CUADERNOS DEL CIMBAGE

Universidad de Buenos Aires
Facultad de Ciencias Económicas

PRONÓSTICOS Y DATA MINING PARA LA TOMA DE DECISIONES. PRONÓSTICO SOBRE LA DESERCIÓN DE ALUMNOS DE UNA FACULTAD

Autor(es): CHINKES E.

Fuente: Cuadernos del CIMBAGE, Nº 20 (Mayo 2018), pp 107-132

Publicado por: Facultad de Ciencias Económicas de la Universidad de Buenos Aires.

Vínculo: http://ojs.econ.uba.ar/ojs/index.php/CIMBAGE/issue/view/173

PRONÓSTICOS Y DATA MINING PARA LA TOMA DE DECISIONES. PRONÓSTICO SOBRE LA DESERCIÓN DE ALUMNOS DE UNA FACULTAD

Ernesto Chinkes
Programa Ejecutivo de Business Intelligence. Escuela de Posgrado Facultad de Ciencias Económicas. Universidad de Buenos Aires
Av. Córdoba 2122- CABA- C1120AAQ- Argentina
echinkes@rec.uba.ar

Recibido 10 de octubre 2017, aceptado 12 de diciembre 2017

Resumen

Para tomar decisiones es necesario entender la realidad de la organización y su contexto. No sólo en la actualidad, sino también en el futuro. Muchas decisiones se basan en supuestos sobre el futuro. Para reducir la incertidumbre sobre el futuro existen los pronósticos usando métodos cuantitativos. Este trabajo analiza estos métodos y como el *data mining*, *machine learning* y las soluciones informáticas, cobijadas en las soluciones denominadas "Analytics", pueden aportar en dicho camino. Para ello más alla de explicar los conceptos que lo sustentan, se expone cómo es posible aplicarlo para pronosticar la deserción de estudiantes para una facultad a partir de aplicar modelos predictivos, mediante la tarea de clasificación, usando una herramienta de *machine learning* que funciona en la nube. También invita a pensar cual es el impacto, que pueden tener estas posibilidades en la era del *big data*.

Palabras clave: *Data mining*, *machine learning*, pronósticos, deserción de alumnos.

FORECAST AND DATA MINING FOR MAKING DECISION. FORECAST ON DROPOUT OF STUDEN OF A FACULTY

Ernesto Chinkes
Programa Ejecutivo de Business Intelligence. Escuela de Posgrado Facultad de Ciencias Económicas. Universidad de Buenos Aires
Av. Córdoba 2122- CABA- C1120AAQ- Argentina
echinkes@rec.uba.ar

Received October 10th 2017, accepted December 12nd 2017

Abstract

In order to make decisions, it is necessary to understand the reality of the organization and its context. Not only today, but also in the future. Many decisions are based on assumptions about the future. To reduce uncertainty about the future, there are forecasts using quantitative methods. This paper analyzes these methods and how data mining, machine learning and IT solutions, including in called Analytics solutions, can contribute in this way. In addition to explaining the concepts that support it, it is explained how it is possible to apply it to predict the dropout of students for a faculty applying predictive models, through the classification task, using a machine learning tool that works in the cloud. Also invites you to think about the impact that these possibilities can have in the big data era.

Keywords: Data mining, machine learning, forecast, dropout of students.

1. EL FUTURO Y LA TOMA DE DECISIONES

Las decisiones que se toman en las organizaciones, y que son las que determinan su destino, ameritan un entendimiento de la realidad de la propia institución así como del contexto en el que ella se desenvuelve. Pero no sólo lo que ha pasado, o está pasando, sino sobre todo comprender como será el futuro, ya que es allí en donde se llevarán a cabo las acciones de las decisiones que se están considerando.

Para decidir los planes de producción o las compras de los productos a comercializar se necesitará estimar previamente cual será el volumen de ventas que se prevé (por ejemplo para el próximo trimestre); o para tomar la decisión de ampliar las instalaciones o construir un nuevo edificio en una universidad será importante saber cuál es la cantidad de alumnos que estarán cursando en los próximos años (estimar nuevos ingresantes, cantidad de alumnos que desertarán, etc.). Estas estimaciones acerca del futuro son los "pronósticos".

Es decir, cotidianamente los directivos deben tomar decisiones sin saber qué es lo que pasará efectivamente en el futuro, por eso tratan de reducir dicha incertidumbre mediante la realización de pronósticos (Render, 2012). Los pronósticos pueden ser totalmente subjetivos o, tal como se abordara en este trabajo, basados en métodos cuantitativos.

Muchas decisiones se basan en pronósticos, ya sea porque se calculan de manera explícita, o porque tiene detrás supuestos que de alguna manera están implicando estimaciones sobre el futuro.

Malos pronósticos pueden dar como resultado malas decisiones y un impacto negativo significativo para la institución. Cuanto más mejoren las estimaciones, se estarán mejorando también los supuestos de las decisiones que se tomen, y por ende mejorando las chances de tomar mejores decisiones. Según la interpretación que los decisores hagan acerca del futuro será la pertinencia de las estrategias, planes y otro tipo de decisiones e iniciativas que finalmente asuman.

Desde ya que, ese tipo de decisiones racionales, basadas en información, no son las únicas existentes en las organizaciones; aspectos como la emoción, la percepción, el prejuicio o la intuición serán un componente importante en la toma de decisiones (Pavesi, 2004), ya que es difícil escindir de la misma estos aspectos del ser humano dentro de un acto humano como es tomar una decisión. No obstante ello, en el momento histórico que toca vivir, no debieran desperdiciarse la posibilidad de disponer de herramientas que ayuden a tener información como nunca antes se tuvo; e inclusive predecir,

con la ayuda de las bases de datos y de algoritmos informatizados, determinados aspectos de la realidad.

En este trabajo se pretende analizar el apoyo que pueden generar algunos modelos de análisis cuantitativos para los pronósticos, pero con el uso de herramientas informáticas dentro de lo que se denomina Analytics, y más particularmente con algoritmos de *data mining* y *machine learning*.

El análisis cuantitativo consiste en definir un problema, desarrollar un modelo, obtener datos de entrada, desarrollar la solución, testearla, analizar los resultados e implementarlos. Los modelos a desarrollar contendrán variables y parámetros; donde una variable es una medida cuantificable que puede variar sus valores (y en algunos casos pueden ser desconocidos), a su vez, un parámetro es una medida cuantificable que es inherente al problema (siempre son conocidos). Los modelos deben ser resolubles, realistas, fáciles de comprender y modificar, y los datos de entrada deben ser obtenibles (Render, 2012).

