Formación Universitaria
Vol. 13(6), 23-36 (2020)
http://dx.doi.org/10.4067/S0718-50062020000600023

Factores asociados al éxito de los estudiantes en modalidad de aprendizaje en línea: un análisis en minería de datos

Gabriela Mancilla-Vela, Paola Leal-Gatica, Aurora Sánchez-Ortiz, y Cristian Vidal-Silva
Departamento de Administración, Universidad Católica del Norte, Antofagasta, Chile.
(correo-e: Gabriela.jmv@hotmail.com; paolalealg@outlook.es; asanchez@ucn.cl; cristian.vidal@ucn.cl)

Recibido Mar. 26, 2020; Aceptado May. 26, 2020; Versión final Jul. 26, 2020, Publicado Dic. 2020

Resumen

Este estudio determina las variables asociadas al éxito de los estudiantes en programas con modalidad de aprendizaje en línea (e-learning). La minería de datos es una fase del descubrimiento de conocimiento en las bases de datos (KKD, por sus siglas en inglés) que corresponde a la aplicación de algoritmos para encontrar patrones ocultos en los datos. El método utilizado es basado en el modelo CRISP-DM (proceso cruzado estándar de la industria para la minería de datos) aplicado a los programas de e-learning impartidos por el Centro de Educación a Distancia de la Universidad Católica del Norte (CED-UCN) en Chile. La muestra utilizada fue de 18.610 sujetos participantes en dichos programas durante 19 años. Los resultados obtenidos indican que las variables que permiten explicar mejor el éxito de los alumnos en programas a distancia son edad, sexo, profesión, nivel de escolaridad y región. Se concluye que estos resultados contribuyen al entendimiento de los factores críticos en la educación a distancia.

Palabras clave: minería de datos; aprendizaje en línea; KDD; CRISP-DM

Factors associated to student success in online learning: a data mining analysis

Abstract

This research study aims to determine variables associated to student success in online learning (e-learning). The knowledge discovery in databases (KDD) consists on applying algorithms to find hidden data patterns. The method used here is the CRISP-DM (cross industry standard process for data mining) and was applied to examine online degree programs at the Distance Education Center of the Northern Catholic University (DEC-NCU) in Chile. The sample was collected from 19 years of teaching and consists of 18,610 students. The results show that the variables that best explain student success are age, gender, degree study, educational level, and locality. It is concluded that these results contribute to improve the understanding of distance education critical factors.

Keywords: data mining; e-learning; KDD; CRISP-DM

INTRODUCCIÓN

El aprendizaje de habilidades y conocimientos es un proceso fundamental para el ser humano, el cual, por los avances en educación y tecnología actuales, es mucho más y fácil para las personas. El aprendizaje en línea es un modelo que ha revolucionado la educación gracias a la inclusión de las Tecnologías de la Información y la Comunicación (TICs) las cuales han hecho que las instituciones educativas estén interesadas en la utilización de nuevas metodologías en el proceso educativo (Sánchez et al., 2009). En la actualidad, existen múltiples estudios que evalúan el éxito de las plataformas tecnológicas de aprendizaje en línea, basados en su gran mayoría en el éxito de los sistemas de información de DeLone y McLean el que mide y evalúa el éxito de los sistemas de aprendizaje electrónico (Alsabawy et al., 2012; Delone y McLean, 2003). Sin embargo, a pesar del rápido crecimiento del aprendizaje en línea, existen una serie de problemáticas que enfrentan las instituciones que imparten cursos en esta modalidad y el origen de las variables que impactan el éxito de los estudiantes en estos sistemas es aún desconocido.

