Modelo predictivo para determinar la graduación de alumnos de carreras de ingenierías aplicando técnicas de minería de datos

Claudio José Carrizo
cjcarrizo77@gmail.com / Universidad Tecnológica Nacional San Francisco
San Francisco, Argentina

Recepción: 1-7-2018 / Aceptación: 21-8-2018

Resumen. La ingeniería cumple un rol fundamental en el desarrollo económico y el bienestar social de un país. Por este motivo, en el 2012 el gobierno nacional de Argentina impulsó el "Plan Estratégico de Formación de Ingenieros 2012-2016 (PEFI)", con el objetivo de incrementar la cantidad de graduados en carreras de ingeniería, en pos de mejorar la industria, la innovación productiva y la expansión económica de dicho país. El propósito del presente trabajo es construir un modelo predictivo a través de técnicas de minería de datos que permita, por un lado, determinar la cantidad de alumnos que pueden graduarse en carreras de ingenierías de UTN Facultad Regional San Francisco, y por otro lado, identificar patrones que puedan incidir en la graduación. Los resultados de este proyecto representarán un aporte a la gestión académica en lo que respecta a la planificación, el seguimiento y el control de las cohortes de alumnos de las carreras de ingeniería.

Palabras clave: modelo predictivo, minería de datos, graduados, ingeniería, UTN

Predictive model for determining students' graduation from engineering undergraduate programs using data mining techniques

Abstract. Engineering plays a fundamental role for the economic development and social welfare of a country. For this reason, the National Government promoted in 2012 the "Plan Estratégico de Formación de Ingenieros 2012-2016 (PEFI)", aiming to increase the number of graduates in engineering careers, in order to improve industrial development, productive innovation and economic expansion of Argentina. The purpose of this work is to build a predictive model through data mining techniques that allow, on one hand, determine the number of students who can graduate in engineering careers at UTN Facultad Regional San Francisco, and, on the other hand, identify patterns that may affect graduation. The results of this project will represent a contribution for the area of academic management, specifically for planning, monitoring and control keeping of engineering student cohorts.

Keywords: predictive model, data mining, graduates, engineering, UTN

1. Introducción

Desde la década de 1960, la población mundial ha crecido en forma exponencial. Este crecimiento ha generado problemas en sistemas energéticos, sanitarios, de telecomunicaciones e infraestructura (Naciones Unidas, 1987).

La ingeniería cumple un papel fundamental para el desarrollo económico y el bienestar social de la sociedad (Bohórquez y Torres, 2016). En el ámbito de la República Argentina, será necesario contar con una mayor cantidad de graduados en carreras de ingeniería para incrementar el desarrollo industrial, la innovación productiva y la expansión económica. Aquellos países que apuesten a la formación de más y mejores ingenieros, estarán apostando a la industrialización y el desarrollo.

Cifras referidas por el Colegio de Ingenieros de la Provincia de Buenos Aires (2016) indican que en China hay un ingeniero por cada 2000 personas; en Francia o Alemania uno por cada 2300; en México o Chile uno por cada 4500; en Brasil uno por cada 6000, mientras que en Argentina la proporción es de uno por cada 6700 habitantes.

En Argentina se dictan 481 carreras de ingeniería; 383 en universidades públicas y 98 en universidades privadas. A partir de un informe realizado por la Secretaría de Políticas Universitarias, se consigna que la UTN forma al 42,75 % de los ingenieros que se gradúan en el país (UTN Facultad Regional Buenos Aires, 2014). Esta cifra marca la importancia relativa que tiene la UTN respecto de las demás universidades que cuentan con carreras de ingeniería, posicionándose como la universidad de ingeniería más grande de la Argentina.

El desarrollo industrial y tecnológico del país necesita ingenieros capaces de liderar las innovaciones de los próximos treinta años. Es por este motivo que la Secretaria de Políticas Universitarias (2012) crea el Plan Estratégico de Formación de Ingenieros 2012-2016 (PEFI), como un compromiso del Ministerio de Educación de la Nación para incrementar la cantidad de graduados en ingeniería en un 50 % en 2016 y en un 100 % en 2021; esto permitiría de alguna manera asegurar en cantidad y calidad los recursos humanos necesarios para apoyar los ejes del Plan Estratégico 2020 con el fin de hacer de Argentina un país desarrollado. El PEFI pretende colocar a la Argentina entre los países con mayor cantidad de graduados en ingenierías de Latinoamérica y para ello se propuso lograr que exista un ingeniero por cada 4000 habitantes. Para llevar a cabo el PEFI se plantearon objetivos sobre la base de tres ejes estratégicos: "Mejora de indicadores académicos", "Aporte al desarrollo territorial sostenible" e "Internacionalización de las ingenierías".

