

## **II Jornadas Internacionales de Estadística Aplicada 5 y 6 de diciembre de 2019**

### **Minería de datos para soporte a decisiones de planificación educativa**

Ezequiel. Montes<sup>1</sup>, Ariel Mogro<sup>1</sup>, José H. Farfán<sup>1</sup>, Mariela Rodríguez<sup>1</sup>

Facultad de Ingeniería – Universidad Nacional de Jujuy<sup>1</sup>.

*ezemontes@gmail.com, arielmogro123@gmail.com, jhfarfan@gmail.com,  
mariela.rodriguez@fi.unju.edu.ar*

#### **RESUMEN**

El rendimiento educativo anual de la provincia es uno de los objetivos relevantes del Ministerio de Educación de la Provincia de Jujuy. La obtención de conocimiento o inteligencia en el ámbito educativo es de gran utilidad para los analistas de esta institución.

El proyecto parte de generar un contenedor de datos filtrados de una extensa base de datos inicial mediante la construcción de un Data Warehouse, el cual contendrá solo información relevante y concisa para un análisis posterior.

Con el uso de la Minería de Datos se pueden predecir los cursos con bajo rendimiento y deserción estudiantil, para así prevenir riesgos y aplicar estrategias en el Ministerio de Educación y en los propios establecimientos. La ventaja de utilizar la Minería de Datos y sus algoritmos radica en el cruce masivo de la información para obtener una mejor perspectiva del rendimiento educativo.

El resultado de todo este proceso se integra a un sistema de apoyo a decisiones (DSS) que permite a los analistas y directores tener una fácil manipulación de datos, como así también proporcionar un ambiente amigable y ayudarles en la toma de decisiones.

**Palabras Claves:** Data Warehouse, Minería de Datos, Rendimiento Educativo

#### **INTRODUCCIÓN**

El Ministerio de Educación de la provincia de Jujuy tiene la necesidad de contar con información de valor que le permita tomar decisiones. La misma debe estar sustentadas de acuerdo al sistema educativo actual, para focalizar las inversiones en los sectores de mayor impacto, tanto en el nivel primario como secundario.

Actualmente, la Secretaría de Planeamiento Educativo del ministerio cuenta con tres grandes bases de datos LUA, RA y Aprender, pero la capacidad de analizar de forma eficiente es reducida, requiere de trabajo manual y muchas veces no llega a procesarse a tiempo para tomar acciones correctivas o generar un seguimiento adecuado.

La planificación que se realiza en este organismo público es basándose en informes que surgen a partir del análisis de datos volcados en una planilla de cálculo; usando métodos estadísticos generales, tales como promedios de notas y porcentajes de aprobados y desaprobados. Sin embargo, este proceso no permite determinar de forma ágil qué cursos, sectores o instituciones tienen bajo rendimiento académico o cuáles cuentan con una alta tasa de deserción estudiantil; siendo éstos indicadores críticos de los problemas que se desean solucionar.

Cuando se realiza un informe y es necesario obtener información disgregada o unificada no se puede realizar en el mismo por lo que sí existe esa necesidad, se debe solicitar un nuevo informe. Esto deriva a su vez en un gasto extra del tiempo que debe invertir el personal abocado a esta tarea.

Según el CEA (Centro de Estudios de la Educación Argentina), casi el 50% de los estudiantes de Jujuy no termina el nivel secundario, este dato es realmente alarmante, es por eso que se plantea con este proyecto analizar y corroborar estos datos estadísticos.

## **METODOLOGÍA**

Para completar todas las etapas del proyecto fue necesario aplicar tres distintas metodologías: Hefesto, CRISP-DM y Scrum.

### **Hefesto**

Es una metodología que permite la construcción de un Data Warehouse de forma sencilla, ordenada y de manera intuitiva. Dada su flexibilidad, la misma puede ser embebida en cualquier ciclo de vida mientras se cumplan las condiciones esenciales para ser implementada. Estas son las razones principales por la cual se adopta esta metodología del presente trabajo final.

Hefesto fue desarrollada por el Ing. Darío Bernabéu con la finalidad de guiar el modelado del Data Warehouse. Con la ayuda de esta metodología es posible obtener, en poco tiempo, una visión integral del problema y del modelo constructivo, suficientemente sólida, para desarrollar una implementación completa del sistema requerido.

La propuesta Hefesto se desarrolla en base a una amplia y exhaustiva investigación realizada, la comparación de metodologías existentes y experiencias previas aportadas por el mismo autor, en el estudio y confección de distintos almacenes de datos.

Esta metodología posee, de acuerdo con su autor Bernabeu (2010), las siguientes características:

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- Se basa en los requerimientos de los usuarios, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- Reduce la resistencia al cambio, ya que involucra a los usuarios finales en cada etapa para que tome decisiones respecto al comportamiento y funciones del DW.
- Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- Es independiente de las herramientas que se utilicen para su implementación.
- Es independiente de las estructuras físicas que contengan el DW y de su respectiva distribución.
- Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente.
- Se aplica tanto para Data Warehouse como para un Data Mart (el cual es una versión más pequeña o reducida de un Data Warehouse).

