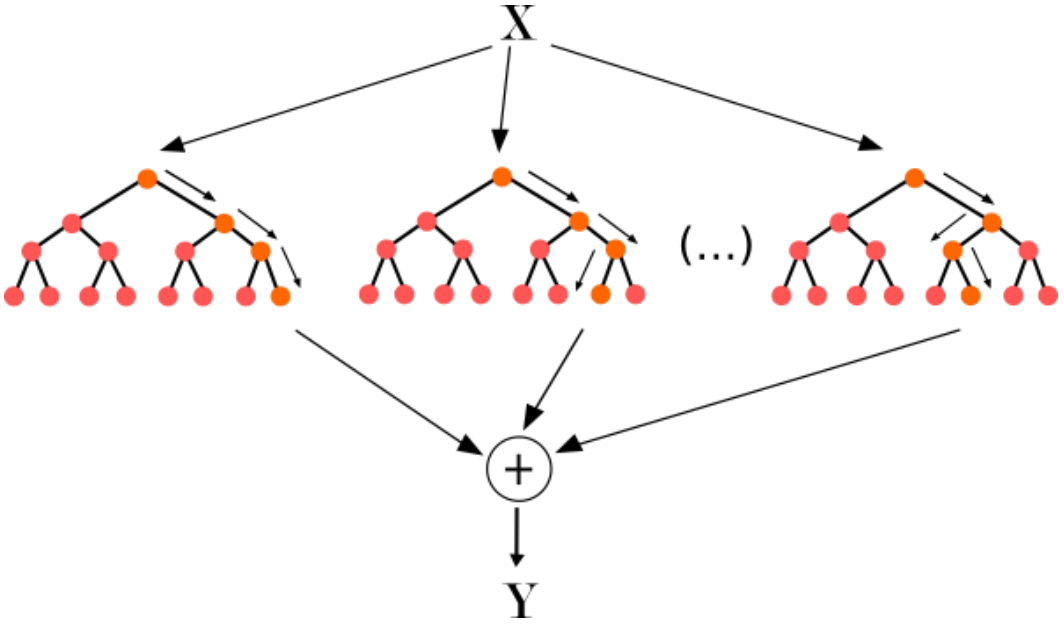
Este trabajo propone un nuevo método de conjunto basado en árboles para problemas supervisados de clasificación y regresión. Básicamente consiste en aleatorizar fuertemente tanto la opción de atributo como el punto de corte mientras se divide un nodo de árbol. En el caso extremo, construye árboles totalmente aleatorios cuyas estructuras son independientes de los valores de salida de la muestra de aprendizaje. La fuerza de la aleatorización se puede ajustar a los problemas específicos mediante la elección adecuada de un parámetro. Evaluamos la solidez de la elección predeterminada de este parámetro, y también proporcionamos información sobre cómo ajustarlo en situaciones particulares. Además de la precisión, la principal fortaleza del algoritmo resultante es la eficiencia computacional. También se proporciona un análisis de sesgo/varianza del algoritmo Extra-Trees, así como una caracterización geométrica y de kernel de los modelos inducidos.



Consideramos que el problema de aprendizaje supervisado en modo de lote estándar, y nos centramos en el aprendizaje de problemas caracterizados por posiblemente un gran número de variables de entrada numéricas y una una única variable objetivo (categórica o numérica).

Los atributos del candidato denotan todas las variables de entrada que están disponibles para un problema determinado. Usamos el término salida para referirnos a la variable objetivo que define el problema de aprendizaje supervisado. Cuando el resultado es categórico, hablamos de un problema de clasificación y cuando es numérico, hablamos de un problema de regresión. La muestra de aprendizaje de términos denota las observaciones utilizadas para construir un modelo, y el término prueba a muestrear las observaciones utilizadas para calcular su exactitud (tasa de error, o error cuadrático medio).

N se refiere al tamaño de la muestra de aprendizaje, es decir, su número de observaciones, y n se refiere a el número de atributos candidatos, es decir, la dimensionalidad del espacio de entrada.



