Desarrollo de algoritmos de aprendizaje automático para la exclusión educativa

Dr. Patricio Rodríguez

prodriguez@ciae.uchile.cl

IV Taller de Gobierno Centrado en el Ciudadano
Brasilia, Brasil
Junio 29, 2017

Laboratorio de Analítica Pública®

Centro de Investigación Avanzada en Educación
Universidad de Chile

CENTRO DE INTELIGENCIA TERRITORIAL
DESIGN.LAB UAI

BID
Banco Interamericano de Desarrollo

Enap

MINISTÉRIO DO PLANEJAMENTO, DESENVOLVIMENTO E GESTÃO
BRASIL
GOVERNO FEDERAL

1 ¿Qué es el aprendizaje automático?

¿Qué es el aprendizaje automático?

Tipos de aprendizaje automático

Una subespecialidad de la Ciencia de la Computación (denominada históricamente "inteligencia artificial") que se ocupa del diseño y desarrollo de algoritmos que permiten inferir comportamientos basados en datos empíricos

El aprendizaje automático puede ser de dos tipos: supervisado y sin supervisión

Tipos de aprendizaje automático

Aprendizaje supervisado

En el **aprendizaje supervisado**:

Se debe inferir una función a partir de un conjunto de ejemplos de entrenamiento

Estos consisten en un conjunto de entradas (en forma de vector) y un conjunto de salidas que son casos exitosos (satisfacen la función)

Los casos exitosos generan una **medida de error** respecto a las predicciones que se quieren hacer

Ejemplos:

Redes neuronales, redes bayesianas, árboles de decisión, *Support Vector Machines, Random Forests*, regresión logística, *Deep learning*, entre otros.

Tipos de aprendizaje automático

Aprendizaje no supervisado

En el **aprendizaje no supervisado**:

- Los “casos exitosos” no se conocen (o no se necesitan)
- No existe retroalimentación para ajustar una función

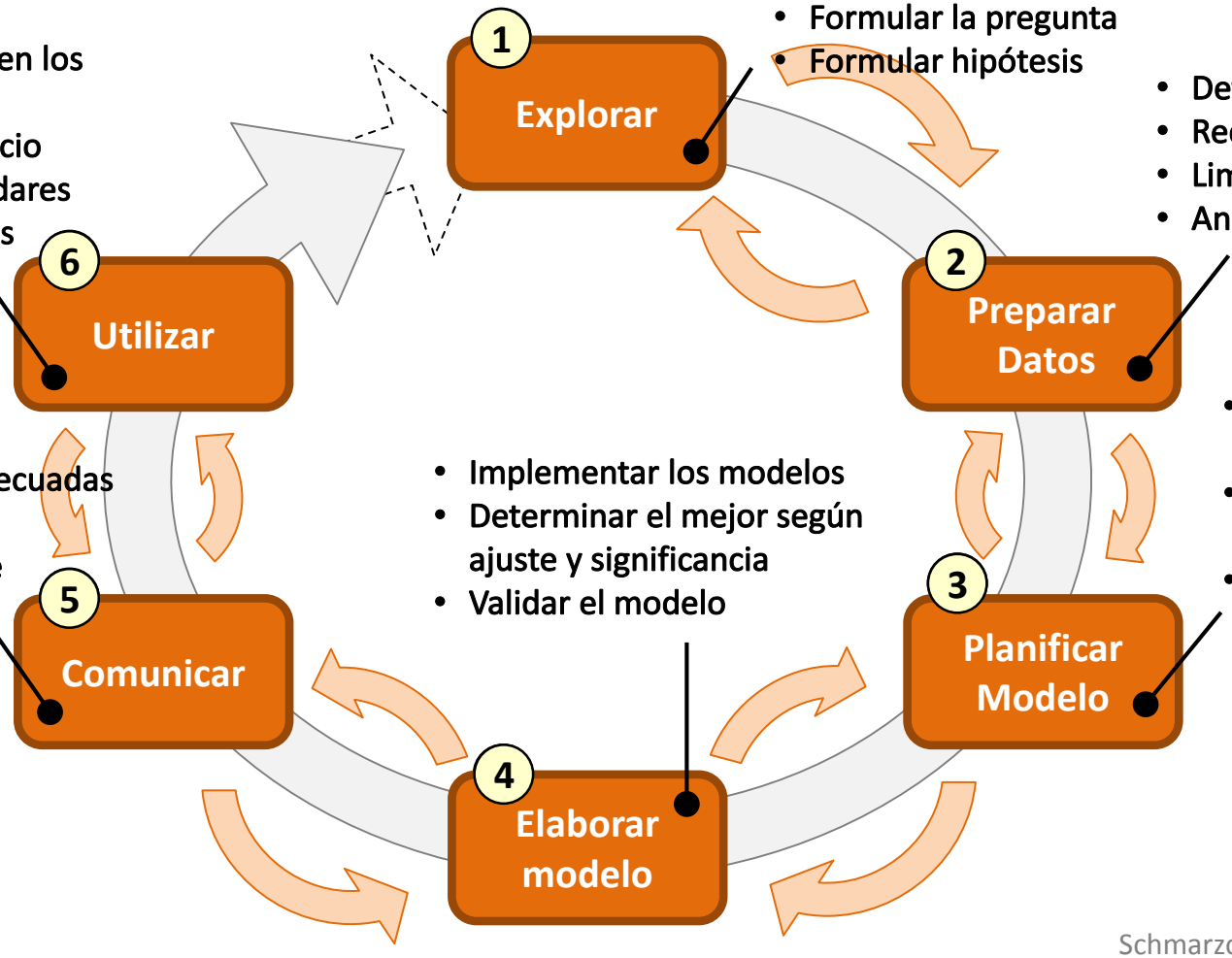
Por lo tanto, el objetivo del algoritmo es **organizar** los datos o **describir su estructura**