Consta de cuatro etapas generales como se observa en la Figura 1:



Figura 1: Bernabeu (2010). Etapas de la metodología Hefesto.

## CRISP-DM

IBM Corporation (2016) señala que CRISP-DM “es un método probado para orientar trabajos de minería de datos”.

Como metodología: Incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

Como modelo de proceso: Ofrece un resumen del ciclo vital de minería de datos, tal como se puede ver en la Figura 2.

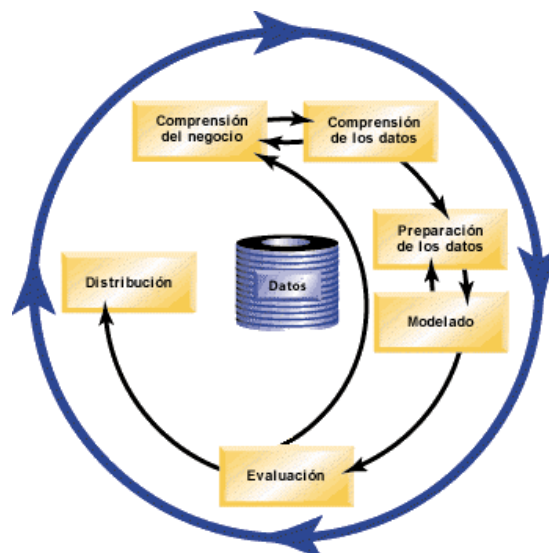


Figura 2: IBM corporation (2016). Etapas de la Metodología Crisp-dm.

En este marco se definen, a continuación, las 6 fases de la metodología:

1. Comprensión del negocio o problema: Es la fase más importante, que comprende los objetivos y requisitos del proyecto, es necesario entender de la manera más completa el problema que se desea resolver, esto permite recolectar los datos correctos y convertir el conocimiento adquirido del negocio en un problema de Data Mining.
2. Comprensión de los datos: Corresponde la recolección inicial de datos para identificar su calidad, descripción y establecer relaciones más evidentes que permitan definir las primeras hipótesis.
3. Preparación de los datos: Una vez realizada la recolección, se procede a la preparación para aplicar sobre ellos las técnicas de Minería de Datos. La preparación de datos incluye la selección de datos, limpieza de datos, integración de datos y cambios de formato. Esta fase se encuentra vinculada con la fase de modelado, ya que los datos requieren ser procesados de diferentes formas dependiendo la técnica elegida.
4. Modelado: En esta fase se seleccionan las técnicas de modelado más apropiadas al problema, por otro lado se determina un modelo de evaluación de los modelos para establecer el grado de eficiencia de los mismos. Los parámetros se utilizan para la generación de modelos, dependen de las características de los datos y el grado de precisión que se quiera lograr con el modelo.
5. Evaluación: Se evalúa el modelo teniendo en cuenta el cumplimiento de los criterios de éxito del problema, por otro lado se usa para revisar el proceso teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior donde se haya posiblemente cometido el error.
6. Implementación: Una vez que el modelo se construye y valida, se transforma en conocimiento obtenido, este se implementa, monitorea y mantiene en su lugar de aplicación.

## Scrum

Scrum es una metodología ágil para la gestión de proyectos, Deemer P. (2009) lo define como “un marco de trabajo iterativo e incremental para el desarrollo de proyectos, productos y aplicaciones. Estructura el desarrollo en ciclos de trabajo llamados Sprints que son iteraciones de 1 a 4 semanas, y se van sucediendo una detrás de otra”. Los Sprints tienen una duración fija con fecha de entrega, en cada inicio del mismo un equipo multifuncional realiza un análisis de aprobación o cambios.

El Product Backlog es una lista ordenada de todo lo que se sabe que se necesita en el producto. Es la única fuente de requisitos para cualquier cambio que se realice en el producto; siendo el Product Owner el responsable del mismo, incluyendo su contenido, disponibilidad y priorización (Sutherland, 2017).

El Product Backlog nunca está completo. Evoluciona a medida que lo hace el producto y el entorno en el que se utiliza. Cambia constantemente para identificar lo que el producto necesita para ser apropiado, competitivo y útil. Este enumera todas las características, funciones, requisitos, mejoras y correcciones que constituyen los cambios a realizar en el producto.

El Sprint Backlog es el conjunto de elementos seleccionados del Product Backlog para el Sprint, más un plan para entregar el Incremento del producto. Éste es la previsión del equipo de desarrollo sobre las funcionalidades a realizar y el trabajo necesario para concluir las.

Las User Stories son incrementos funcionales que resultan de la división del trabajo a realizar. Éstos son establecidos por el equipo junto al cliente o Product Owner.

## Proceso de obtención del conocimiento

Hefesto facilitó en cierta medida la definición del almacén de datos, permitiendo la construcción del mismo a partir de los requerimientos de información del cliente. No obstante, estos requerimientos escalaron debido a la necesidad de emplear la información en un DSS y analizarla desde múltiples perspectivas. De esta forma, la aplicación de esta metodología se tornó un poco más compleja dado que se tuvo que trabajar con casi 200 tablas repartidas entre dos esquemas distintos, de las cuales se usaron finalmente 32 para extraer la información. Así mismo, se requirió de múltiples procesos ETL, entre los que se destaca el proceso ETL de hechos con 19 pasos.