En el 2016 la Secretaría de Políticas Universitarias (SPU) publicó una estadística acerca de la cantidad de graduados de ingeniería en donde se establece que se graduaron 7470 ingenieros de diferentes especialidades. Según el lema del PEFI, para asegurar un desarrollo sostenible del modelo productivo y del sistema científico y tecnológico, se necesita alcanzar una graduación de alrededor de 10 000 ingenieros por año, proyección que se definió teniendo en cuenta las

actuales necesidades insatisfechas y las áreas aún no desarrolladas que se desearía consolidar. Por lo tanto, de acuerdo con la estadística proporcionada por la SPU, queda en evidencia que los intentos del gobierno nacional por incrementar las tasas de graduación en carreras de ingenierías no trajeron los resultados esperados (Fernández, 2018).

En el ámbito local, más precisamente en la ciudad de San Francisco (provincia de Córdoba), se emplaza la Facultad Regional San Francisco, una de las 32 facultades regionales que forman parte de la Universidad Tecnológica Nacional. En este contexto, en el periodo 2015-2017 se llevó a cabo la ejecución de un Proyecto de Investigación y Desarrollo (PID), que fue evaluado externamente y homologado por la UTN según disposición SCTyP 380/15.

Durante la ejecución del PID, más precisamente en la elaboración del estado del arte, pudimos detectar que existen muchos trabajos de investigación relacionados con la temática que expone este trabajo, solo que la mayoría de ellos están enfocados en la construcción de modelos que permitan determinar la deserción estudiantil o el rendimiento académico de alumnos universitarios, través del uso de técnicas de minería de datos (Fisher, 2012; Valía *et al*., 2017; La Red, Karanik, Giovannini y Scappini, 2009; Porcel, Dapozo y López, 2016).

El objetivo de este trabajo consiste en determinar la cantidad de alumnos que pueden graduarse en un plazo de tiempo de ocho años en las carreras de ingeniería que se dictan en la UTN Facultad Regional San Francisco (ingenierías de Sistemas de Información, Química, Electrónica y Electromecánica), e identificar los patrones que pueden incidir en la graduación de estos alumnos, tomando como año de ingreso el 2012 y como año de egreso el 2018.

Para lograr este objetivo se propuso la construcción de un modelo predictivo a través de la utilización de técnicas de minería de datos (Pérez, 2015), que permita a través de los resultados pertinentes, no solo determinar la cantidad de alumnos que pueden graduarse en un plazo de ocho años, sino también detectar los patrones que inciden en la graduación de un alumno de ingeniería.

Los resultados de este proyecto contribuyen en forma de un instrumento útil para el área de gestión académica de la UTN Facultad Regional San Francisco, ya que permitirán proveer información acerca de la cantidad de alumnos que pueden graduarse en carreras de ingeniería en un tiempo promedio de ocho años y además brinda la posibilidad de poder trabajar específicamente sobre los patrones que inciden en la graduación de alumnos en las carreras de ingeniería para poder mejorar las tasas de graduados en los próximos años.

2. Metodología

Para guiar el desarrollo del proyecto, se utilizó la metodología CRISP-DM (Goicochea, 2009) para minería de datos, en donde se llevaron a cabo en una primera instancia las fases de comprensión y preparación de los datos y luego, en una segunda instancia, se desarrolló la fase de modelado.

En cuanto al diseño de la investigación, en primera medida se precedió a relevar, analizar y clasificar los datos obtenidos a partir del sistema de gestión académica SysAcad, a fin de poder construir el perfil del alumno de ingeniería de UTN, teniendo en cuenta sus aspectos personales, laborales, académicos y socioeconómicos. Por otra parte, se precedió a relevar, analizar, evaluar y seleccionar las técnicas de minería de datos que más se adapten para armar el modelo predictivo que permita obtener los resultados antes mencionados.

2.1 Técnica de recolección de datos

Los datos correspondientes a alumnos de carreras de ingenierías de UTN Facultad Regional San Francisco se recolectaron a través de archivos en formato Excel separados por comas (CSV), en donde cada uno de los archivos contenían los atributos y registros de alumnos de carreras de ingenierías de la Facultad Regional San Francisco, en el periodo 1970-2018, que consiguieron su graduación o no.

En lo que respecta a las técnicas y herramientas de minería de datos a ser caracterizadas, se utilizó la técnica de análisis documental a través de relevamiento bibliográfico, publicaciones, *papers*, etcétera.

2.2 Técnicas de procesamiento y análisis de datos

Para el análisis y procesamiento de los datos relevados se utilizó la herramienta de minería de datos *open source* denominada *RapidMiner Studio*[1], lo que permitió no solo analizar y procesar los datos, sino también construir y evaluar el modelo predictivo a través de la técnica de minería de datos denominada árboles de decisión.