En este trabajo no se pretende profundizar sobre cuál es el rol, el grado y la relevancia que se le debe dar a estas herramientas dentro del proceso decisorio, sino simplemente demostrar las posibilidades que en la actualidad pueden proveer para predecir determinados aspectos del futuro, y para ello se trabajará con un caso: *predecir la deserción de alumnos que tendrá una facultad al año siguiente*.

2. LOS PRONÓSTICOS Y LOS MODELOS MATEMÁTICOS

Como se mencionó, muchas decisiones que se toman en las organizaciones están basadas en pronósticos. Para pronosticar es posible basarse en los datos históricos de la misma variable que se desea estimar, formando lo que se denomina una serie de tiempo.

Una serie de tiempo es un conjunto de observaciones de una variable medida en puntos sucesivos en el tiempo o a lo largo de periodos sucesivos. El objetivo de los pronósticos es predecir el valor futuro de una variable en la serie de tiempo (Anderson, 2011).

Según Render (Render, 2012) hay ocho pasos para hacer un pronóstico:

1. Determinar el uso que se dará al pronóstico (cuál es su objetivo).

2. Elegir los ítems que serán pronosticados.

3. Determinar el horizonte temporal del pronóstico (corto plazo: 1-30 días, mediano: 1 mes a un año, largo plazo: más de un año).
4. Elegir el/los modelo/os de pronóstico.
5. Obtener los datos o información necesaria para hacer el pronóstico.
6. Validar el modelo.
7. Hacer el pronóstico.
8. Implementar los resultados.

Estos pasos son necesarios a la hora de concebir, diseñar e implementar un sistema de pronóstico; pero una vez que el mismo esté implementado, cuando se desea generar pronósticos regularmente, los datos serán colectados en forma rutinaria y el pronóstico podrá realizarse en forma automática.

Tal como se menciona en el paso 4, se debe elaborar un modelo. Para ello se usan métodos, los que se pueden clasificar entre cuantitativos y cualitativos.

Los primeros (cuantitativos) se utilizan cuando: 1) se dispone de información pasada sobre la variable que se pronosticará, 2) la información puede cuantificarse, y 3) es razonable suponer que el patrón del pasado seguirá ocurriendo en el futuro. (Anderson, 2011)

Estos a su vez pueden clasificarse entre, métodos de serie de tiempo y causales. En los métodos de serie de tiempo los datos históricos se restringen a valores pasados de la variable que se intenta pronosticar. El objetivo es descubrir un patrón en los datos históricos y luego extrapolarlo hacia el futuro. Los métodos de elaboración de pronósticos causales se basan en el supuesto de que la variable que se trata de pronosticar exhibe una relación de causa y efecto con una o más variables (Anderson, 2011). Es en este último tipo en el que se trabajará.

En estos métodos causales, se introduce otra variable o variables que no es el tiempo, para intentar explicar el comportamiento de la variable que se desea predecir.

El análisis de regresión, por ejemplo, es una técnica estadística donde se desarrolla una ecuación matemática que establezca cómo se relacionan las variables en juego. La variable a predecir se llama variable dependiente o de respuesta (representada generalmente con la letra “y”). La variable o variables que se utilizan para predecir el valor

de la variable dependiente se llaman variables independientes o pronosticadores (representadas generalmente con la "x"). El análisis de regresión que involucra una sola variable independiente y una variable dependiente, para el cual la relación entre las variables se aproxima por medio de una recta, se llama regresión lineal simple. El análisis de regresión que integra dos o más variables independientes se conoce como análisis de regresión múltiple (Anderson, 2011).

Como ejemplo de otro tipo de modelos causales, en una sección posterior, se trabajará un modelo que permita predecir si los estudiantes de una facultad en la universidad, desertarán o continuarán sus estudios el año siguiente. Para ello se intentará encontrar un modelo matemático que establezca como se relacionan distintas variables (variables independientes) como la edad de los alumnos, lugar donde viven, sexo, desempeño que tuvieron durante el año anterior, año de ingreso, etc., con la variable dependiente (desertor). Luego a partir de ese modelo, y conociendo los valores de las variables independientes para un momento futuro (por ejemplo el año próximo) podrá estimarse si dichos alumnos abandonarán sus estudios en la facultad durante el periodo futuro considerado.

En los métodos de elaboración de pronósticos cuantitativos se requieren datos históricos sobre la variable de interés, de modo que no pueden aplicarse cuando no se dispone de estos datos. Además, aun cuando se disponga de dichos datos, un cambio significativo en las condiciones del entorno que afectan la serie de tiempo puede hacer cuestionable el uso de los datos pasados en la predicción de valores futuros de las series de tiempo (Anderson, 2011).

Es en estos casos, por lo tanto, es donde cobran mayor significatividad los métodos cualitativos que necesitan del juicio de los expertos. Si bien no se trabajarán aquí estos últimos casos, interesa evidenciar con esto las limitaciones de los métodos cuantitativos, y los cuidados que hay que tener con estas predicciones. Es decir, que son una ayuda de gran valor, pero que deben ser consideradas siempre entendiendo sus limitaciones.

3. EL DATA MINING Y MACHINE LEARNING

La minería de datos (*data mining*) puede definirse como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias que surgen al examinar grandes cantidades de datos (Pérez López, 2007).

Machine learning son métodos computacionales que usan la experiencia para mejorar el desempeño o para hacer predicciones, donde "experiencia" se refiere a la información del pasado que se encuentra disponible, lo que típicamente llevará considerar los datos electrónicos que fueron colectados y puestos a disposición del análisis (Mohri, 2012).

Machine learning y *data mining* suelen emplear los mismos métodos, y en ambos casos la idea es usar algoritmos (matemáticos) que sean capaces de "aprender" un modelo, para luego usar el mismo para resolver problemas. Es decir, usan algoritmos matemáticos para descubrir relaciones que están implícitas en los datos (Hernandez Orallo, 2004).

Si bien existen distintas definiciones sobre ambos, es más fácil encontrar en las mismas sus coincidencias y no tanto sus diferencias, pero podría intentarse una diferenciación mencionando que en *machine learning* generalmente se intenta reproducir el conocimiento ya conocido, se centra en el aprendizaje, mientras que en *data mining* la tarea clave es descubrimiento del conocimiento, es decir lo que previamente se desconocía. Los datos fueron producidos por "algo" que puede ser aprendido, en el caso del *data mining* los datos son la materia prima en base a lo cual se quiere descubrir algo nuevo (Mannila, 1996).

Por ejemplo, en el primer caso, podría considerarse un modelo de *machine learning* que permite la clasificación automática del correo electrónico (como spam o no), mediante al aprendizaje de las clasificaciones que se hicieron previamente. En el segundo, el uso de *data mining* para descubrir cuáles son las características de los clientes que le generan a la empresa las mayores utilidades.