El algoritmo Extra-Trees construye un conjunto de árboles de decisión o regresión no podados de acuerdo con el procedimiento clásico de arriba a abajo. Sus dos principales diferencias con otros métodos de conjunto de árboles son que divide los nodos eligiendo puntos de corte totalmente al azar y que utiliza toda la muestra de aprendizaje (en lugar de una réplica de bootstrap) para hacer crecer los árboles.

El procedimiento de división de los árboles extra para los atributos numéricos se indica a continuación.

Tiene dos parámetros: K, el número de atributos seleccionados aleatoriamente en cada nodo y nmin, el tamaño mínimo de la muestra para dividir un nodo. Se utiliza varias veces con la muestra de aprendizaje original (completa) para generar un modelo de conjunto (denominamos por M el número de árboles de este conjunto). Las predicciones de los árboles se agregan para obtener la predicción final, por mayoría de votos en los problemas de clasificación y media aritmética en los problemas de regresión.

Los parámetros K, nmin y M tienen diferentes efectos: K determina la fuerza del proceso de selección de atributos, nmin la fuerza f promediando el ruido de salida, y M la fuerza de la reducción de la varianza de la agregación del modelo de conjunto. Estos parámetros podrían adaptarse a las características del problema de forma manual o automática (por ejemplo, mediante validación cruzada).

Sin embargo, preferimos utilizar los ajustes predeterminados para ellos a fin de maximizar las ventajas computacionales y la autonomía del método. Se estudian estos

ajustes predeterminados en términos de robustez y suboptimalidad en diversos contextos.

Para especificar el valor del parámetro principal K, usaremos la notación ETK, donde K es sustituido por "d" para decir que se utilizan los ajustes por defecto, por "∗" para denotar los mejores resultados obtenidos sobre el rango de posibles valores de K, y por "cv" si K se ajusta por validación cruzada.

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.

Conclusiones

El tratamiento inteligente de datos puede ser una herramienta muy útil para denunciar desigualdades y ayudar a combatirlas, o como se dice en el artículo de nuestro estudio, para que el gobierno pueda repartir mejor los recursos, de forma que se puedan centrar donde más se necesita y esto tenga un impacto positivo tanto para la ciudadanía en cuestión.

Uno de los debates a la hora del desarrollo del estudio fue que se pudiera hacer un mal uso del mismo, discriminando a las personas con mayor probabilidad de abandono escolar. Para intentar evitar esto, se trasladó todas las cuestiones morales que suponen el estudio de una parte de la población a los respectivos ministerios que pidieron el desarrollo del mismo, para que así se garantice siempre un buen uso de los datos del proyecto desarrollado.

Esta es una cuestión que se debería tener en cuenta para cualquier estudio que implique a personas, ya que se trata de mejorar la vida de las personas, no de discriminarlas, aunque esto está en mano de quien maneje los datos del estudio, a veces estos estudios son manejados por grandes empresas para su propio interés, lo que entra en conflicto con lo expuesto anteriormente.

Bibliografía

Blog de Ana Valdivia:o:
https://valdilab.wordpress.com/2018/10/03/datos-para-el-bien-comun/

Descripción del proyecto desarrollado por Ana Valdivia (et al):
http://www.dssgfellowship.org/project/identifying-factors-driving-school-dropout-and-improving-the-impact-of-social-programs-in-el-salvador/

Código fuente del estudio relevante por Ana Valdivia (et al):
https://github.com/dssg/el-salvador-mined-public

Blog de Tableau donde se explica la aplicación de business intelligence para diagnosticar qué estudiantes del distrito de colegios público “Des Moines” están en riesgo de abandono:
https://www.tableau.com/solutions/customer/des-moines-public-school-district-improves-intervention-programs-predictive

Calendario escolar preliminar 2018:
http://www.mined.gob.sv/descarga/CALENDARIO_22_ENERO_2018_PRELIMINAR.pdf

Extra Trees Classifier:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html