Por su parte, CRISP-DM brinda el marco para realizar la minería de datos. Al aplicarla, varios de los pasos ya habían sido abordados en Hefesto, como la fase de comprensión de negocio y la preparación y limpieza de datos.

Completando las metodologías, se encuentra Scrum, que brindó un marco flexible desde el planteo de los requerimientos y permitió concretar y validar el desarrollo sprint a sprint.

## DESARROLLO

### Exploración de datos

El trabajo utilizó un DataSet que cuenta con 160.409 registros de alumnos matriculados al inicio de año (80.806 varones y 79.603 mujeres) y a fin de año la matrícula totaliza 157.706 alumnos (79.260 varones y 78.446 mujeres). Con un total de 141.283 alumnos promovidos (71.708 varones y 69.575 mujeres).

A continuación se muestran y describen algunas de las distribuciones de los datos de la cursada en función de dimensiones tales como curso, ámbito y sector utilizando la herramienta Google Data Studio.

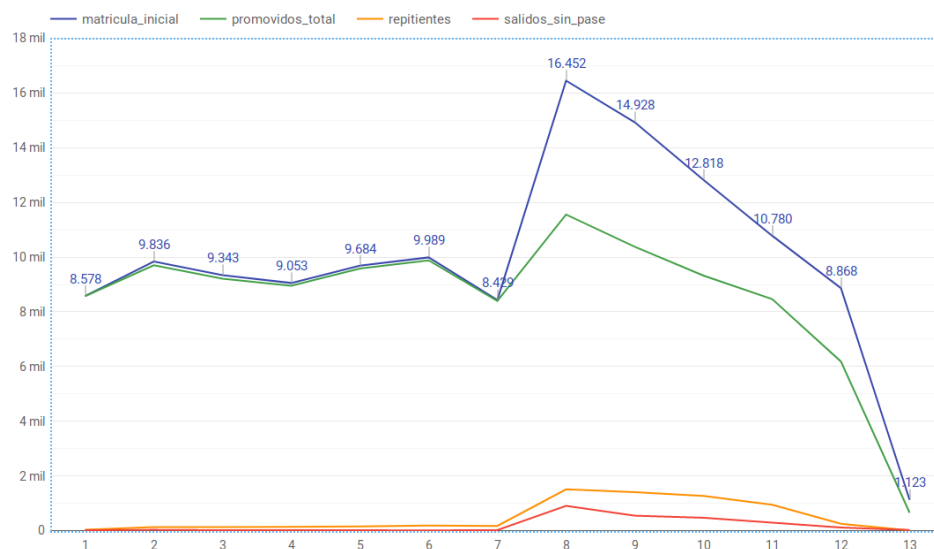
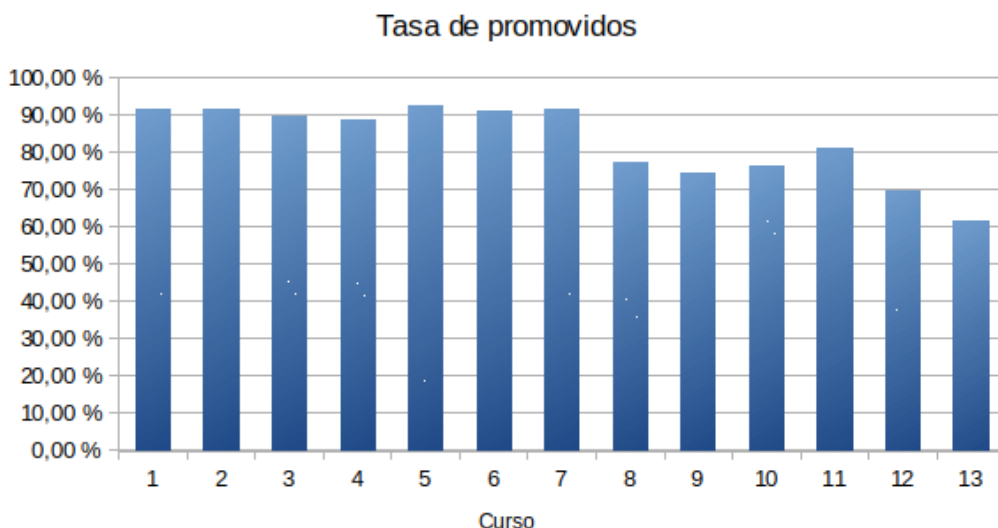