Ejemplos: los algoritmos de agrupamiento (*clustering*) como *k-means*

2

Ciclo de vida del análisis de datos

2 Ciclo de vida del análisis de datos



- Explorar los datos disponibles
- Formular la pregunta
- Formular hipótesis

- Determinar datos necesarios
- Recopilar datos
- Limpiar datos
- Analizar consistencia de datos

- Determinar qué variables explican y cuáles predecir
- Seleccionar posibles modelos y algoritmos a utilizar
- Definir métricas de desempeño

- Implementar los modelos
- Determinar el mejor según ajuste y significancia
- Validar el modelo

- Interpretar resultados
- Generar visualizaciones adecuadas para comunicar resultados
- Hacer recomendaciones de mejora

- Tomar decisiones basadas en los resultados
- Definir estándares de servicio
- Metas para alcanzar estándares
- Planificar y asignar recursos

Schmarzo (2012): Data Analytics Lifecycle

Ciclo de vida del análisis de datos

Ejemplo – Exclusión educativa

¿Qué es?: cuando el sistema educativo **falla** en proveer a sus alumnos servicios que produzcan aprendizaje exitoso, o cuando éstos no puedan seguir avanzando en sus cursos

Resultados observables: deserción, abandono o repitencia

¿Cuándo?: durante la transición desde la educación básica a la educación media y alrededor de los 15 años (en Chile)

Ciclo de vida del análisis de datos

Ejemplo – Exclusión educativa

Cuando desertan, los estudiantes quedan con un nivel educacional **menor** que el que podrían alcanzar. Esto produce:

Mayor ↑

- Brecha de ingresos entre desertores y graduados
- Tasa de desempleo e inactividad
- Tasa de criminalidad
- Gasto fiscal en sistema de salud y asistencia social
- Nivel de pobreza

Menor ↓

- Ingreso fiscal por impuestos
- Productividad
- Crecimiento económico
- Expectativa de vida
- Cohesión social y participación ciudadana

2 Ciclo de vida del análisis de datos

Ejemplo – Exclusión educativa

Desertores 2011	Mujeres	Hombres	Total
15-19 años	74.571	68.347	142.918

9,9% matrícula 15 — 19 años
3,84% matrícula total

Costos asociados con deserción	Impacto Deserción Escolar	Costo social de por vida por una cohorte que deserta (Millones USD 2015)			Costo Social (% del PIB)	Beneficio social por % reducción
		Mujeres	Hombres	Total		↓1%
Ingresos (sueldos)	Mayor brecha en ingresos de trabajo	619,975	1.996,518	**2.616.493**	1,2%	26.165
Impuestos (IVA)	Menor ingreso fiscal [illegible]	[illegible]	[illegible]	[illegible]	0,3%	14.994
Empleo	Menor empleo	-	-	**597.792**	0,7%	5.978
Asistencia Social	Mayor gasto público a subsidios monetarios	324,354	310,277	**634.631**	0,3%	6.346
Salud	Mayor gasto público en salud (FONASA)	-	-	**151.731**	0,1%	1.517
Total		**2.096.004**	**2.654.541**	**5.500.069**	**2,6%**	**55.001**

2,6% del PIB de Chile

Cifras estimadas en 2011 expresadas en USD 2015

Ciclo de vida del análisis de datos

1 – Explorar

Ejemplo – Exclusión educativa

La deserción y el abandono que requieren ser abordados por la brecha de oportunidades que generan en un segmento de la población

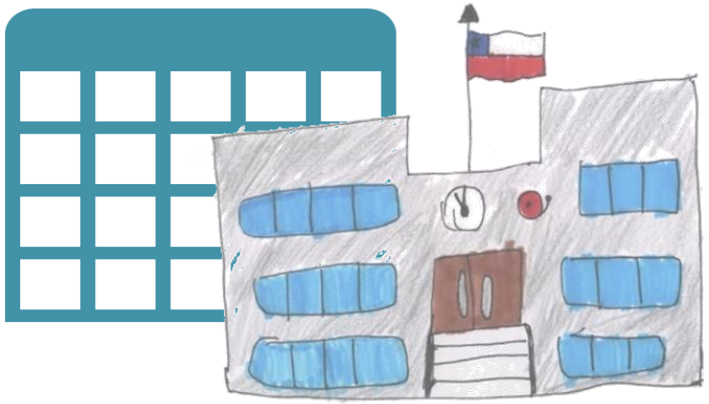
Un programa de intervención para detectarla precozmente para evitarla tiene muy altas posibilidades de ser rentable según las estimaciones alcanzadas:

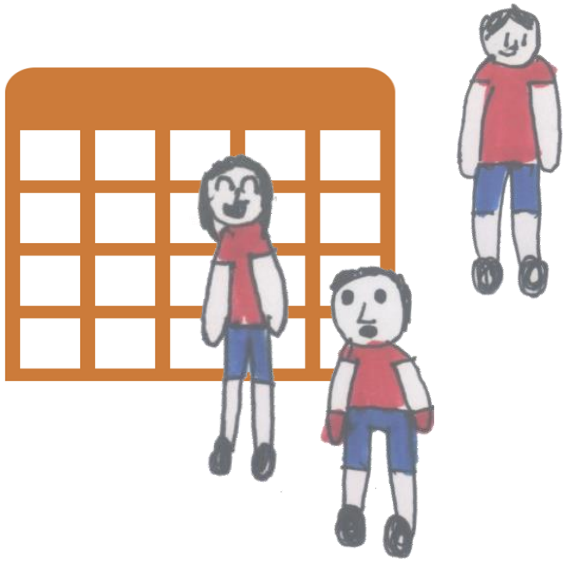
¿Cómo podemos generar un mecanismo para detectar anticipadamente si un(a) joven podría desertar?