En la actualidad existen herramientas como la minería de datos que permiten la identificación de patrones de comportamiento en los datos y que podrían aportar en la identificación de factores asociados al éxito del aprendizaje en línea (Herrera et al., 2019). Este estudio llena el vacío en términos de las variables asociadas al éxito de los estudiantes en programas con modalidad de aprendizaje en línea adaptando la metodología de Minería de Datos CRISP-DM (Proceso cruzado estándar de la industria para la minería de datos) al descubrimiento de las variables de éxito (IBM, 2020). La importancia de esta investigación radica en que en Chile no existen investigaciones previas que identifiquen las determinantes de éxito de los estudiantes que optan por estudiar a distancia y el aprendizaje en línea podría satisfacer las necesidades, características y requisitos de potenciales estudiantes, que deseen optar por esta modalidad de estudio (Ronteltap y Eurelings, 2002). Así, esta investigación revela detalles acerca de variables asociadas con el rendimiento de los estudiantes en los distintos cursos, diplomados y postítulos en el Centro de Educación a Distancia de la Universidad Católica del Norte (CED-UCN) entre los años 2000 y 2018.

Wani (2013), destaca la importancia del aprendizaje electrónico en la educación superior ya que crea un entorno de capacitación virtual donde los alumnos pueden lograr diferencias significativas en su desarrollo gracias a este entorno de aprendizaje virtual. El aprendizaje en línea es resultado de la adopción y el uso de tecnologías de la información en la educación (Torras y Bellot, 2018). El éxito de los sistemas de aprendizaje en línea se asocia a diferentes variables tales como el nivel de uso de Tecnologías de Información y las modalidades de los programas. En la actualidad existen múltiples estudios que evalúan el éxito de las plataformas tecnológicas en aprendizaje en línea, basados en su gran mayoría en el modelo de éxito de los sistemas de información de DeLone y McLean el que mide y evalúa el éxito de los sistemas de aprendizaje electrónico (Alsabawy et al., 2012; Delone y McLean, 2003). Sin embargo, a pesar del rápido crecimiento del aprendizaje en línea, existen una serie de problemáticas que enfrentan las instituciones que imparten cursos en esta modalidad y el desconocimiento de las variables que impactan el éxito de los estudiantes en estos sistemas es aún desconocido. En la actualidad existen herramientas como el minería de datos que permiten la identificación de patrones de comportamiento en los datos y que podrían aportar en la identificación de factores asociados al éxito del aprendizaje en línea.

Hoy en día, la educación superior a distancia genera oportunidades al alumno para responder mejor a las exigencias actuales del mercado, junto con una disminución notable de barreras geográficas, económicas y de tiempo para quienes no pueden acceder a cursos de su interés de forma presencial (Ustrov, 2019). De esta forma, un aprendizaje en línea de calidad contribuiría al desarrollo de Chile. Cidral et al. (2018) estudiaron el éxito de las plataformas de aprendizaje en línea en Brasil concluyendo que la satisfacción y el uso de las plataformas por los usuarios son críticos en este resultado.

El éxito o fracaso de los estudiantes en los programas a distancia ha sido estudiado por varios autores entre los que destaca García (2019) quien analizó las causas de la deserción online universitaria en un análisis sistemático de la literatura y definió una clasificación de las variables que podrían causarlas. Las categorías que se definieron fueron: estudiante, institución, docentes, medios, grado de integración social y académica. Este mismo autor indica que el conocimiento de las características personales, sociales y demográficas pueden ser muy importantes para predecir el éxito y el fracaso de estudiantes en programas de aprendizaje en línea.

El descubrimiento de conocimiento en bases de datos (KDD), comúnmente conocido como minería de datos es un proceso para el descubrimiento de patrones y el modelado predictivo en grandes bases de datos (Fayyad et al., 1996). KDD hace un amplio uso de métodos de minería de datos, procesos automatizados y algoritmos que permiten el reconocimiento de patrones (Nájera y De la Calleja, 2017). Característicamente, la minería de datos implica el uso de métodos de aprendizaje automático (Machine Learning) desarrollados en el dominio de la inteligencia artificial (Cummins, 2019). La minería de datos se puede definir como el