3. Fases de comprensión y preparación de datos

A través de la metodología CRISP-DM se realizó la fase de compresión y preparación de los datos. En primera instancia, se construyó el perfil del alumno de carreras de ingenierías a través de datos personales, laborales, académicos y socioeconómicos, en donde se especificó un conjunto de atributos iniciales que se pueden visualizar en la tabla 1.

1. Ver *Data Science Behind Every Decision* en https://rapidminer.com/ y *Analytics* en https://www.microsystem.cl/plataforma/rapidminer/

Tabla 1
Atributos iniciales del alumno de la carrera de ingeniería

Atributo	**Tipo de valores**
Legajo	Numérico (6)
Fecha nacimiento	Date (dd/mm/aaaa)
Sexo	String
Nombre estado civil	String
Procedencia cuidad	String
Procedencia provincia	String
Residencia ciudad	String
Residencia provincia	String
Ocupación	String
Trabaja	String
Cantidad familiares	Numérico (2)
Cantidad hijos	Numérico (2)
Año ingreso	Numérico (4)
Cantidad materias aprobadas	Numérico (4)
Cantidad materias regularizadas	Numérico (4)
Aplazos	Numérico (4)
Promedio	Decimal (2)
Instrucción padre	String
Instrucción madre	String
Ocupación padre	String
Ocupación madre	String

Elaboración propia

Luego se llevó a cabo la descripción y exploración de los datos a través de la herramienta RapidMiner Studio. Por último, para concluir la etapa de comprensión de datos, se realizó una verificación de la calidad de los datos relevados a través de los siguientes criterios, medidos en porcentajes: datos perdidos, errores de datos, errores de mediciones, incoherencias de codificación y metadatos erróneos.

En la fase de preparación de los datos se llevó a cabo la selección de los datos relevantes de entrada no solo para el modelo de entrenamiento sino también para el modelo predictivo; a esto también se lo conoce como vista minable o *dataset*. La metodología CRISP-DM propone

dos formas de selección de datos relevantes para objetivos de minería de datos: una de ellas es la selección de elementos (filas) y la otra es la selección de atributos o características (columnas). En este trabajo la selección de datos relevantes de entrada se realizó tomando como criterio los atributos iniciales que han pasado por un proceso de verificación de calidad de datos cuyos resultados de verificación han sido atributos con nivel de calidad de datos aceptable. Cabe destacar que se han descartado atributos relacionados con el anonimato del alumno como, por ejemplo, legajo, ciudad, provincia.

Los datos de entrada seleccionados (también conocidos como vista minable o *datatset*) se pueden visualizar en la tabla 2.

Tabla 2
Atributos seleccionados del alumno de la carrera de ingeniería

Atributo	**Tipo de valores**
Edad	Numérico (2)
Sexo	String
Estado civil	String
Trabaja	String
Cantidad familiares	Numérico (2)
Cantidad hijos	Numérico (2)
Año ingreso	Numérico (4)
Cantidad materias aprobadas	Numérico (4)
Cantidad materias regularizadas	Numérico (4)
Aplazos	Numérico (4)
Promedio	Decimal (2)

Elaboración propia

Por último, para concluir la fase de preparación de datos se procedió a realizar limpieza y formateo de los datos de entrada seleccionados.

4. Fase de modelado

En la fase de modelado se llevó a cabo la construcción de un modelo predictivo a través de la herramienta RapidMiner Studio. Para ello se elaboró en primera instancia un modelo de entrenamiento (figura 1) en donde se tomó como entrada de datos un dataset con atributos

seleccionados en la fase de preparación de datos, los cuáles contenían datos históricos de alumnos ingresantes a carreras de ingeniería en la UTN Facultad Regional San Francisco entre los años 1970 y 2011. Cabe destacar, que a este dataset se le agregó el atributo "Graduado", que indica si el alumno consiguió su graduación o no durante el periodo mencionado.

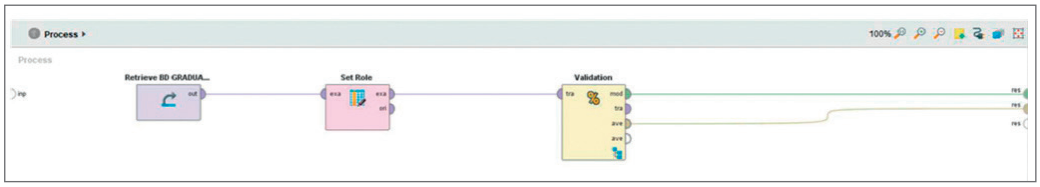


Figura 1. Modelo de entrenamiento

Elaboración propia

Para realizar la validación de los datos de entrada se utilizó el operador VALIDATION SPLIT (figura 2), dentro del mismo se realizó la fase de entrenamiento (training) y prueba (testing) del modelo.