Ambos conceptos no son nuevos, y mucho menos las técnicas y disciplinas que los sustentan (Hernandez Orallo, 2004). Lo que le ha dado un gran impulso en los últimos tiempos, y principalmente bajo el nombre de *machine learning*, es la gran capacidad de datos que se encuentran disponibles y las posibilidades de procesamiento y almacenamiento informático que son posibles en la actualidad.

Para potenciar aún más ambos conceptos puede agregarse la baja en los costos de procesamiento y almacenamiento, y la nube que permite obtener grandes capacidades de ellos sin realizar una inversión inicial importante. Estos aspectos, sumados a la facilidad de uso del software para implementar las técnicas matemáticas en que se sustentan es una mezcla que pone a estas soluciones en un momento óptimo para ser aprovechado, y que no se ha dado en otro momento de la historia.

Tal como se mencionaba más arriba en ambos casos se usan los mismos métodos. Existen modelos predictivos y descriptivos. Los primeros sirven para estimar valores futuros de una variable, en cambio los segundos sirven para entender las propiedades, patrones y relaciones de los datos examinados. En los modelos predictivos podemos mencionar las denominadas tareas de regresión y de clasificación. En el caso de los modelos descriptivos las de agrupamiento (clustering), reglas de asociación y análisis correlacional (Hernandez Orallo, 2004).

En este trabajo se desea profundizar su aplicación en los pronósticos (predicción), y en particular ejemplificarlo mediante la deserción de los alumnos universitarios en un futuro. Por ello se debe considerar un modelo predictivo, donde se podrían estar usando tareas de regresión o de clasificación.

En la clasificación, cada instancia o registro de los datos analizados pertenece a una clase, la cual se indica mediante el valor de un atributo que se denomina la clase de la instancia, el resto de los atributos (que se consideren relevantes para el modelo) se usan para predecir la clase. El objetivo del algoritmo será "tomar la experiencia" de las clases que fueron clasificadas en el pasado, para poder predecir la clase de nuevas instancias de las que se conocerá de antemano el resto de los atributos.

La regresión considera elaborar una función que pueda asignar a cada instancia un valor numérico. Es decir se genera una función que en base a los valores de los atributos de la instancia genera un valor numérico para dicha variable a predecir.

Para el caso de los modelos descriptivos, el "agrupamiento" consiste en obtener grupos de instancias a partir de los datos (se habla de grupos y no de clases). En lugar de analizar datos etiquetados (con su clase), como sucede en la clasificación, lo que se quiere es generar dichas "etiquetas" en base a los datos. Es decir que las etiquetas no existen en forma previa. Otra tarea de modelos descriptivos es la "correlación" sirve para evaluar el grado de similitud entre el comportamiento de dos variables numéricas. Por último otra tarea que puede mencionarse es la de "reglas de asociación" que es similar a la correlación, ya que intenta identificar relaciones entre variables, pero en este caso categóricas.

Cabe aclarar que para cada tarea (regresión, clasificación, agrupamiento, correlación o asociación) se pueden usar distintas técnicas/métodos (con sus algoritmos) para su realización, como ser árboles de decisión, redes neuronales, algoritmos genéticos,

aprendizaje bayesiano, etc.; y por otro lado el mismo método podrá se usando para distintas tareas. A forma de ejemplo puede observarse la siguiente tabla obtenida de (Hernandez Orallo, 2004) donde se cruzan tareas y con algunos de los principales algoritmos:

Modelos / Tareas / Algoritmos	**PREDICTIVO**		**DESCRIPTIVO**		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones
Redes neuronales	X	X	X		
Arboles de decisión ID3, C4.5, C5.0	X				
Arboles de decisión CART	X	X			
Redes de Kohonen			X		
Regresión lineal y logarítmica		X			X
Kmeans			X		
Naives Bayes	X				
Vecinos más próximos	X	X	X		
Algoritmos genéticos evolutivos	X	X	X	X	X
Máquinas de vectores de soporte	X	X	X		

Tabla 1. Tareas y sus principales algoritmos

Tal como se mencionó previamente, dado que se está considerando resolver modelos predictivos, se trabajará en la sección 5 con la tarea de clasificación, y para ello se evaluarán distintos algoritmos, como redes neuronales, arboles, etc.

4. LAS BASES DE DATOS Y LA PERSPECTIVA DEL BIG DATA.

Obtener los datos adecuados para un modelo es fundamental. Si los datos no son precisos y completos, no importa que tan bueno sea el modelo, el resultado obtenido no será de utilidad (Render, 2012).

Algunos temas que deben resolverse son:

a) Cobertura de los datos: deben existen datos digitales registrados de todos los aspectos que se necesiten para construir un modelo aceptable. Para ello deben:
   - existir registros digitales sobre los hechos, cosas y/o personas que tienen los atributos que permiten establecer la relación de causalidad con la variable independiente,
   - se están registrando los atributos que son necesarios sobre cada hecho, cosa o persona.
   - Se están registrando con el nivel de granularidad y semántica que se necesitan.

b) Integración: para generar modelos complejos muchas veces se necesitan usar datos que provienen de distintas fuentes. Dichos orígenes deben integrarse en forma homogénea. Esta es una tarea muy compleja, máxime cuando se necesitan asegurar un alto nivel de calidad. Dicha integración por lo general ya está resuelta cuando la institución pueden montar estas soluciones sobre un data warehouse[1], pero cuando no existe, lograr esta integración puede ser el gran desafío que defina si el modelo a generar será o no factible.

c) Calidad de los datos: para tener datos "limpios" se necesita que los mismos puedan pasar por procesos de filtrado, destilado y manipulación antes de ser usados en un modelo. Pero independientemente de los algoritmos de "limpieza" que puedan usarse para emprolijar los datos, en muchos casos la calidad pasará por la decisión de que datos pueden usarse o cuáles no cumplen las condiciones mínimas necesarias de pertinencia, completitud y exactitud.

[1] *Data warehouse* es una base de datos que se diseña y administra con el objetivo de brindar información para la toma de decisiones en las organizaciones. Las mismas se actualización mediante procesos muy cuidados llamados ETL que permiten integrar datos de múltiples orígenes persiguiendo la homogeneidad y calidad de los datos (Chinkes, 2008).

Desafortunadamente la validación de los resultados de un modelo no es lo mismo que la validación de los datos que serán usados. Es decir que no sirve de nada validar los resultados, sino nos podemos asegurar la calidad de los datos que serán ingresados. Si los resultados del input son cuestionables, entonces el resultado del modelo es cuestionable.

5. UN MODELO PARA PRONOSTICAR LA DESERCIÓN

El objetivo buscado con este modelo de pronóstico es identificar la cantidad de alumnos que abandonarán sus estudios en la Facultad el año próximo. Se busca que el modelo permita individualizarlos, y ello permitirá también estimar su número total. Obtener estas estimaciones puede servir de base para distintas decisiones como la asignación de docentes y aulas, o la posibilidad de realizar intervenciones a tiempo para intentar que dicho número disminuya.