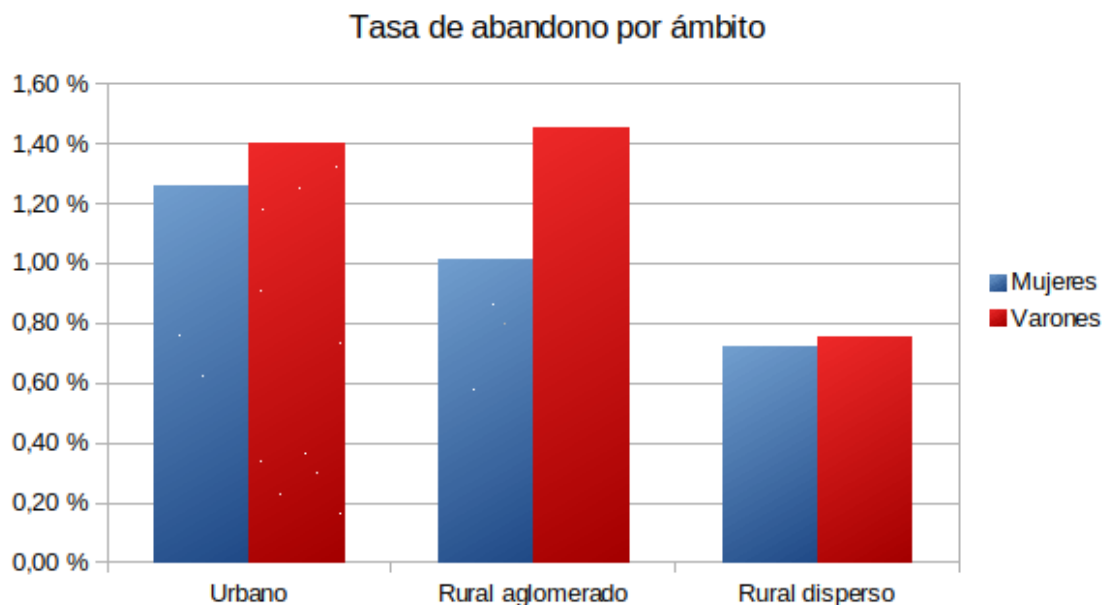
Figura 3: Cantidad de alumnos y situación por año/curso.

Cada año se matriculan una cantidad de alumnos en cada curso. A final de la cursada, un porcentaje de estos promociona, otros abandonan (salidos sin pase) y muchos regresan a repetir el curso el año siguiente. En la Figura 3 se muestran estas cantidades y se observan ciertos comportamientos particulares.



**Figura 4: Tasa de promovidos por curso.**

En la Figura 4 describe el porcentaje de promovidos de cada curso en la provincia, donde en el eje X están descrito los cursos de todos los establecimientos y en el eje Y la tasa de promovidos. Los valores 8, 9, 10, 11, 12 y 13 corresponden a los cursos del nivel secundario: 1er año, 2do año, 3er año, 4to año, 5to año y 6to año en caso de las escuelas técnicas. Se puede observar en la Figura 4, que los cursos con baja tasa de promoción son los 5to y 6to año del nivel secundario.



**Figura 5: Tasa de abandono por ámbito.**

En la Figura 5 se describe la tasa de abandono separado por ámbito. En el eje X se puede apreciar los distintos ámbitos y en el eje Y la tasa de abandono de todos los establecimientos de la provincia. Se puede observar que la tasa de abandono en todos los ámbitos es mayor en los varones, tomando en cuenta todos los establecimientos y grados de la provincia de Jujuy.

Tasa de promovidos por sector

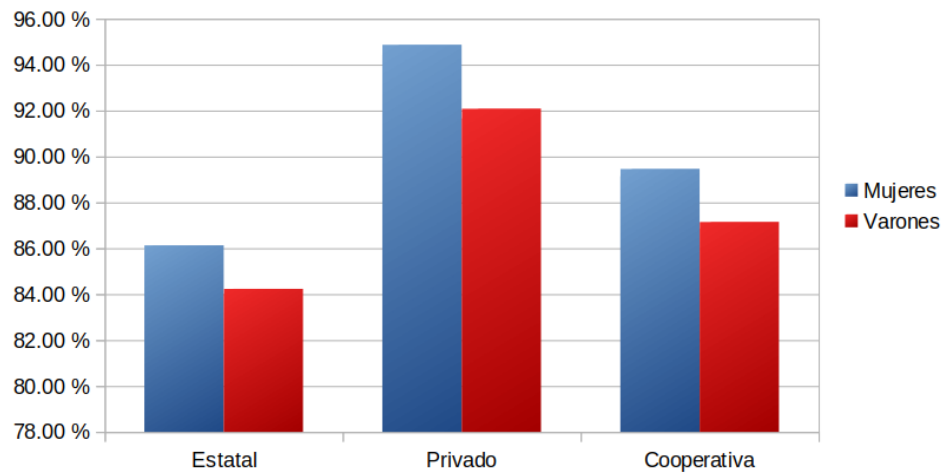


Figura 5: Tasa de promovidos por sector.

En la Figura 6 se puede observar la tasa de promovidos divididas por sector y separados por género. En el eje X se encuentran los sectores (Estatál, Privado y Cooperativa) y en el eje Y los porcentajes de promoción. Observando el gráfico se deduce claramente que la mayor tasa de promoción en todos los sectores es mayor en las mujeres.