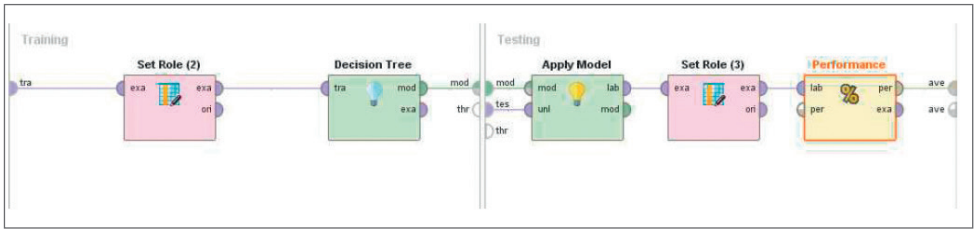
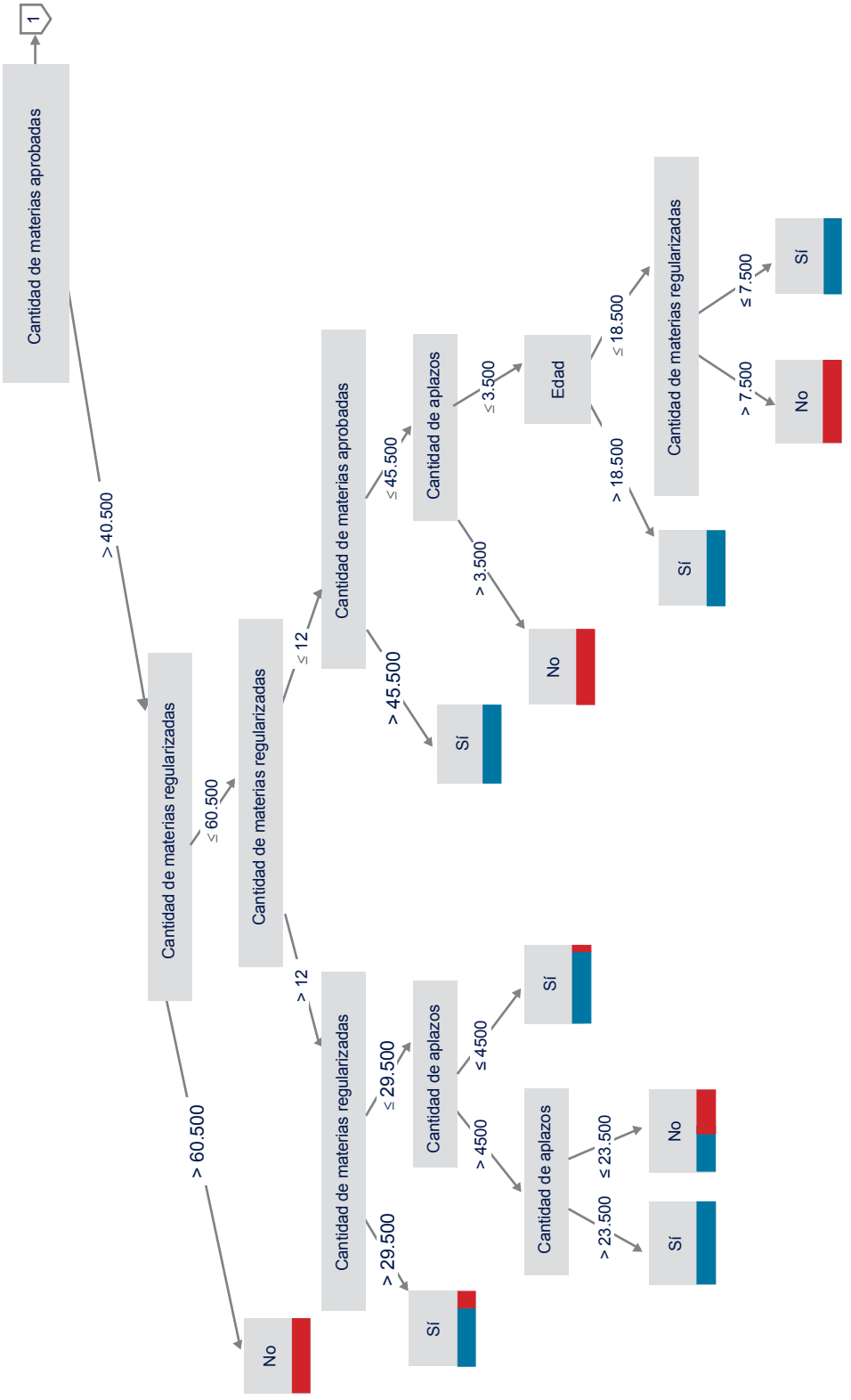