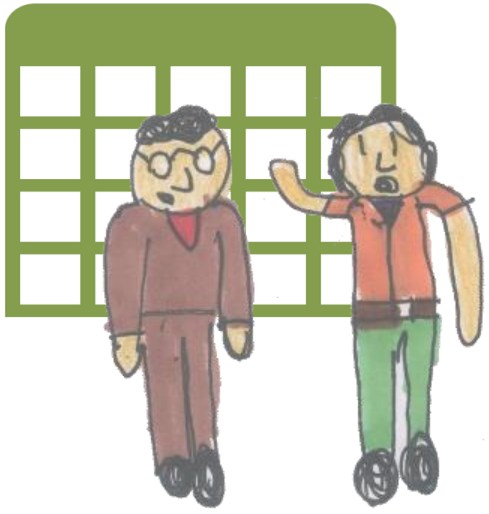
Propuesta: modelo predictivo para la deserción escolar

Ciclo de vida del análisis de datos

2 – Preparar datos







2

Ciclo de vida del análisis de datos

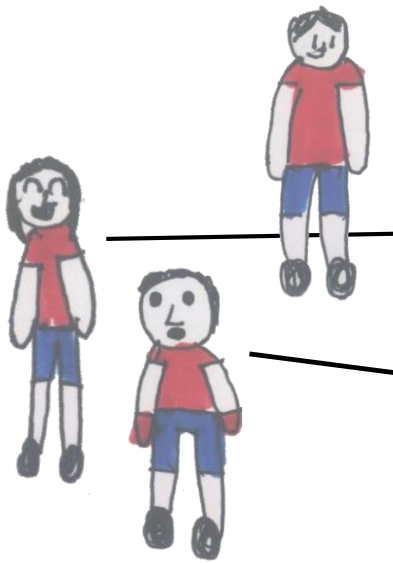
3 – Planificar el modelo

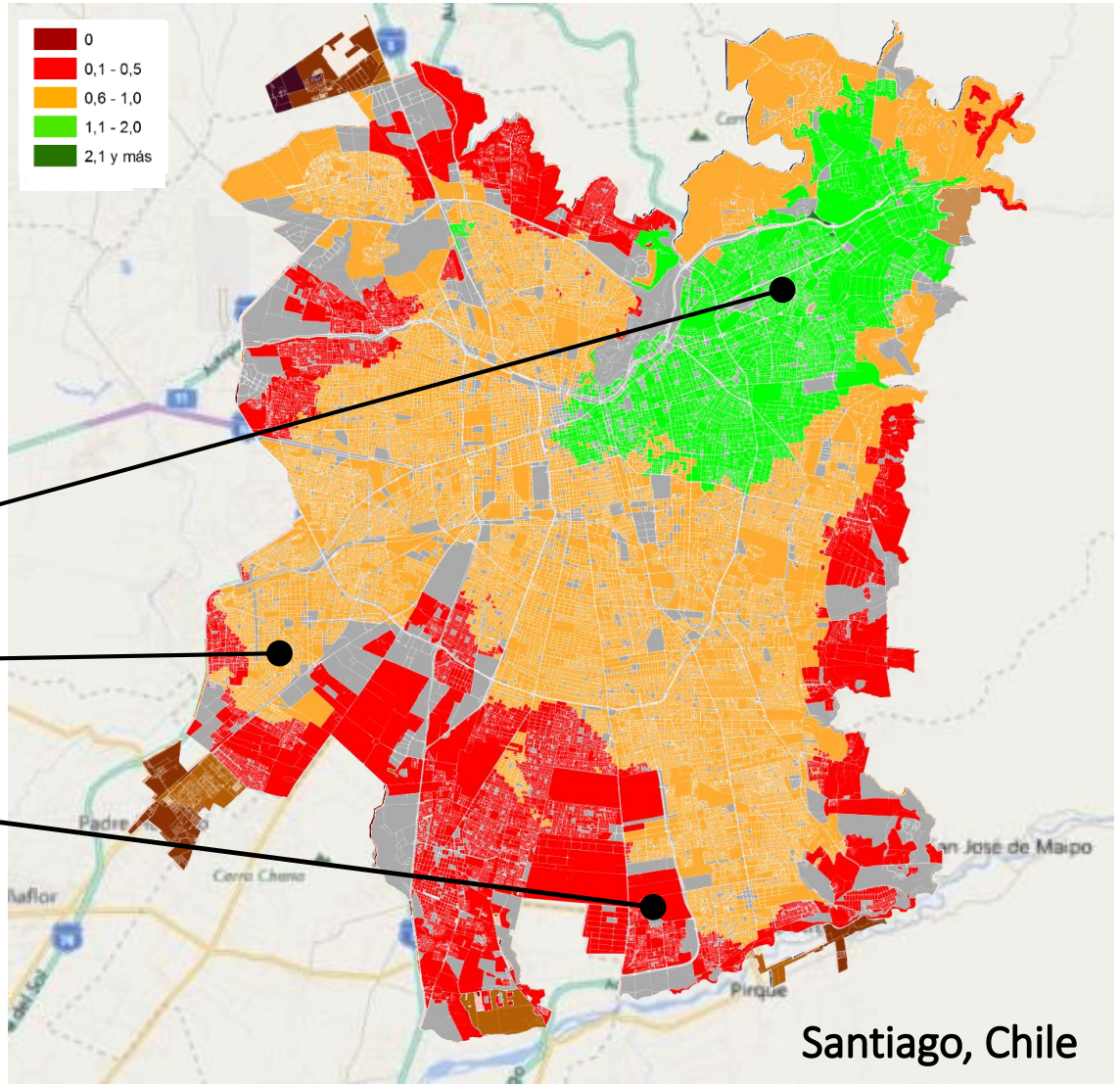
Donde viven

Que nivel de ingreso familiar tienen

Distancia promedio a servicios

Acceso a colegios de alto estándar

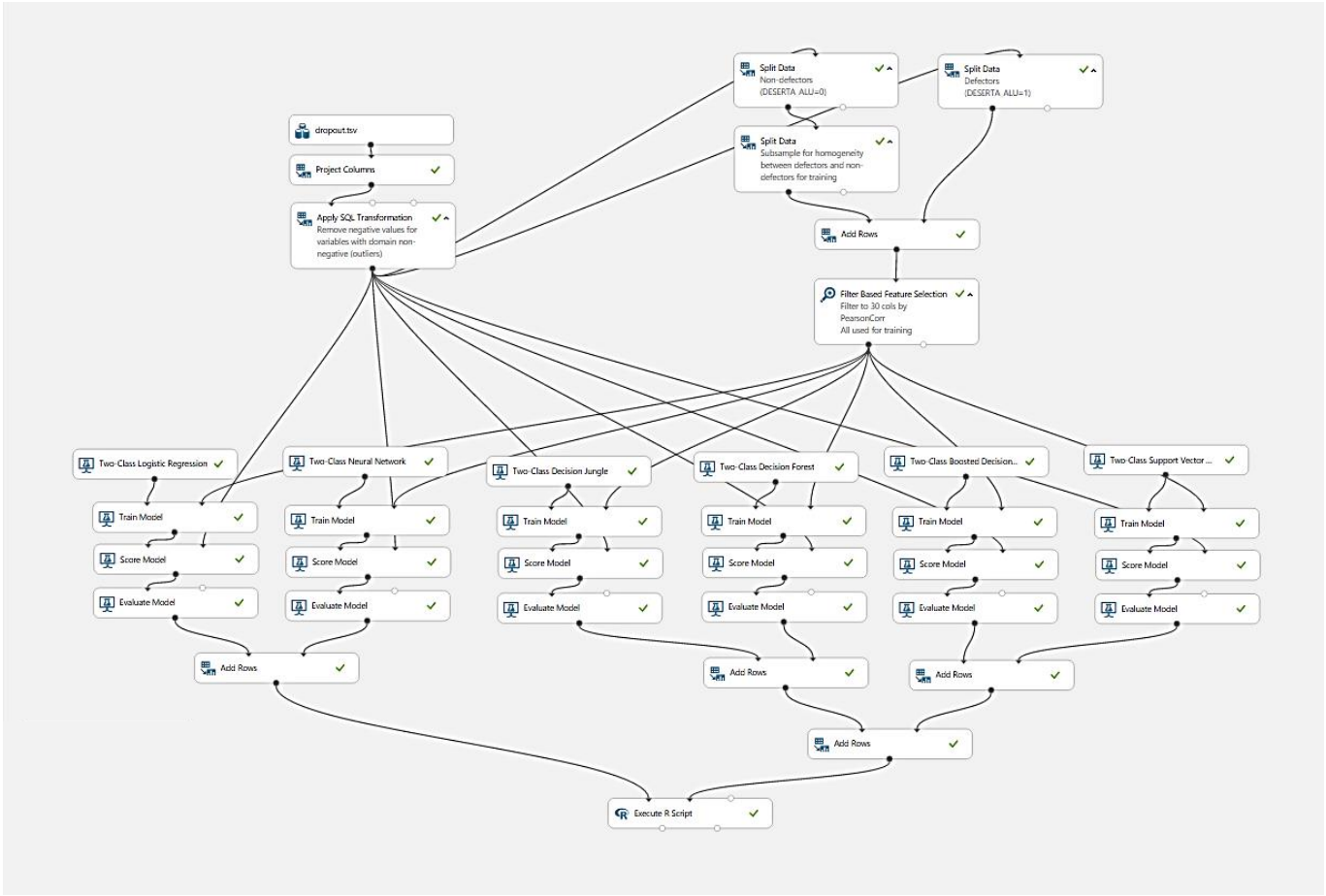




2

Ciclo de vida del análisis de datos

4 – Elaborar el modelo