proceso que utiliza técnicas estadísticas, matemáticas, artificiales y de aprendizaje automático para extraer conocimiento e identificar información pertinente y conocimiento relacionado oculto en grandes volúmenes de datos sin procesar (Sugiyarti et al., 2018). Técnicamente, la minería de datos es el proceso de encontrar correlaciones o patrones entre miles de campos en grandes bases de datos (Nájera y De la Calleja, 2017). La minería de datos encuentra estos patrones y relaciones utilizando herramientas y técnicas de análisis de datos para construir modelos, de ahí el aprendizaje automático (Witten et al., 2005). Scheuer y Mclaren (2011), proponen un modelo para identificar los factores más influyentes que pueden predecir el rendimiento de los estudiantes, no solo se predice el estado de aprobación o reprobación de los estudiantes, sino también como se considera el rendimiento del estudiante (excelente, bueno o promedio). Las principales etapas del modelo propuesto en el contexto de la predicción del rendimiento de los estudiantes se muestran en la Figura 1.

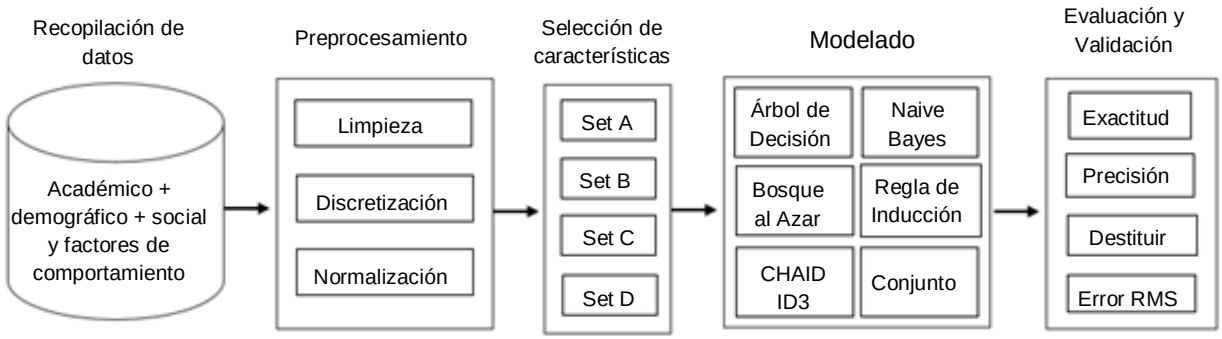

Fig. 1: Modelo propuesto para predecir los factores más influyentes de los estudiantes en riesgo (Scheuer y Mclaren, 2011)

Educational Data Mining (minería de datos educacional) se preocupa por desarrollar, investigar y aplicar métodos computarizados para detectar patrones en grandes colecciones de datos educativos, patrones que de otro modo serían difíciles o imposibles de analizar debido al enorme volumen de datos en el que existen. Hernández-Blanco et al. (2019), indican que la minería de datos educacional se preocupa por desarrollar, investigar y aplicar el aprendizaje automático, la minería de datos y los métodos estadísticos para detectar patrones en grandes colecciones de datos educativos que de otro modo serían imposibles de analizar. En este sentido también indican que en los últimos años, el uso de técnicas de aprendizaje profundo ha surgido en el campo de la minería de datos educacional.

Los datos de interés no se limitan a las interacciones de estudiantes individuales con un sistema educativo sino que también se pueden incluir datos administrativos y demográficos (por ejemplo: género, edad, calificaciones) (Scheuer y Mclaren, 2011). El principal objetivo de este estudio es determinar las variables asociadas al éxito de estudiantes en programas con modalidad de aprendizaje en línea con la utilización de la metodología CRISP-DM (IBM, 2020) según los datos de una universidad en Chile. En este estudio las causas de éxito y/o fracaso se analizan desde la categoría de variables asociadas al estudiante, ya que se toman sus características demográficas y de desempeño.

APRENDIZAJE EN LÍNEA Y MINERÍA DE DATOS

El proceso de aprendizaje en línea tiene como característica la capacidad de registrar la mayoría de las variables asociadas al proceso de aprendizaje lo que va desde los datos de ingreso de los alumnos al uso y eficiencia de la plataforma. Estos grandes volúmenes de datos proveen la oportunidad de análisis utilizando herramientas de descubrimiento de conocimiento en los datos. El KDD (Knowledge Discovery Database) o descubrimiento de conocimiento en base de datos nace en virtud de la necesidad de conocer patrones que se esconden en los grandes volúmenes de datos que los sistemas de información almacenan en general, y que es información vital para el proceso de toma de decisiones en las organizaciones (Hendrickx, et al., 2015) (ver Figura 2). Una de las fases más importantes del KDD es conocida como minería de datos, que corresponde a la aplicación de algoritmos para encontrar patrones de comportamiento ocultos en los datos (Fayyad et al., 1996; Moro et al., 2011).