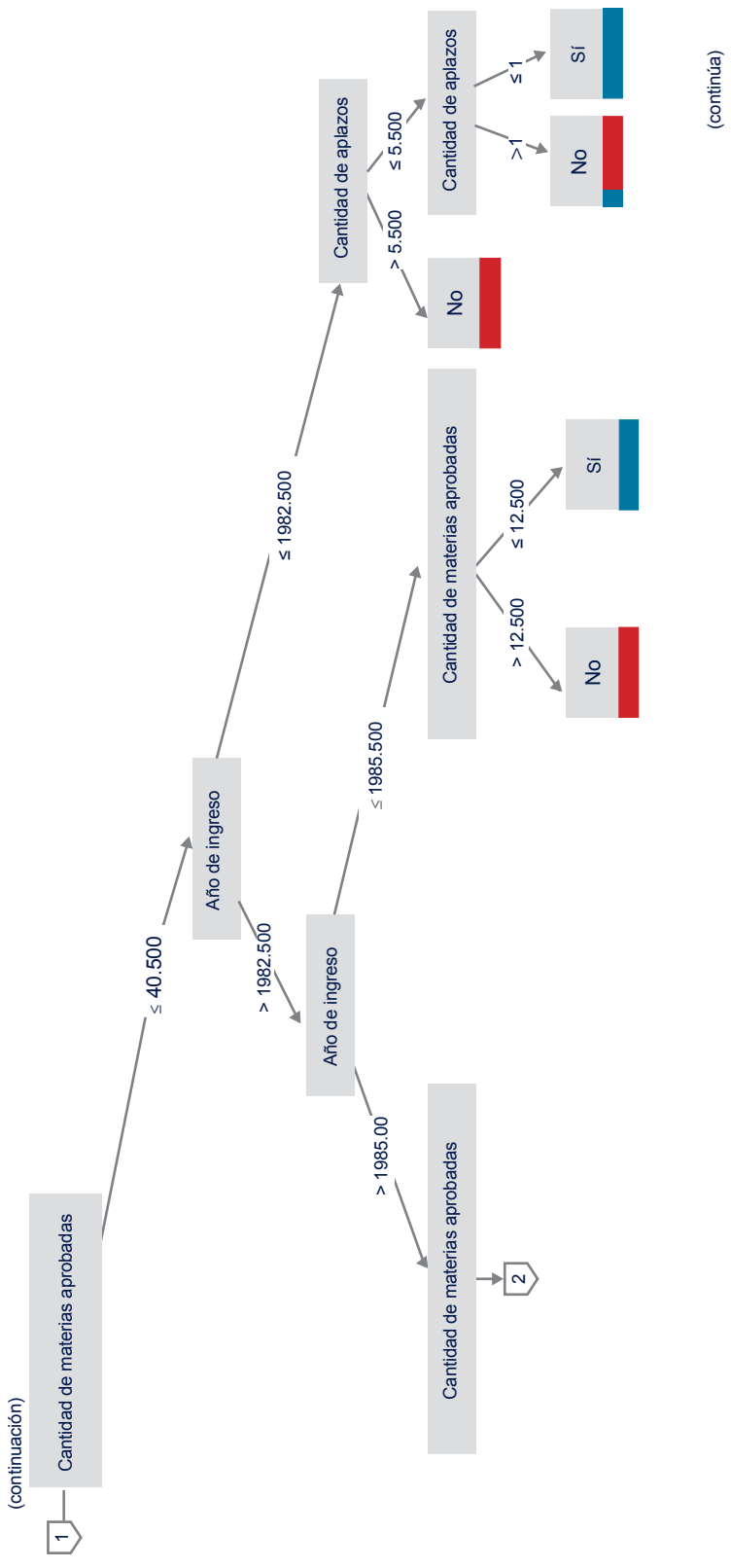
Figura 2. Modelo de entrenamiento (operador VALIDATION SPLIT)

Elaboración propia

En la fase de entrenamiento se utilizó la técnica de minería de datos denominada árboles de decisión (figura 3), seleccionada luego de realizar un relevamiento bibliográfico de diferentes técnicas como análisis de varianza, análisis discriminante, árboles de decisión, modelos de regresión, redes neuronales y series temporales, en donde se caracterizó a cada una de ellas según diferentes criterios, como en qué consiste la técnica, cómo funciona, cómo se aplica, etcétera. Luego se elaboró una matriz de doble entrada (criterios/técnicas), en donde se realizó una comparación entre las distintas técnicas relevadas para luego seleccionar la técnica más adecuada para este proyecto. Luego de la comparación realizada se seleccionó la técnica de árboles de decisión como la más adecuada para aplicar al modelo predictivo de este trabajo.