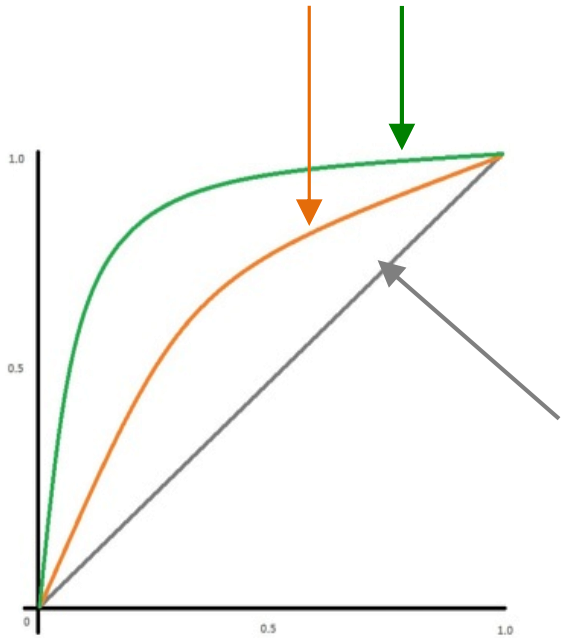
Azure Machine Learning Studio

2

Ciclo de vida del análisis de datos

4 – Elaborar el modelo : evaluar ajuste

Áreas sobre la diagonal indican una ganancia en la capacidad predictiva del algoritmo



Curva ROC

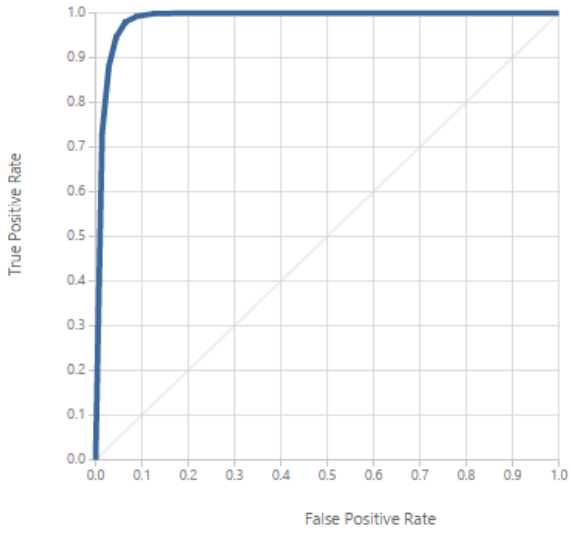
La diagonal representa un algoritmo aleatorio sin capacidad de predicción

2 Ciclo de vida del análisis de datos

4 – Elaborar el modelo : evaluar ajuste

		Predicción	
		1	0
Real	1	99,23%	0,77%
	0	8,92%	91,08%

Falsos negativos