La aplicación de las técnicas de minería de datos, tiene dos fines fundamentales: construir de modelos y detectar patrones (Hand et al., 2001). La construcción de modelos busca producir un resumen del conjunto de datos para identificar y describir las principales características. La detección de patrones busca identificar pequeñas desviaciones de la norma, para detectar patrones inusuales de comportamiento a través del descubrimiento de patrones y reglas y de búsquedas por contenidos. Cuando no es posible construir modelos

para el conjunto de datos, se pueden buscar patrones de comportamiento. El descubrimiento de patrones y reglas busca encontrar combinaciones y/o asociaciones de atributos que ocurren con frecuencia en transacciones de bases de datos (por ejemplo productos que se adquieren juntos). Este problema se ha atacado mediante el uso de técnicas basadas en reglas de asociación.

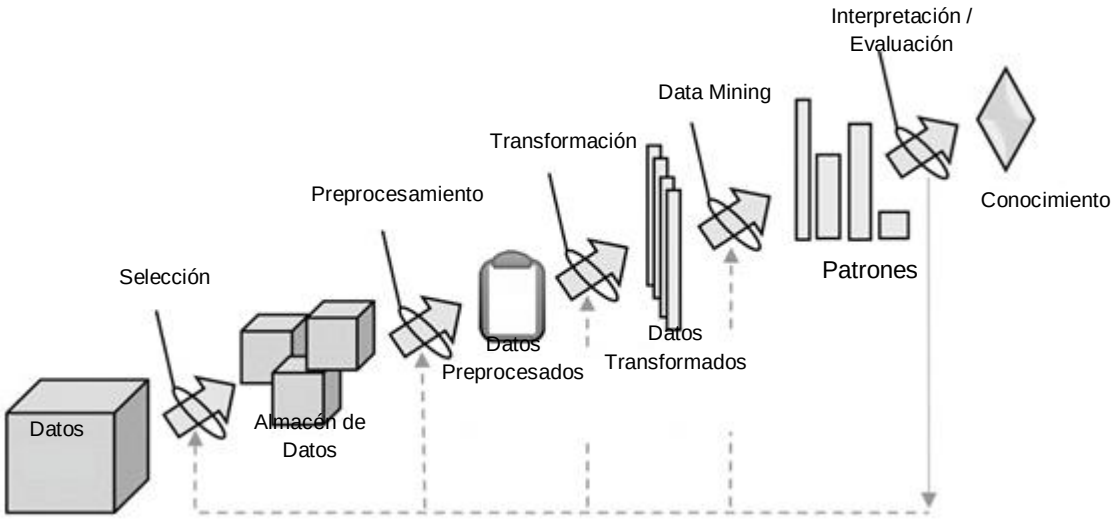

Fig. 2: Proceso de la metodología KDD (Hendrickx, et al., 2015).

El desarrollo de un proyecto usando minería de datos debe ser estructurado por lo cual es necesario contar con una metodología de desarrollo. La metodología CRISP-DM es una de las más eficientes (Daderman y Rosander, 2018; Huber et al., 2018). En esta metodología, el proceso está organizado en seis fases. Las tareas generales se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. El objetivo del CRISP-DM es permitir a diferentes empresas usar el mismo vocabulario, metodología y herramientas en las actividades de minería de datos. Las seis etapas de la metodología se presentan en la Figura 3.