Cantidad de materias aprobadas
1
> 40.500
Cantidad de materias regularizadas
> 60.500
No
≤ 60.500
Cantidad de materias regularizadas
> 12
Cantidad de materias regularizadas
> 29.500
Sí
≤ 29.500
Cantidad de aplazos
> 4500
Cantidad de aplazos
> 23.500
Sí
≤ 23.500
No
≤ 4500
Sí
≤ 12
Cantidad de materias aprobadas
> 45.500
Sí
≤ 45.500
Cantidad de aplazos
> 3.500
No
≤ 3.500
Edad
> 18.500
Sí
≤ 18.500
Cantidad de materias regularizadas
> 7.500
No
≤ 7.500
Sí

(continúa)


(continuación)
1
Cantidad de materias aprobadas
≤ 40.500
Año de ingreso
> 1982.500
Año de ingreso
≤ 1982.500
Cantidad de aplazos
≤ 5.500
Cantidad de aplazos
≤ 1
Sí
> 1
No
> 5.500
No
≤ 1985.500
Cantidad de materias aprobadas
≤ 12.500
Sí
> 12.500
No
> 1985.00
Cantidad de materias aprobadas
2
(continúa)

(continuación)

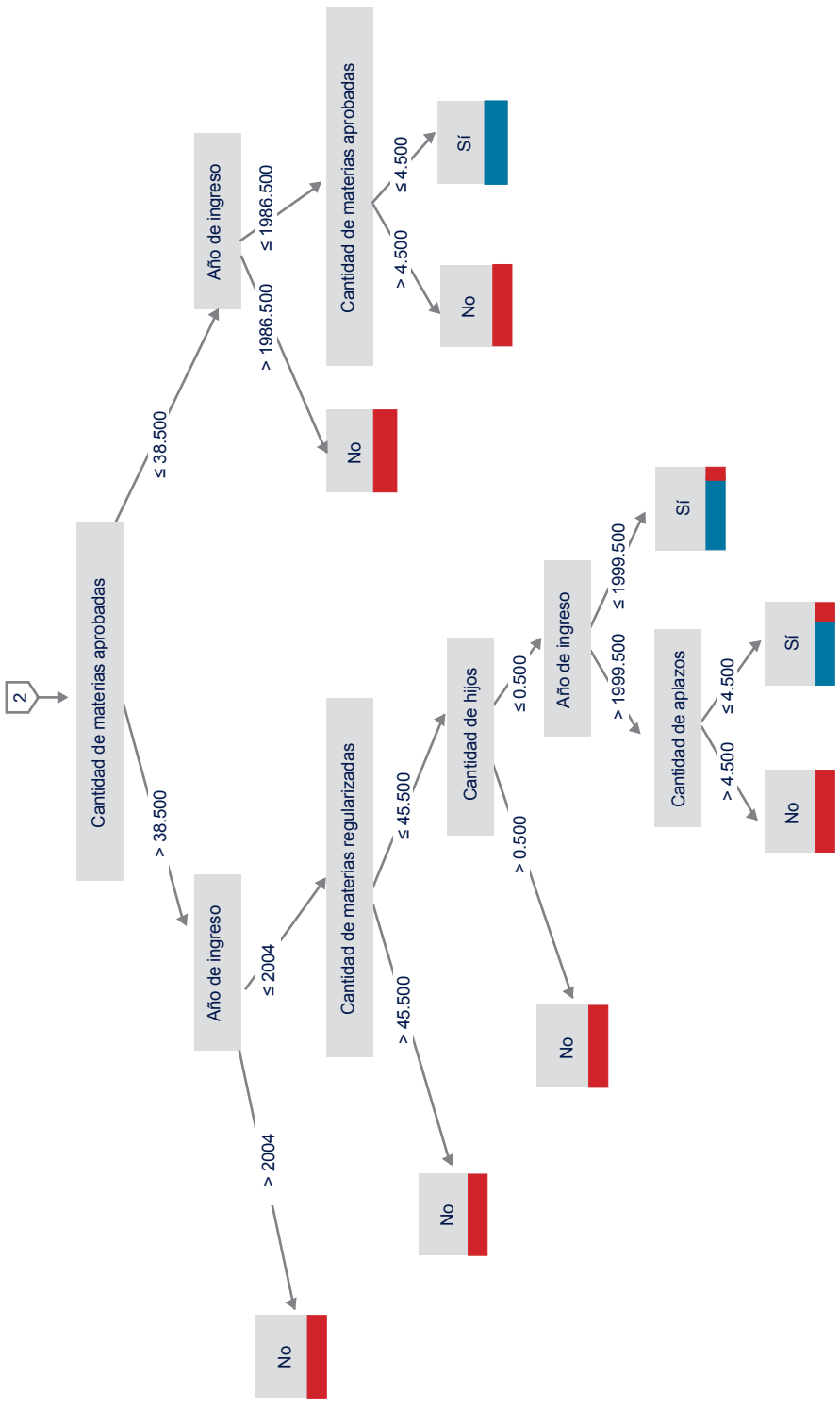


Figura 3. Árbol de decisión resultante
Elaboración propia

En la fase de prueba, se utilizó el operador performance vector (figura 4) para determinar el rendimiento del modelo de entrenamiento (figura 2); este rendimiento se analiza a través de lo que se conoce como matriz de confusión.

accuracy: 97.53%

	true Si	true No	class precision
pred. Si	231	20	92.03%
pred. No	9	914	99.02%
class recall	96.25%	97.86%	

Figura 4. Operador performance vector (matriz de confusión)
Elaboración propia

Por último, se realizó la construcción del modelo predictivo (figura 5), tomando como entrada el modelo de entrenamiento resultante, al cual se le conectó a través del operador Apply Model, el *dataset* con datos de alumnos de carreras de ingenierías que hayan ingresado entre el año 2012 y el año 2018 y que no hayan conseguido su graduación hasta la fecha.