tasa de **positivos** predichos correctamente

$$\frac{VP}{VP+FP}$$

tasa de **observaciones** correctamente predichas

$$\frac{VP+VN}{\sum Casos}$$

tasa de **verdaderos positivos** predichos correctamente

$$\frac{VP}{VP+FN}$$

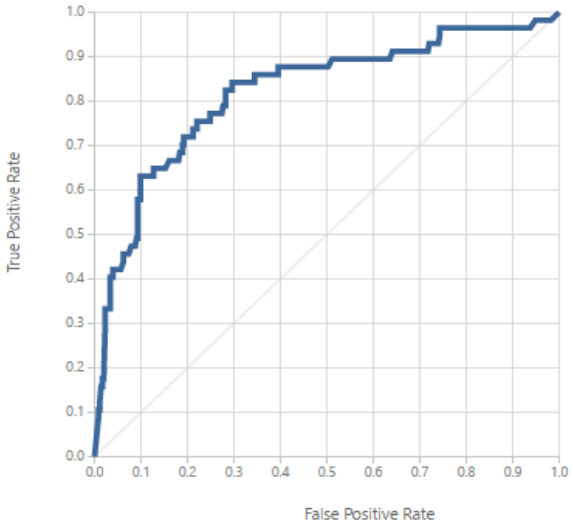
Algorithm	Accuracy	Precision	Recall
Logistic Regression	0.925667	0.266003	0.860471
Decision Jungle	0.926418	0.274812	0.908246
Decision Forest	0.937317	0.318159	0.979712
Boosted Decision Tree	0.927707	0.282431	0.937827
SVM	0.847723	0.068369	0.328665