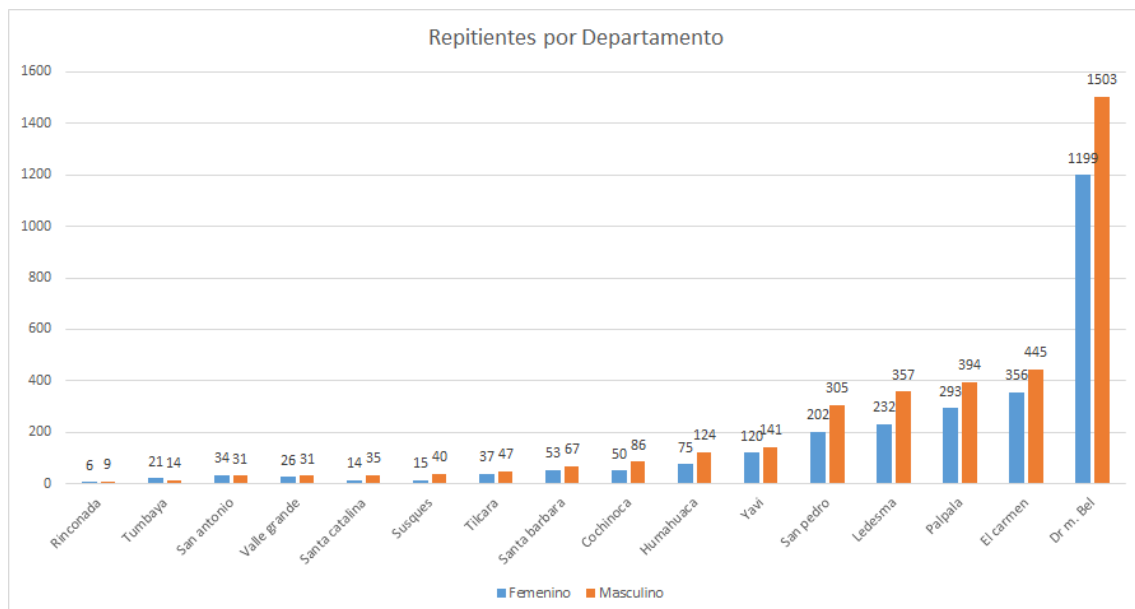


Figura 7: Repitientes por Departamento.

En la Figura 7, se puede apreciar la cantidad de repitientes en cada departamento de la provincia, separados por género. Los valores más altos, se dan proporcionalmente a la cantidad de matriculados. En el eje Y, se encuentran el total de repitientes, en el eje X los departamentos de la provincia.

## Seleccionar técnica del modelado

El modelado debe realizarse de acuerdo a las características del problema a resolver, en este trabajo final es pertinente usar predicción de los indicadores.

La herramienta H2o proporciona distintas técnicas de modelado. Varias de éstas técnicas se orientan a la predicción, por lo que se puede modelar con distintos algoritmos y compararlos para encontrar cuál el más óptimo y adecuado al problema.

## Construcción del modelo

Como paso inicial, se corre la función AutoML para obtener una orientación de modelos “óptimos” con sus respectivos hiperparámetros. Este algoritmo, como se muestra en la Figura 8, emplea el Dataset de entrenamiento (training frame) y el de validación (validation frame) para realizar los entrenamientos. Adicionalmente permite especificar la columna de respuesta (label), el máximo de modelos a construir, tiempo máximo de ejecución y las columnas y algoritmos a excluir.

**Run AutoML**

Project Name:

Training Frame:

Response Column:

Weights Column:

Ignored Columns

Showing page 1 of 1. 10 ignored.

Column	Type
<input checked="" type="checkbox"/> id	INT
<input type="checkbox"/> ambito_id	INT
<input checked="" type="checkbox"/> arancelado_id	INT
<input type="checkbox"/> categoria_id	INT
<input checked="" type="checkbox"/> confesional_id	INT
<input type="checkbox"/> departamento_id	INT
<input type="checkbox"/> dependencia_id	INT
<input type="checkbox"/> establecimiento_id	INT
<input type="checkbox"/> grado_id	INT

☒ All ☐ None

Only show columns with more than 0 % missing values.

Validation Frame:

Leaderboard Frame:

Balance classes: ☐

Exclude these algorithms

<input checked="" type="checkbox"/> GLM
<input type="checkbox"/> DRF
<input type="checkbox"/> GBM
<input checked="" type="checkbox"/> XGBoost
<input type="checkbox"/> DeepLearning
<input checked="" type="checkbox"/> StackedEnsemble

Figura 8: AutoML en H2o.ia.

Esta configuración arroja el siguiente resultado:

#### MODELS

models sorted in order of mean\_residual\_deviance, best first

model_id	mean_residual_deviance	rmse	mse	mae	rmsle
0 GBM_4_AutoML_20190812_181628	0.01133304472274819	0.10645677396365245	0.01133304472274819	0.04899908322527786	0.06764059101163096
1 GBM_3_AutoML_20190812_181628	0.011354257078097897	0.1065563563476994	0.011354257078097897	0.049175487663590256	0.06787421828615178
2 GBM_grid_1_AutoML_20190812_181628_model_5	0.011469450597878	0.10709552090483523	0.011469450597878	0.04793760907042443	0.06805142732886733
3 GBM_grid_1_AutoML_20190812_181628_model_51	0.011519645829197634	0.10732961301149667	0.011519645829197634	0.049798202712244936	0.0683153019085299
4 GBM_2_AutoML_20190812_181628	0.011598536398456713	0.10769650132876515	0.011598536398456713	0.05054427586739577	0.0685500775846194
5 GBM_grid_1_AutoML_20190812_181628_model_58	0.01162209268594294	0.10780581007507406	0.01162209268594294	0.05061462947468851	0.06871161385933731
6 GBM_grid_1_AutoML_20190812_181628_model_72	0.011632179806131149	0.10785258367851532	0.011632179806131149	0.04945851470261331	0.06867425832888158
7 GBM_grid_1_AutoML_20190812_181628_model_33	0.01168932692877807	0.10811719071811877	0.01168932692877807	0.04895435780810941	0.06876009227326901
8 GBM_grid_1_AutoML_20190812_181628_model_42	0.011689872663603261	0.10811971450019307	0.011689872663603261	0.047613153058683244	0.06819496333308626
9 GBM_grid_1_AutoML_20190812_181628_model_1	0.011749825341828815	0.10839661130233184	0.011749825341828815	0.0552729740071798	0.06963470328489257
10 GBM_1_AutoML_20190812_181628	0.011927392243015495	0.10921260111825692	0.011927392243015495	0.05020622370200583	0.06938937365792533
11 GBM_grid_1_AutoML_20190812_181628_model_28	0.01192823590677741	0.10921646353355986	0.01192823590677741	0.05245669541686536	0.06951256973600178
12 GBM_grid_1_AutoML_20190812_181628_model_45	0.012022372520449054	0.10964658006727367	0.012022372520449054	0.05623147659683476	0.07037749459947314
13 GBM_grid_1_AutoML_20190812_181628_model_56	0.01204365123882662	0.10974357037579295	0.01204365123882662	0.05709601919741492	0.0703785699923997
14 GBM_grid_1_AutoML_20190812_181628_model_60	0.012105619385725767	0.11002553969749827	0.012105619385725767	0.0564005607494299	0.07096152808316332
15 GBM_grid_1_AutoML_20190812_181628_model_38	0.012162059269781866	0.11028172681719246	0.012162059269781866	0.05548537304043428	0.07045845094122616
16 GBM_grid_1_AutoML_20190812_181628_model_22	0.012200530696281214	0.1104560124949349	0.012200530696281214	0.05703778632382434	0.07089799029874523
17 GBM_grid_1_AutoML_20190812_181628_model_57	0.012314442626598048	0.11097045835085141	0.012314442626598048	0.057017883932680105	0.07094983090881243
18 GBM_grid_1_AutoML_20190812_181628_model_10	0.012417724361147837	0.11143484356855282	0.012417724361147837	0.05875556793249472	0.07178647916669528
19 GBM_5_AutoML_20190812_181628	0.012417849980051339	0.11143540720996778	0.012417849980051339	0.05339961615031124	0.07108162206813372
20 GBM_grid_1_AutoML_20190812_181628_model_37	0.012439764248764737	0.11153369109271305	0.012439764248764737	0.05530542548860042	0.0710994340028426
21 GBM_grid_1_AutoML_20190812_181628_model_50	0.012479834147724005	0.11171317803967447	0.012479834147724005	0.05478408468854976	0.07147931055113348
22 GBM_grid_1_AutoML_20190812_181628_model_27	0.012554287680452933	0.11204591773220894	0.012554287680452933	0.05744854580196895	0.07143060854568163

Figura 9: Resultado AutoML.

Como se muestra en la Figura 9, los modelos predictivos obtenidos, a través de la técnica de GBM han ofrecido resultados más precisos. Estos modelos son sometidos a iteraciones y ajustes de parámetros para tratar de reducir el error cuadrático medio.

## Generar plan de prueba

Una vez entrenado el modelo, se mide su rendimiento mediante el error cuadrático medio. No obstante, para comprobar realmente la capacidad que tiene el modelo de realizar predicciones con nuevos datos, es necesario probarlo con un conjunto de datos distintos. Por lo que se debe reservar parte del Dataset principal para esto.

Usando este criterio, en el presente proyecto se dividió el Dataset dejando el 80% de los datos para el Dataset de entrenamiento y un 20% para el Dataset de validación. Esta división implica además que los ejemplos se seleccionan de forma aleatoria para completar cada subconjunto de datos.

Cuando se realiza un entrenamiento se especifican los Datasets de entrenamiento y validación, y con ellos la herramienta calcula automáticamente el MSE para los mismos. Con los MSE de cada modelo, se realiza una comparación para determinar cuál es el más preciso.

## Selección de parámetros

Para realizar el modelado se seleccionan los parámetros que se determinan con AutoML y se trata de obtener modelos de mayor precisión ajustando estos hiperparámetros. En la Figura 10 se muestran los parámetros del mejor modelo encontrado para la tasa de promoción, de la misma manera se realiza para los demás indicadores.

▼ MODEL PARAMETERS

Parameter	Value
model_id	GBM_4_AutoML_20190812_181628
training_frame	automl_training_hechos_80
validation_frame	hechos_20
nfolds	5
keep_cross_validation_models	false
keep_cross_validation_predictions	true
score_tree_interval	5
fold_assignment	Modulo
response_column	tasa_promocion
ignored_columns	id, arancelado_id, confesional_id, nivel_id, subvencion_id, matricula_inicial, promovidos, salidos_sin_pase, repitientes, tasa_abandono
ntrees	64
max_depth	15
min_rows	100

Figura 10: Hiperparámetros tasa de promoción GBM.