Hay experiencias interesantes realizadas en algunas universidades en otros lugares del mundo que inclusive, incluyen estas predicciones dentro de las herramientas que usan los docentes, permitiéndoles a estos conocerlo durante la propia cursada generando alertas mediante semáforos (Arnold, 2012)

5.1. Los datos

Para predecir la deserción que sucederá durante el próximo año se trabajará con los datos provenientes del data warehouse de la Universidad, que dispone de datos de los alumnos y su actividad, donde se encuentra registrados muchos años de su desempeño. Aquí se trabajará con registros a nivel individual, pero que se manejarán mediante claves que cuidan su anonimato, ya que por otro lado no aportaría nada a este trabajo disponer de la identidad de los mismos. Por otro lado se está considerando un conjunto de datos entre 2007 y el 2014.

El dataset

Para poder trabajar la herramienta y generar un modelo es necesario generar un set de datos (*dataset*) con formato de tabla. Por lo cual se generó una, en formato CVS con cabecera, que contuviera los alumnos activos por año académico. Se consideró como activos a los que tuvieron registrada al menos una actividad académica en la institución en ese año (una inscripción a la unidad académica, inscripción a materia, que rindieran un examen o hicieran la rematriculación anual que solicita la universidad).

Su clave primaria (es decir lo que identifica unívocamente cada fila) se la consideró como la concatenación entre el identificador del alumno y el año de actividad. Es decir que cada fila del *dataset* es un alumno en un año académico específico de actividad.

En la tabla también se generó un atributo calculado que se denominó "abandono", y que tiene distintos valores según correspondiera (A, B, C, E o N):

Para los que abandonaron al año siguiente del año que se está analizando (A, B o C):

> **A**: cuando no tuvo actividad al año siguiente, pero si tuvo actividad luego.
>
> **B**: cuando no tuvo actividad los dos años siguientes, pero si tuvo actividad al tercero.
>
> **C**: cuando no tuvo actividad los tres años siguientes al año analizado.

Para los que no abandonaron (E o N)

> **E**: No abandonó, porque egresó (existieron registro de que egresó en una carrera, por lo tanto es evidente que no abandonó).
>
> **N**: No desertó porque no pertenece a ninguna de las otras categorías.

Claramente lo deseable es usar solo los de la letra C, como los que desertaron, y considerar los que tienen la A y la B como del grupo de los que no abandonaron. El problema es que para aplicar dicho criterio solo puede aplicarse cuando tengo ya tres años de historia. Es por ello que se han guardado también los datos de A y B, ya que si los números son similares podrían usarse estos y tener siempre un modelo que trabaje con datos más actualizados.

La tabla obtenida tiene tantas filas como alumnos/años se estén considerando. En primera instancia se toma el rango de años 2007 al 2014 para los alumnos de la facultad de ciencias económicas (421.282 filas). En base a evaluar los datos, tal como se verá más adelante, se fue acotando para considerar solo 3 años (los años 2011 a 2013). El tomar como último año 2013 se debe a que dado que es necesario poder calcular si los alumnos abandonaron se necesita por lo menos dos años de actividad posterior, y tomar estos años daba mayor tranquilidad de esta situación (quedan por lo tanto 145.390 filas).

El resto de los atributos que se decidieron tomar para conformar la tabla para cada alumno son:

- Del perfil del estudiante: edad para dicho año, sexo, localidad de residencia, horas que trabaja y máximo nivel de estudio de padres.
- Años: año analizado, cantidad de años desde que inicio al carrera (año de ingreso ->permanencia)
- Desempeño del alumno: cantidad de materias aprobadas desde el inicio, cantidad de materias aprobadas en el año y promedio desde el inicio.

Variable dependiente:

Para poder trabajar con un modelo de clasificación, se genera un atributo que será la variable dependiente. El mismo se denominó "abandono (2)", cuyos posibles valores serán "S" (cuando abandonó) o "N" (cuando no lo hizo). Más adelante durante la generación del modelo se explicará cómo se generó dicho atributo tomando como base el del atributo abandono.

5.2. Generación del modelo de pronóstico

Si bien existen distintas herramientas de software que permiten generar modelos predictivos como los descriptos (como RapidMinner, Weka, etc.) se ha decidido usar un servicio en la nube que ofrece Microsoft denominado Azure *Machine learning*. El mismo, al igual que los otros mencionados, puede ser usado en forma gratuita.

Esta herramienta, por su parte es un buen ejemplo, para demostrar varios de los argumentos planteados en este trabajo en forma previa donde se plantea como existe una tendencia a facilitar su uso a un número mayor de personas no expertas en los algoritmos y en herramientas informáticas, y cómo la infraestructura de procesamiento y almacenamiento deja de ser un limitante (ya que todo corre en la nube).

En primer lugar se creará dentro de la plataforma, un nuevo "Experimento", dentro del mismo es donde se armará, eligiendo las opciones que ofrece la herramienta, el experimento que permita generar el modelo predictivo que se pretende.

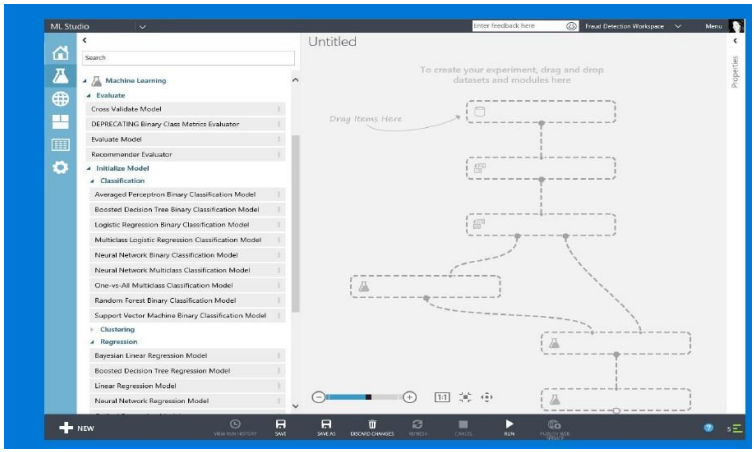

Figura 1. Pantalla de Azure *Machine learning* para un nuevo experimento

Dado que el primer paso del modelo es tomar los datos del *dataset* que se había conseguido, será necesario importar dicho *dataset* para que quede guardado como una de las "fuentes de datos" que estén disponibles desde el entorno en la nube. Para ello se hace la carga, a través de la opción correspondiente, tomando el archivo que disponíamos guardado en forma local (en formato CVS).