2 Ciclo de vida del análisis de datos

4 – Elaborar el modelo : evaluar ajuste

		Predicción	
		1	0
Real	1	84,21%	15,79%
	0	31,36%	68,64%

Falsos negativos



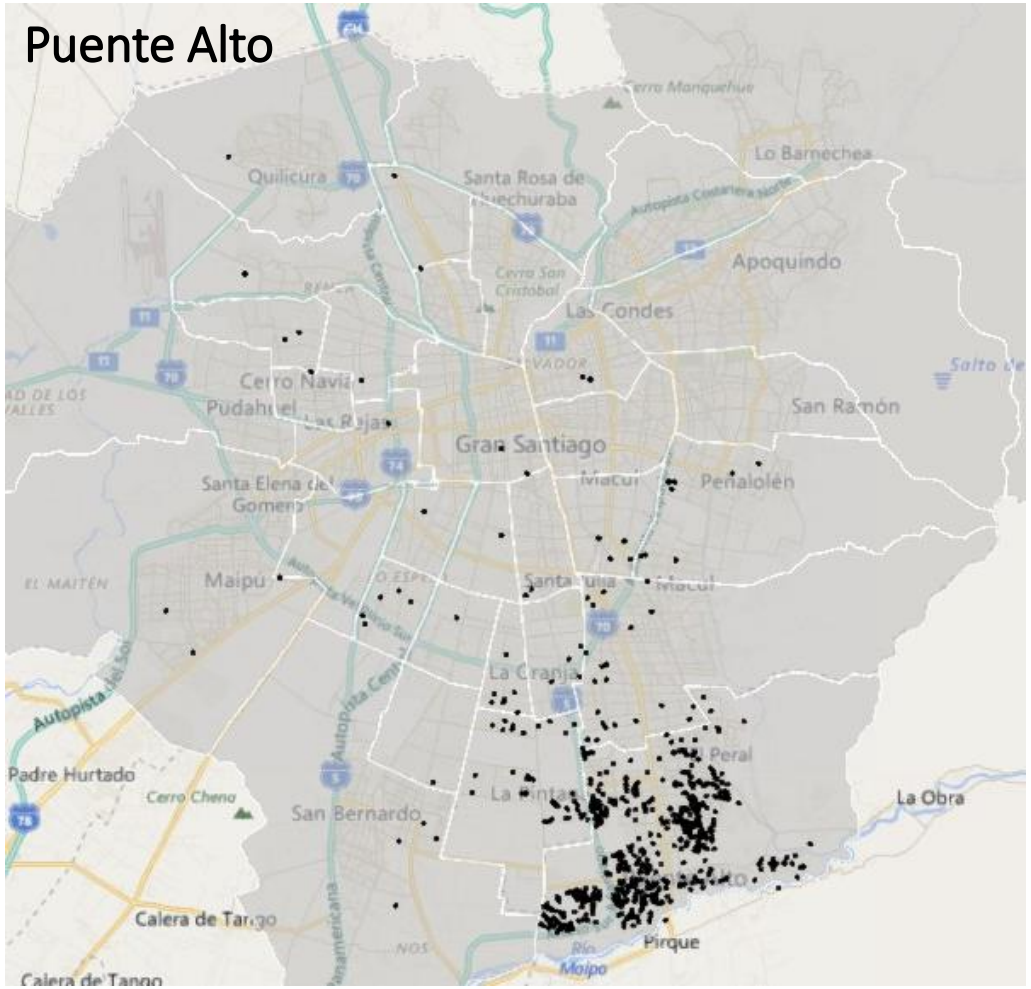
Algorithm	Accuracy	Precision	Recall
Logistic Regression	0.742138	0.301887	0.8
Neural Network	0.874214	0	0
Decision Jungle	0.712895	0.305732	0.842105
Decision Forest	0.742092	0.321168	0.77193
Boosted Decision Tree	0.68239	0.243697	0.725
SVM	0.512579	0.114094	0.425

2 Ciclo de vida del análisis de datos

5 – Comunicar

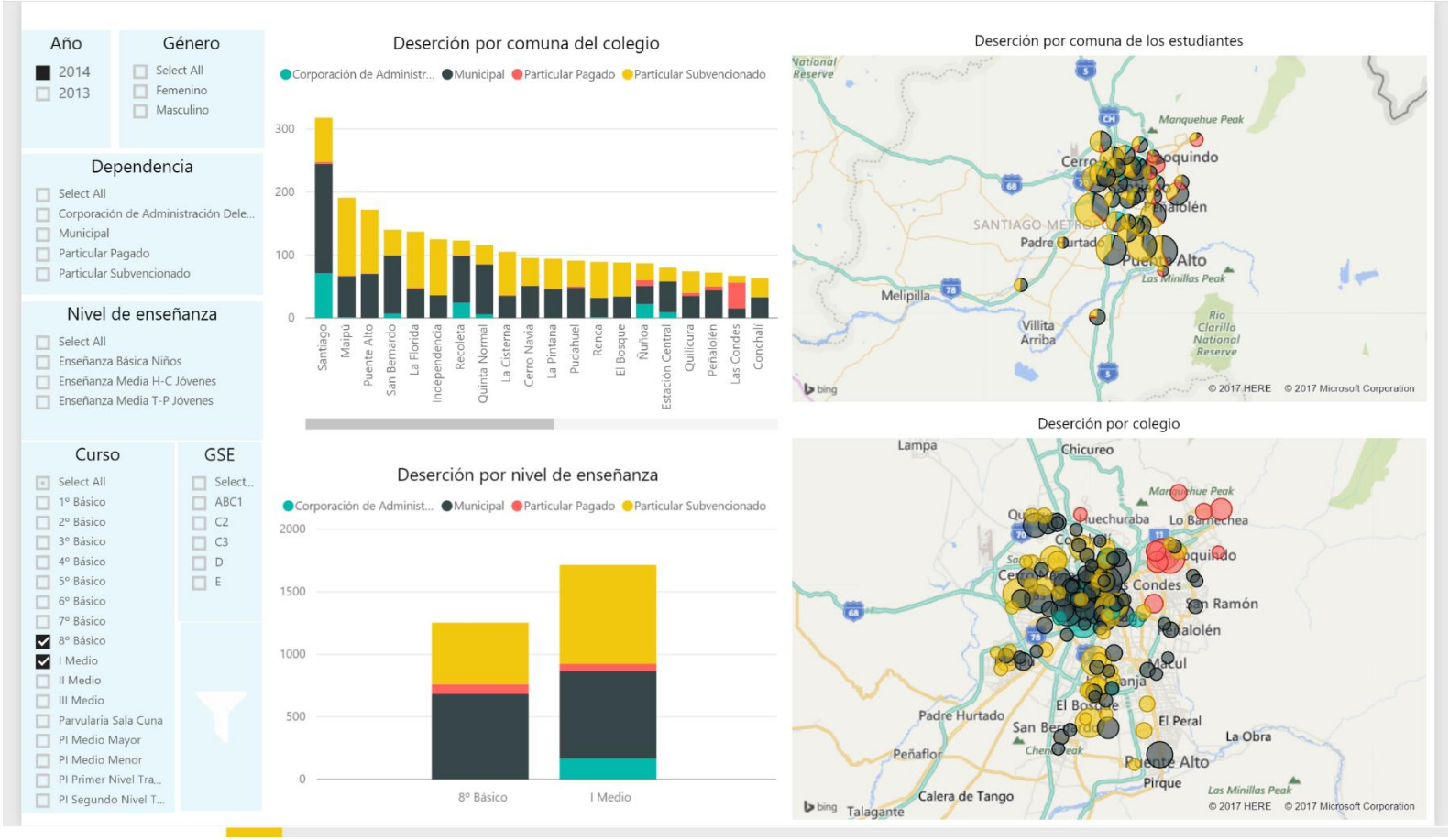
Manzanas donde viven estudiantes con $P \geq 0{,}5$ de desertar según las comunas donde estudian