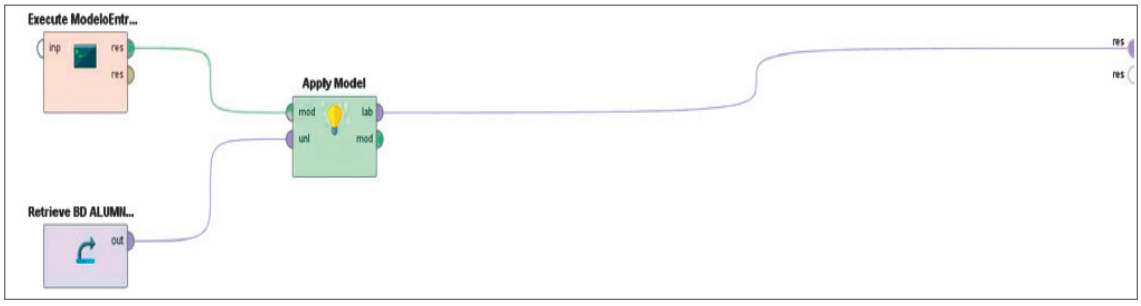


Figura 5. Modelo predictivo resultante
Elaboración propia

5. Resultados

En la fase de comprensión de datos, más precisamente en la verificación de calidad de los datos iniciales, se evidenció en el aspecto socioeconómico (ocupación madre/padre, instrucción madre/padre) y en el aspecto laboral (ocupación, tipo de trabajo), un alto porcentaje de datos perdidos, por lo que estos atributos debieron se descartados al momento de seleccionar los atributos que formaron parte de la vista minable como datos de entrada para el modelo de entrenamiento.

En la etapa de modelado, se analizaron en primera instancia los resultados que arrojó la matriz de confusión (figura 4) a través de la aplicación de la técnica de minería de datos árbol de decisión, en donde se pudo observar que el modelo de entrenamiento arrojó una tasa de aciertos

global del 97,53 %, esto significa, que el modelo entrenado tiene un alto rendimiento y es eficiente. En segunda instancia se realizó un análisis de los resultados arrojados por el operador apply model dentro del modelo predictivo (figura 5), en donde se evidencia que la variable predicha o de predicción denominada "Graduado", indica que en el periodo 2012-2018 se graduarán ocho alumnos de carreras de ingeniería de UTN Facultad Regional San Francisco (figura 6).

Prediction (Graduado)	Confidence (Sí) ↓	Confidence (No)
Sí	1	0
Sí	1	0
Sí	1	0
Sí	1	0
Sí	1	0
Sí	1	0
Sí	1	0
Sí	0.895	0.105
No	0.467	0.533
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998
No	0.002	0.998