Luego dentro del experimento creado, que en este caso se ha denominado "Abandono de alumnos", se arrastra esta fuente de datos que creamos a la zona de la herramienta donde se arma el experimento. Esta opción, como el resto de las que deberán realizarse, se hacen arrastrando los elementos existentes desde el panel de la izquierda al sector donde se arma el experimento (del lado derecho).

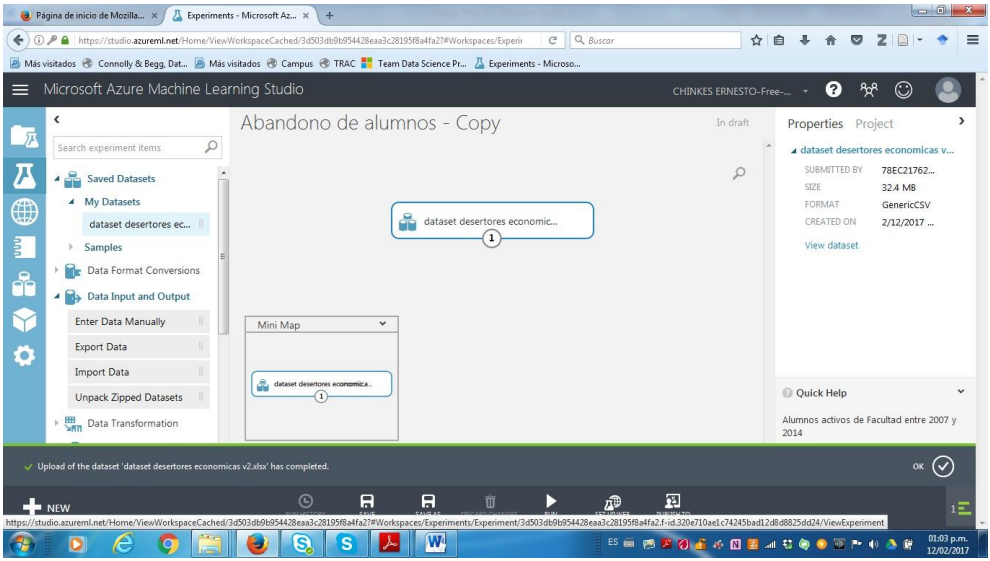

Figura 2. Pantalla de Azure *Machine learning* incorporando *dataset* en el experimento

Una vez ubicado el *dataset* en el panel del experimento, pueden visualizarse los valores eligiendo dicha opción, tal como se ve en la siguiente figura.

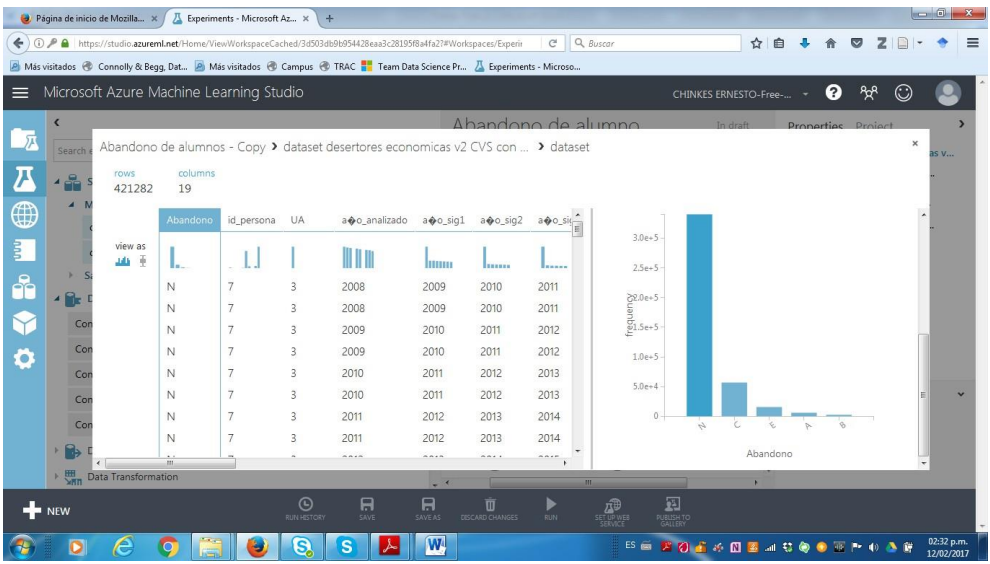

Figura 3. Pantalla de Azure *Machine learning* donde se visualizan los valores del *dataset*

Esta visualización sirve para evaluar que manipulaciones de datos realizar, en forma a las acciones más relacionadas con el modelo; así como tomar decisiones sobre atributos y filas que se usarán. En nuestro caso, a partir de visualizar los valores en los distintos atributos se fueron tomando las siguientes decisiones:

a) dejar los años expuestos (2011 a 2013),
b) sacar aquellos atributos que no parecían muy confiables o que no servirían para generar el modelo (id_persona, año de egreso, año de emisión, provincia del domicilio, etc.),
c) detectar algunas filas duplicadas y eliminarlas. También emprolijar aquellos alumnos que no tenían un año de ingreso definido.
d) Se creó el atributo permanencia (año analizado menos año de ingreso), dejando luego afuera el año de ingreso.
e) evaluar los porcentajes de la variable "abandono", en base a lo cual se decidió la cantidad de dos años sin actividad como criterio para la deserción.

Para ello se fueron eligiendo distintas opciones de manipulación de datos (para seleccionar columnas, seleccionar filas, y generar algunas transformaciones), que no se explicarán en detalle en este trabajo.

Luego se debe elegir un algoritmo, dentro del tipo de tarea elegida. En nuestro caso es la "clasificación", y los algoritmos disponibles se encuentran en "Inicialice Model->Classification".

En el siguiente gráfico pueden observarse los distintos tipos de algoritmos que ofrece el servicio según el tipo de tarea a desarrollar en el modelo.

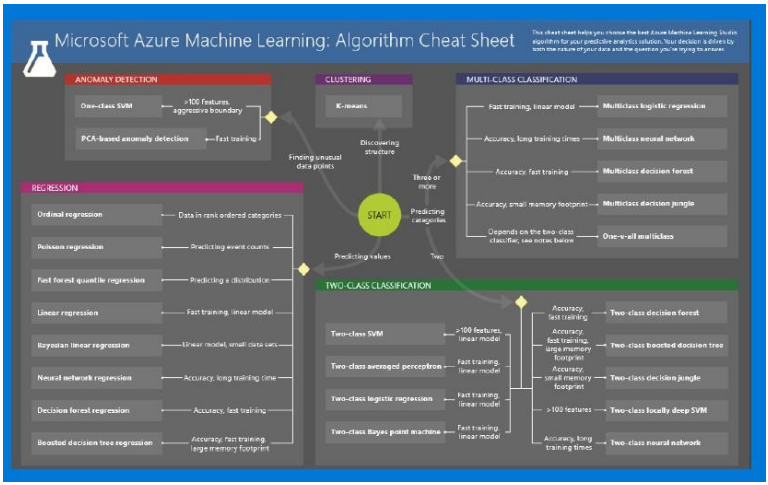

Figura 4. Algoritmos que ofrece Azure *Machine learning* para distintas tareas

En nuestro caso la primera opción fue probar con "Two-Class Bayes Point Much".