Comuna donde estudia	Total
Puente Alto	683
Santiago	588
Maipú	530
La Florida	486
La Cisterna	412



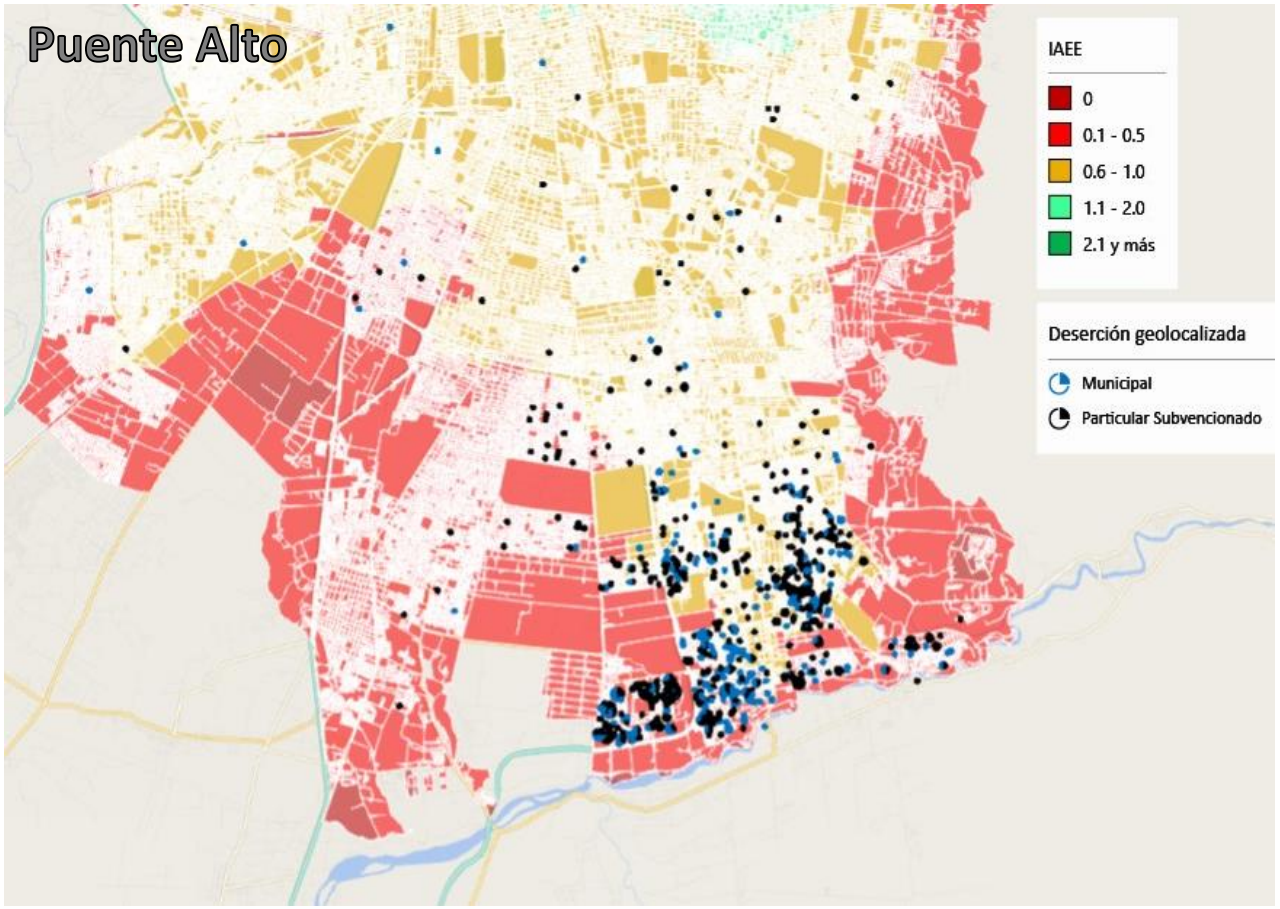
Ciclo de vida del análisis de datos

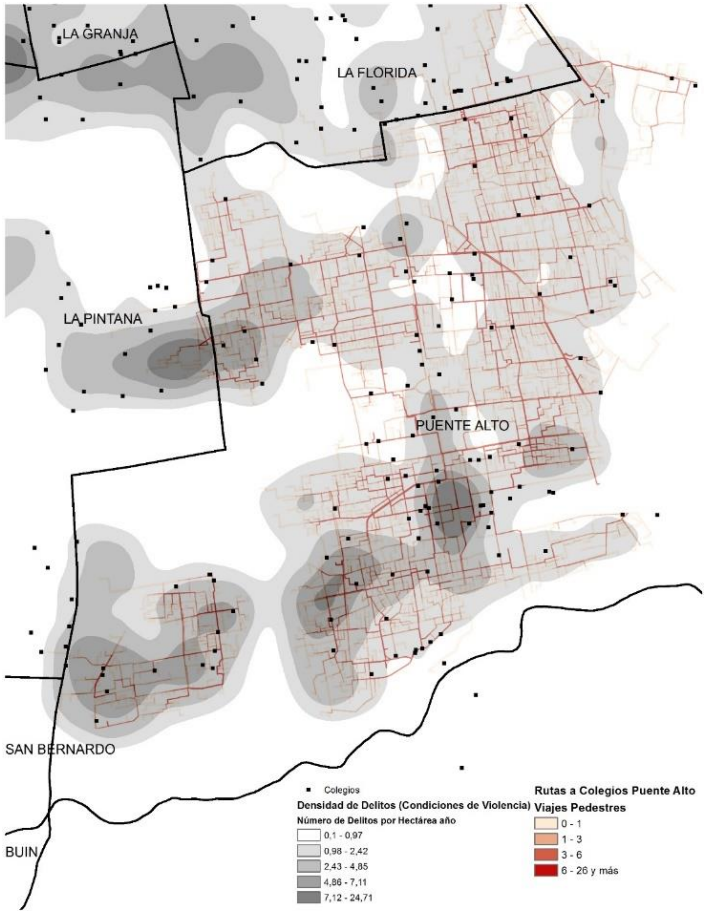
5 – Comunicar



2 Ciclo de vida del análisis de datos

5 – Comunicar





Imputación colegio cercano (20 min)

Pedestre

2

Ciclo de vida del análisis de datos

6 – Utilizar

Ciclo de vida del análisis de datos

6 – Utilizar

Usando la predicción del **sistema de alerta temprana**:

Diseñar, implementar y evaluar intervenciones para los estudiantes según su perfil y nivel de riesgo

Evaluar los resultados de las intervenciones: **replicar y escalar**

Con los resultados de las distintas intervenciones implementadas:

Generar un modelo predictivo para que escoja la intervención más costo-efectiva de según el perfil del (la) estudiante

¡Esto requiere recursos y compromiso de mediano plazo!

3 Consideraciones finales

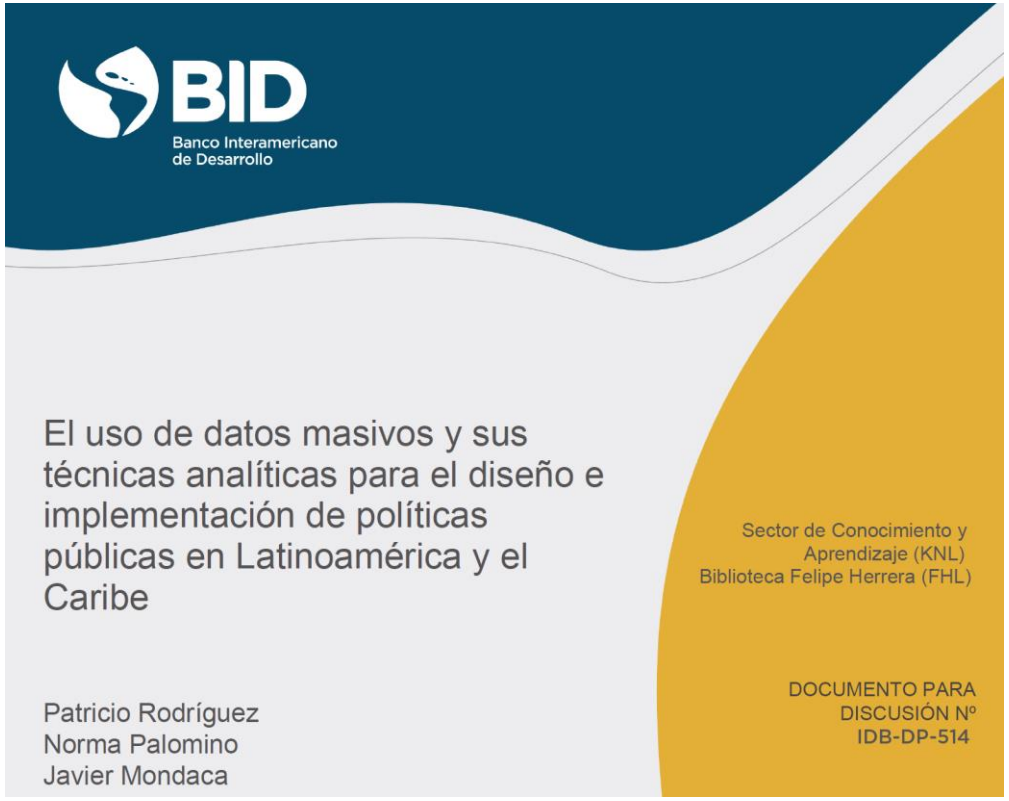
Consideraciones finales

El desarrollo de modelos predictivos requiere necesariamente tener la capacidad de tomar la predicción y hacer uso de ella: **convertirla en algo práctico**

Por lo tanto, la **asignación de recursos** debe considerar tanto el desarrollo de los algoritmos como el de las intervenciones asociadas a su uso

Esas intervenciones, con seguridad, serán más caras que el desarrollo del modelo predictivo

Para mayor información …



https://publications.iadb.org/handle/11319/8276