Cabe destacar que si bien el algoritmo GBM posee una gran cantidad de parámetros, los que se destacan sobre los demás son *ntree*, *min\_rows* y *max\_depth*. Ya que las pruebas empíricas realizadas con los demás no muestran una influencia significativa en el comportamiento de los modelos; razón por la cual no se incluyen en este trabajo. En la Tabla 1 se detallan los valores iniciales de los parámetros para el entrenamiento de cada indicador.

<i>Tabla de parámetros iniciales</i>	<i>ntrees</i>	<i>max_depth</i>	<i>min_rows</i>
Tasa de promoción	65	15	100
Tasa de abandono	33	15	100
Repitientes	241	1	1

Tabla 1: Parámetros iniciales.

## Evaluación del modelo

La herramienta H2o ofrece, para cada modelo generado, un gráfico de la progresión del MSE de entrenamiento y de validación.

A continuación, se analiza las gráficas de la aplicación de la técnica GBM a la tasa de promoción. Como se muestra en la Figura 11, las curvas del MSE caen significativamente en los primeros 10 árboles y luego en menor grado hasta que deja de disminuir entre los 60 y los 65 árboles. Con el ajuste de parámetros se determina el comportamiento y si existen mejoras para el MSE.

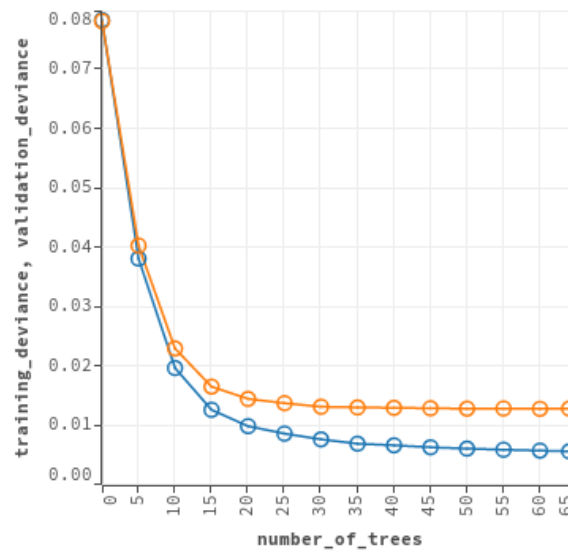


Figura 11: Evolución del MSE de GBM de la tasa de promoción.

En la aplicación de GBM para la tasa de promoción, como se observa en la Figura 12, se muestra que las curvas de MSE las cuales caen gradualmente, sin tanta acentuación. Al finalizar cada curva, no se puede determinar fehacientemente si es posible que siga mejorando o no. Por lo que invita a ajustar los parámetros y volver a entrenar.

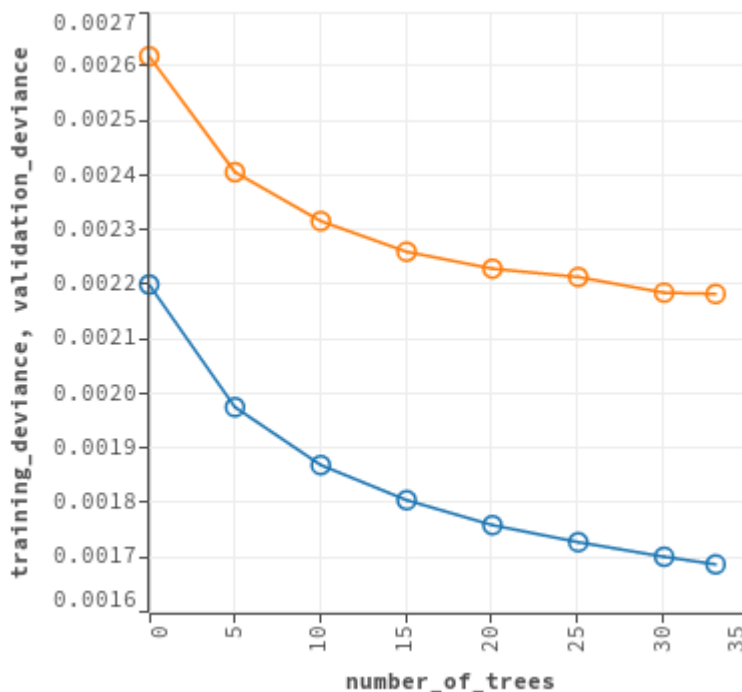


Figura 12: Evolución del MSE de GBM de la tasa de abandono.

En caso de la aplicación al indicador Repitientes. Como se muestra en la Figura 12, las curvas disminuyen rápidamente en los primeros 30 árboles. Luego se estabiliza con una pequeña curvatura hasta alcanzar los 80 árboles donde el MSE de validación se mantiene constante y despegado del MSE de entrenamiento, lo que evidencia un sobreajuste en el modelo.