Dado que se están usando modelos supervisados es necesario dividir las filas del *dataset*. Una porción de las filas se usarán para el entrenamiento (en este caso se eligió configurarlo para que use el 70 % de los datos), y la otra parte para la evaluación, donde se usará el 30 % restante de las filas del *dataset*. Eso se hace a través de la opción: "Data tranformation->Sample and Split->Split Data"

El modelo va a "entrenar" mediante la opción "*Machine learning*->Train->Train Model", usando el algoritmo elegido previamente y con la porción del *dataset* destinada para ello (70 %). En esta opción de "Train Model" debe indicarse cuál es el atributo que se va a predecir. En nuestro caso "abandono (2)".

Cabe aclarar que este atributo Abandono (2), se generó mediante una opción previa que se eligió que permite la agrupación de los valores que existen en un atributo (atributo "abandono"). En el atributo abandono, tal como se había mencionado podrían existir los valores A, B, C, E, o N. En el atributo Abandono (2), se pusieron como S a los valores C y B (considerando que son los que abandonaron), al resto como N (es decir que no abandonaron). Se tomaron los valores C y B, ya que pareció razonable considerar un horizonte de dos años para asumir que al no tener actividad se lo consideraba como un alumno que había desertado. Por otro lado observando los datos, el porcentaje de aquellos que luego de dos años sin actividad, volvían a tener actividad en el tercer año era prácticamente insignificante (menos del 1%).

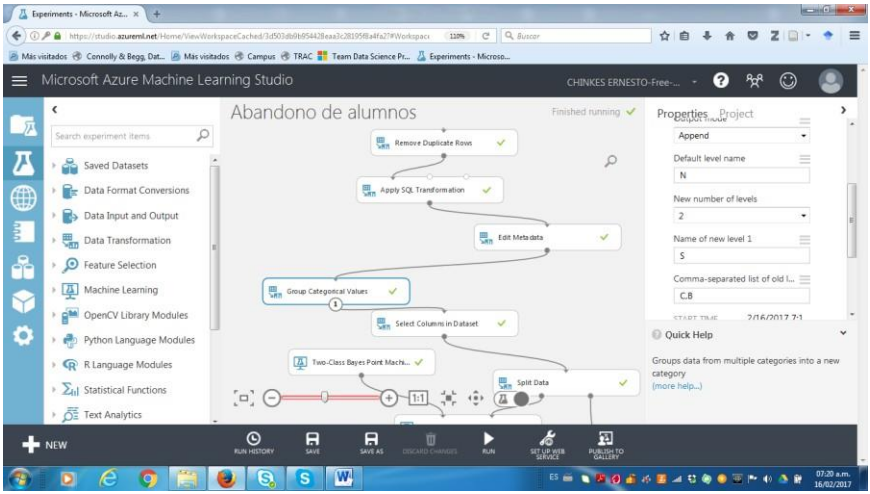

Figura 5. Pantalla de Azure *Machine learning* donde se agregaron opciones de manipulación de datos, el algoritmo del modelo y la opción de Split data.

El resultado del modelo de entrenamiento, es el que se pasa al de testeo usando el set de datos destinados a la evaluación (30 % restante). Para ello y obtener el score se usa la opción "*Machine learning*->Score->Score Model". En la figura puede verse como quedan relacionadas las opciones.

Figura 6. Pantalla de Azure *Machine learning* donde se ve como se incorporan y relacionan el algoritmo con las opciones de entrenamiento y score.

Por último hay que agregar la opción de "*Machine learning*->Evaluate->Evaluate Model" para evaluar la precisión de nuestro modelo, en base a los datos obtenidos del score.

A partir del resultado inicial, se fueron agregando y cambiando atributos, y probando distintos algoritmos de clasificación, hasta llegar al que finalmente se muestra.

Es posible trabajar en un modelo con dos algoritmos a la vez, con el objetivo de compararlos. A continuación se muestra en la figura el modelo cuando al algoritmo anterior se lo comparó con el de árboles de decisión.

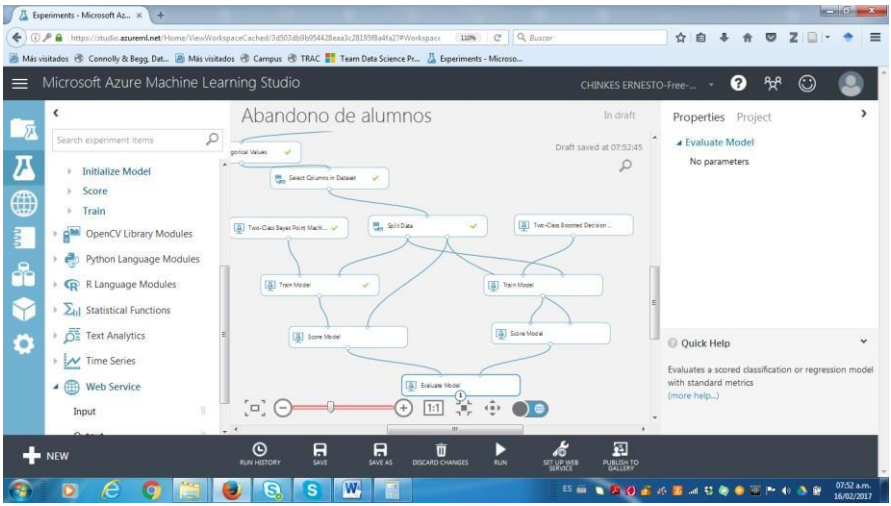

Figura 7. Pantalla de Azure *Machine learning* donde se evalúan dos algoritmos entre sí en el mismo experimento

Este segundo algoritmo dio mejores métricas que el primero, y se decidió luego probar con un tercer algoritmo (redes neuronales), que dio mejores resultados que los dos primeros, siendo el que finalmente se dejó.

Una vez, que luego de este proceso iterativo de búsqueda de los mejores resultados de eficacia del modelo, se encuentra el adecuado la herramienta permite publicar el mismo y dejarlo operativo (corriendo en la nube mediante un servicio web). Este servicio lo que hace es que cuando se le introducen los valores de las variables independientes del modelo (por ejemplo para los alumnos del año corriente), se pueda estimar el valor de la variable dependiente (si abandonará o no el año próximo).