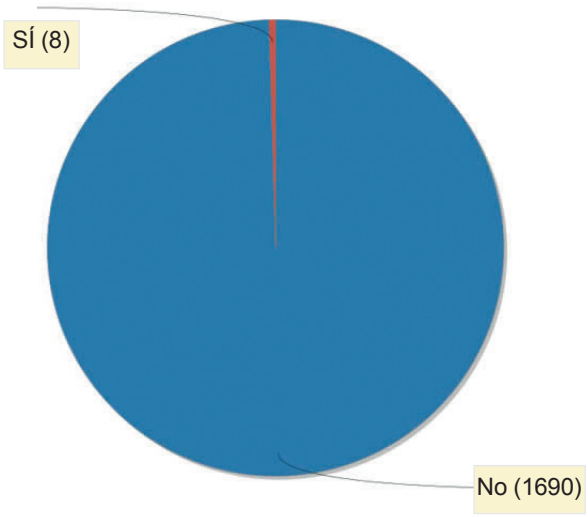


Figura 6. Resultado del Modelo predictivo resultante
Elaboración propia

Por último, en el resultado del árbol de decisión (figura 3) se observan patrones relacionados con aspectos académicos (cantidad de materias aprobadas y regularizadas, cantidad de aplazados, año de ingreso) y aspectos personales (edad, cantidad de hijos).

6. Conclusiones y trabajos futuros

Según lo expuesto anteriormente se pudo construir un modelo predictivo satisfactorio (con un nivel aceptable de tasa de aciertos), que permite determinar la graduación de alumnos de carreras de ingeniería en la UTN Facultad Regional San Francisco a través de la aplicación de la técnica de minería de datos de árboles de decisión.

En cuanto a lo que respecta a la detección de patrones que puedan incidir en la graduación de alumnos, observamos que el árbol de decisión refleja que tienen mayor relevancia los atributos relacionados con aspectos académicos del alumno, mientras los atributos vinculados a aspectos personales tienen menor relevancia.

Por último, cabe destacar que este es un modelo preliminar y que podría mejorarse si se aplica un set de datos con mayor cantidad de atributos y registros o bien a través de la aplicación de una técnica de minería de datos diferente, como puede ser el análisis de *clustering*.

Referencias

Bohórquez, K., y Torres, J. (2016). Cómo la ingeniería puede ayudar a la Sociedad. Recuperado de https://es.calameo.com/read/0048672043c00a6d40983

Colegio de Ingenieros de la Provincia de Buenos Aires. (2016). ¿Por qué faltan ingenieros? Recuperado de http://www.colegioingenieros2.org.ar/web/index.php/novedades/archivo-de-novedades/porque-faltan-ingenieros

Fernández, M. (24 de enero del 2018). Egresan 8 mil ingenieros por año frente a 34 mil graduados de sociales, abogacía y psicología. Recuperado de https://www.infobae.com/educacion/2018/01/24/psicologos-y-abogados-pero-no-ingenieros-en-algunas-disciplinas-clave-se-reciben-menos-de-25-alumnos/

Fischer, E. (2012). *Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios*. Tesis de maestría. Universidad de Chile. Recuperado de http://repositorio.uchile.cl/bitstream/handle/2250/111188/cf-fischer_ea.pdf?sequence=1

Goicochea, A. (2009). CRISP-DM: Una metodología para proyectos de minería de datos (artículo de blog). Recuperado de https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de- datos/

La Red, D., Karanik, M., Giovannini, M. y Scappini, R. (2009). Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios. *XI Workshop de Investigadores en Ciencias de la Computación: WICC 2009, 7-8 de mayo*. San Juan: Universidad Nacional de San Juan. Recuperado de http://sedici.unlp.edu.ar/handle/10915/ 53320

Naciones Unidas (1987). Día mundial de la población. Recuperado de http://www.un.org/es/events/populationday

Pérez, M. (2015). *Minería de datos a través de ejemplos*. México. Alfaomega.

Porcel, E., Dapozo, G. y López, M. (2016). Hacia un modelo predictivo de rendimiento académico utilizando minería de datos en la UTN-FRRe. *XVIII Workshop de Investigadores en Ciencias de la Computación: WICC 2016, 14-15 de abril*. Concordia: Universidad Nacional de Entre Ríos. Recuperado de http://sedici.unlp.edu.ar/handle/10915/19846

Secretaría de Políticas Universitarias del Ministerio de Educación (2012). *Plan Estratégico de Formación de Ingenieros (PEFI)*. Recuperado de http://pefi.siu.edu.ar/

UTN Facultad Regional Buenos Aires (5 de junio de 2014). La UTN forma más del 40% de los ingenieros que se gradúan en el país. Recuperado de https://www.frba.utn.edu.ar/dia-de-la-ingenieria-la-utn-forma-mas-del-40-de-los-ingenieros-que-se-graduan-en-el-pais/

Valía, L., Rostagno, J., Berto, E., Boero, D., Zelko, K., Viscusso, S., ..., Amar, E. (2017). Modelo de deserción universitaria en los primeros años de la Carrera Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Rosario. *Congreso Nacional de Ingeniería Informática. Sistemas de Información: CONAIISI 2017, 2 de noviembre*. Santa Fe: Universidad Tecnológica Nacional.