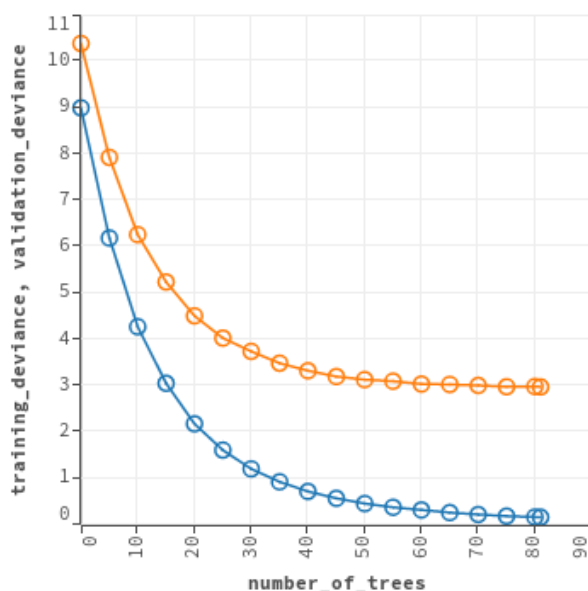


Figura 12: Evolución del MSE de GBM de repitientes.

### Revisión de los parámetros

Luego de definir los ajustes de hiperparametros, se procede a la revisión de parámetros y resultados obtenidos.

<i>Tasa de promoción</i>	<i>ntrees</i>	<i>max_depth</i>	<i>min_rows</i>	<i>MSE</i>
Iteración 1	65	15	100	0,013274
Iteración 2	30	15	100	0,014103
Iteración 3	100	15	100	0,012942
Iteración 4	100	15	50	0,012790
Iteración 5	100	20	50	0,012850

Tabla 2: Ajustes de parámetros para tasa de promoción.

Como se puede observar en la Tabla 2, la iteración cuatro presenta el error más óptimo. A partir de la siguiente fase se tiene en cuenta únicamente este modelo con sus respectivos parámetros.

<i>Tasa de abandono</i>	<i>ntrees</i>	<i>max_depth</i>	<i>min_rows</i>	<i>MSE</i>
Iteración 1	33	15	100	0,002184
Iteración 2	50	15	100	0,002163
Iteración 3	50	15	50	0,002049
Iteración 4	50	10	50	0,002054
Iteración 5	50	50	50	0,002045

Tabla 3: Ajustes de parámetros para tasa de abandono.

Como se visualiza en la Tabla 3, la iteración cinco presenta el resultado más óptimo. A partir de la siguiente fase se tendrá en cuenta únicamente este modelo con sus respectivos parámetros.

<i>Repitientes</i>	<i>ntrees</i>	<i>max_depth</i>	<i>min_rows</i>	<i>MSE</i>
Iteración 1	241	13	1	2,939062
Iteración 2	100	15	1	2,987184
Iteración 3	241	20	1	2,872974
Iteración 4	241	15	10	3,453025
Iteración 5	241	25	1	2,932054

Tabla 4: Ajustes de parámetros para repitientes.

Como se puede observar en la Tabla 4, se genera el resultado más óptimo en la tercera iteración.

## CONCLUSIONES

De acuerdo a los requerimientos solicitados, se logró determinar los indicadores educativos más relevantes (tasa de promoción, tasa de abandono y cantidad de repitientes), siendo éstos los usados para realizar el análisis de la eficiencia interna del sistema educativo.

Se utilizaron los indicadores para medir la eficiencia mediante cálculos de la información provista del Relevamiento Anual, construyendo un Data Warehouse con estos. Para la construcción se utilizó un proceso de ETL con la herramienta Spoon.

A partir del DataWarehouse, se pueden generar los informes sobre los indicadores del ciclo lectivo anterior. No obstante, el factor clave para el éxito del prototipo fue la inclusión de los modelos predictivos generados por la minería de datos, los cuales se determinaron en gran medida gracias a la ayuda que brinda la herramienta H2o a través de su función AutoML, con la cual se determinó que la determinaba un mayor grado de precisión fue la técnica GBM. La precisión de estos modelos, permite concluir que las técnicas de minería de datos poseen importantes ventajas en el modelado del comportamiento del flujo de alumnos (tasa de promoción, tasa de abandono y repitientes).

Dado que la predicción de los indicadores de forma anticipada agrega un alto valor para el análisis y la generación de nuevo conocimiento para la toma de decisiones. Es importante destacar también que la herramienta H2o permite exportar los modelos entrenados, los cuales pueden embeberse en una aplicación Java Web. Por lo tanto, gracias a la funcionalidad mencionada, fue posible realizar la integración de éstos modelos en el DSS.

## BIBLIOGRAFIA

- Bernabeu R. D. (2010). *Hefesto: Metodología para la Construcción de un Data Warehouse*. Córdoba, Argentina.

- IBM Corporation. (2016). *Manual CRISP-DM de IBM SPSS Modeler* [Gráfico]

Recuperado de:

[public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CR](https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf)

[ISP-DM.pdf](https://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf). Consultado el 07/09/2018.

- Deemer P., Benefield G., Larman C., & Vodde B. (2009). *Información básica de SCRUM. California: Scrum Training Institute.*
- Sutherland K. & Schwaber J. (2017). *The Scrum Guide*. Recuperado de:  
<https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-US.pdf>