Tener un servicio web disponible en la nube también habilita a que el modelo construido pueda ser usando incorporándolo en algún sistema que esté conectado a Internet, y predecir en tiempo real la variable ante el input de datos que le envíe el propio sistema, y de esta manera ejecutar una acción o recomendación en base al resultado.

5.3. Resultados

Tal como se mencionó, el modelo generado termina con una opción de evaluación que permite medir la eficacia del modelo. Desde la opción de evaluación de modelos puede verse las métricas y evaluar su desempeño. En el caso de la clasificación binaria (que es la que aquí se usó) existirán resultados con una etiqueta denominada positiva, y el

resto de los resultados con la que se denomina negativa. En nuestro caso los que tienen la etiqueta positiva son los que no abandonaron, y los alumnos que abandonaron tendrán la etiqueta negativa. Por otro lado existirán predicciones que durante el score acertaron con su etiqueta (verdaderos), y otros que no lo hicieron (falsos).

Las métricas que muestra para los modelos de clasificación, son las siguientes:

- **Accuracy**: es las medida de bondad de la clasificación, tomando la proporción de resultados verdaderos sobre el total (positivos verdaderos + negativos verdaderos) / total).
- **Precision:** es la proporción de resultados verdaderos positivos sobre todos los resultados positivos (positivos verdaderos) / (positivos verdaderos + positivos falsos)
- **Recall:** es la fracción de los resultados correctos devueltos por el modelo. (verdaderos positivos / (total de eventos positivos: verdaderos positivos + falsos negativos)
- **F-score:** es el peso promedio de precision y recall entre 0 y 1, donde el valor ideal es 1.
- **AUC:** mide el área debajo de la curva con positivos verdaderos en el eje Y, y falsos positivos en el eje X. Esta métrica es útil porque provee un sólo número que permite comparar modelos de diferentes tipos.

A continuación puede verse el resultado que arrojo el modelo para el primer algoritmo: Two-Class Bayes Point Much, pero habiendo realizando la corrida con un solo año: 2011.

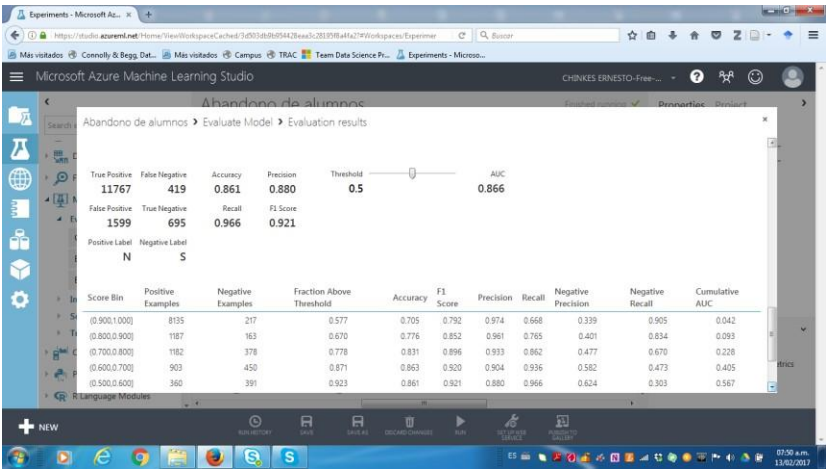

Figura 8. Pantalla de Azure *Machine learning* donde muestra las métricas de la evaluación algoritmo two-class Bayes Point Much

Luego se agregaron más años para la corrida (2011 a 2013) y se probaron los otros algoritmos, tal como se explicó previamente, siendo las redes neuronales los mejores resultados, por lo tanto se eligió este último.

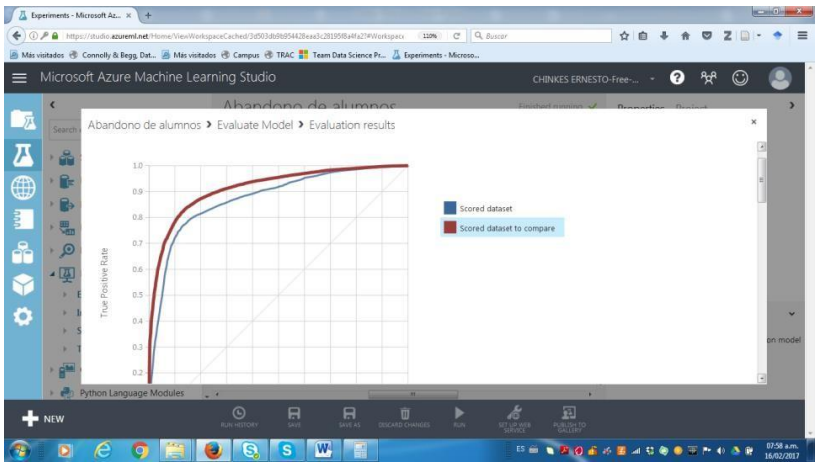

Figura 9. Pantalla de Azure *Machine learning* donde compara gráficamente las curvas de los algoritmos Two-class Neural network con el de Two-class Boosted Decision Tree

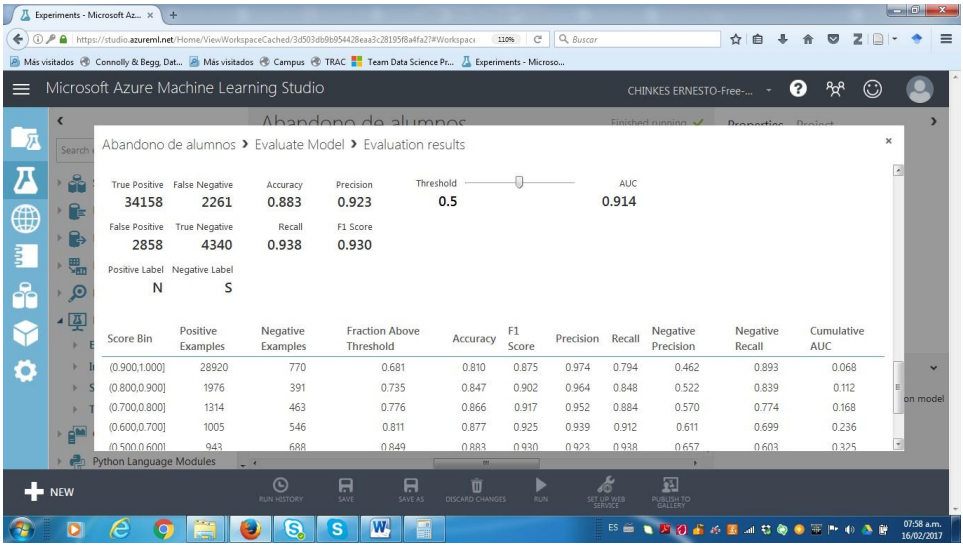

Figura 10. Pantalla de Azure *Machine learning* donde muestra las métricas de la evaluación del algoritmo

Two-class Neural network.

La muestra para la evaluación es de 43.617 filas, que es el 30 % de las filas que contenía el *dataset* para los años 2011 a 2013. La herramienta en forma automática toma uno de los valores de la variable independiente "abandono (2)" como etiqueta positiva y la otra como negativa. En este caso considera la etiqueta positiva a la "N" (que es cuando los alumnos permanecen el próximo año) y la etiqueta negativa la "S" (que es cuando abandonan).

Las métricas para el modelo final que usó el algoritmo de redes neuronales son las siguientes:

- **Accuracy**: (verdaderos positivos + negativos verdaderos) / total):
  34.158 + 4.340 / 43.617=**0,883**
- **Precision:** (verdaderos positivos) / (verdaderos positivos + falsos positivos)
  34.158/(34.158+2.858)=**0,923**
- **Recall:** (verdaderos positivos / (verdaderos positivos + falsos negativos))
  34.158/(34158+2261)=**0,938**
- **F1Score=0,930**

El modelo también detectó, entre 43.617 estudiantes considerados en la evaluación, que 6.601 (4.340 negativos verdaderos + 2.261 negativos falsos) abandonarían al año siguiente (un 15% del total). El total de real de abandono fue del 16 %, con 7.198 alumnos (4.340 negativos verdaderos + 2.858 positivos falsos). Es decir, que tomando este criterio la estimación tiene más del 91% de acierto respecto de estos casos.

Las métricas vistas previamente (Accuracy, precisión, Recall y F1Score) describen al modelo como uno con un muy buen desempeño, pero cabe destacar que esas métricas están más enfocadas en los valores de la etiqueta positiva. Para el caso que nos ocupa, que es predecir que alumnos podrían abandonar el año siguiente, y poder individualizarlos, es interesante alertar que de los 6.601 casos que predijo el modelo, acertó en un 66% (medida de precisión pero para los casos negativos), ya que 4.340 efectivamente lo terminarían haciendo (negativos verdaderos) y 2.261 no (negativos falsos), es decir que el resto fue compensado por los positivos falsos. Es decir, que existe un 34% de casos de abandono de la predicción que serían erróneos.

6. CONCLUSIÓN

La toma de decisiones necesita realizar estimaciones sobre el futuro, ya que dichos pronósticos explican el contexto en el que se basarán las decisiones. Es decir, que la mayoría de las decisiones asumen predicciones (explícitas o implícitas), ya sea que se estimen mediante métodos cualitativos, cuantitativos o sin método alguno.

El modelo y la herramienta planteada en este trabajo son un ejemplo de cómo los métodos cuantitativos, apoyados por las tecnologías de la información y en la era del *big data*, pueden ser una ayuda muy importante para la toma de decisiones; siempre que quienes lo usen entiendan sus limitaciones y el significado de sus predicciones.

Predecir la deserción de los estudiantes, en la medida que debiera tomarse una decisión como la asignación de tutores a estudiantes en riesgo (con el objeto de minimizar el abandono), podría optarse por hacerlo al 15 % del total que asigna el modelo. Al tomar esta decisión debe tenerse conciencia que se asignarán tutores a estudiantes que finalmente no desertarían, y que por otra parte, no se asignarán tutores a otros que también estaban en riesgo pero que no fueron detectados por el modelo.

La idea es intentar modelos con el menor nivel de error posible y tener conciencia de las limitaciones de los métodos de predicción. En nuestro ejemplo es posible mejorar la métrica de precisión (para las etiquetas negativas), tratando de aumentar el 0,66, mediante el uso de otros algoritmos y también ajustando los parámetros de los mismos. También puede mejorar incluyendo otras variables de datos disponibles que pueden ser relevantes, como el porcentaje de desocupación en el año, el porcentaje de crecimiento o decrecimiento del PBI, cantidad de materias cursadas simultáneamente en el año, etc.

Respecto de las limitaciones es interesante destacar que, más allá de que se logren modelos que detenten métricas muy altas como una precisión del 0,90, 0,95, o más, sería peligroso que la institución concluyera que tiene la forma de saber anticipadamente cual será el destino de sus estudiantes el año próximo, ya que existen muchas variables que no se están tomando en cuenta, y porque las mismas en el futuro podrían desempeñarse en forma dispar a lo que lo hicieron en el pasado.

En todo método de pronóstico las variables dependientes que se van a predecir, se ven influenciadas por infinidad de otras variables que pudieron no ser contempladas, inclusive porque muchas de ellas pudieron ser imperceptibles en el pasado. Sin embargo el *big data*, con su capacidad de almacenar datos de cada vez más eventos, cosas y personas, y en mayores cantidades; y el avance de la computación y la nube para aumentar la capacidad de procesar algoritmos complejos cada vez más rápido, obliga a repensar como pueden ayudar las soluciones informáticas de *data mining* y *machine learning*; sobre todo cuando las mismas pueden quedar incluidas dentro de los procesos. Ello permitirá alertar, hacer recomendaciones e inclusive, en algunos casos, tomar la propia decisión. Todo esto a sabiendas, que tal como sucede con los humanos, podrán cometerse algunos errores.

BIBLIOGRAFIA

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2011). Métodos cuantitativos para los negocios (11a. ed.). México, D.F.: CENGAGE Learning.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: using learning analytics to increase student success. Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 267–270). ACM.

Chinkes, E. (2008). Business Intelligence para mejores decisiones de negocio. Buenos Aires: EDICON.

Elias, T. (2011). Learning analytics: Definitions, Processes and Potential.

Hernández Orallo, J., Ramírez Quintana, M. J., & Ferri Ramírez, C. (2004). Introducción a la minería de datos. Madrid: Pearson Prentice Hall.

Hodgkinson, G. P., & Starbuck, W. H. (Eds.). (2008). The Oxford handbook of organizational decision making. Oxford ; New York: Oxford University Press.

Mannila, H. (1996, June). Data mining: *machine learning*, statistics, and databases. In Scientific and Statistical Database Systems, 1996. Proceedings., Eighth International Conference on (pp. 2-9). IEEE.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of machine learning. Cambridge, MA: MIT Press.

Pavesi, P. F. J., Bonatti, P., & Avenburg, D. (2004). La decisión: su teoría y práctica: aplicaciones conceptuales, casos. Buenos Aires: Grupo Editorial Norma.

Pérez López, C., & Santín González, D. (2008). Minería de datos: técnicas y herramientas. Madrid: Paraninfo Cengage Learning.

Render, B., Stair, R. M., & Hanna, M. E. (2012). Quantitative analysis for management (11th ed). Upper Saddle River, N.J: Pearson Prentice Hall.