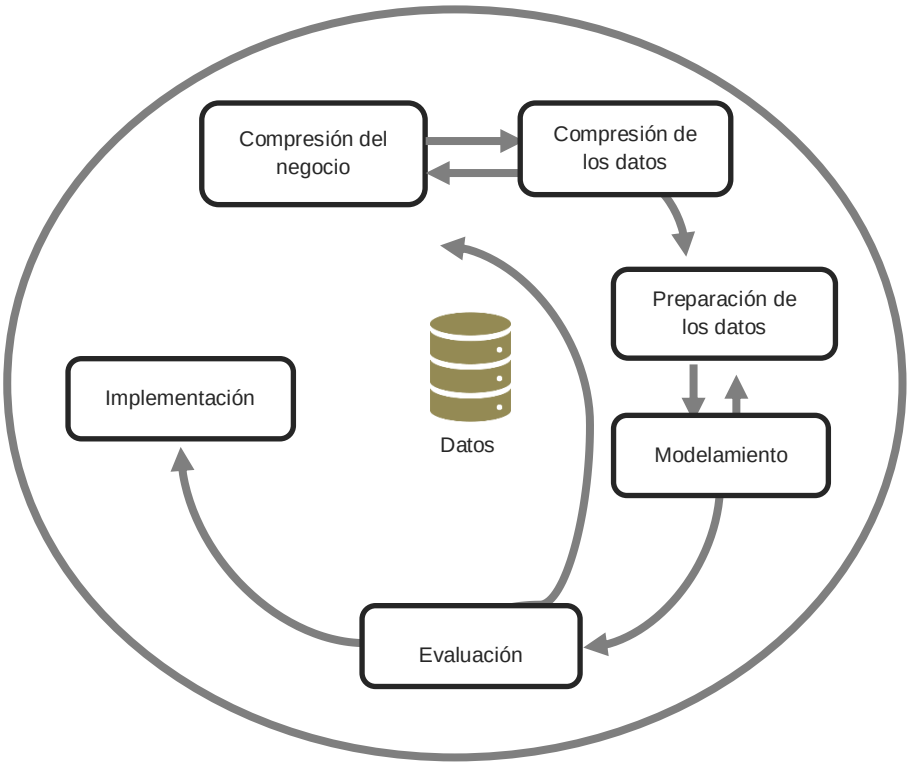

Fig. 3: El proceso de la Metodología CRISP-DM (IBM, 2020)

El significado y características de las seis etapas presentadas en la Figura 3 se explican en forma más detallada en lo que sigue: 1) Fase Comprensión del Negocio (es la primera fase análisis del problema, incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial o institucional); 2) Fase Comprensión de los Datos (es la segunda fase de análisis de datos, comprende la recolección inicial de datos, identificando la calidad de los datos); 3) Fase de Preparación de los Datos (la preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato); 4) Fase de Modelamiento (en la fase de modelado se seleccionan las técnicas de modelado más apropiadas para el proyecto. Una vez seleccionada las técnicas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos); 5) Fase de Evaluación (en la fase de evaluación, se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo); 6) Fase de Implantación (en esta etapa, además de la implantación del modelo, los resultados deben presentarse y documentarse de forma comprensible, para lograr un incremento del conocimiento).

MINERIA DE DATOS EN EDUCACIÓN A DISTANCIA

El presente estudio emplea la metodología de análisis del caso apoyada por minería de datos para conocer las condiciones iniciales del estudiante los cuales realizan programas con modalidad de aprendizaje en línea en el Centro de Educación a Distancia de la Universidad Católica del Norte (CED-UCN), permitiendo entender su éxito. Este estudio se realizó con datos del ingreso y resultados finales de estudiantes de CED-UCN entre los años 2000 y 2018 los que en su totalidad ascienden a 12.264 estudiantes. Las etapas del estudio se desarrollaron a partir del modelo CRISP-DM, se analizará la información en las bases de datos y se aplicaran las herramientas correspondientes. Particularmente se usan las técnicas y algoritmos de minería de datos como árbol de decisión, estadísticas descriptivas y redes neuronales. La herramienta computacional a utilizar es SPSS Statistics 22 (Ball-Rokeach y Hoyt, 2001). Los beneficios de esta técnica proporcionan un fácil entendimiento en la toma de decisión en minería de datos.

Conocimiento de la institución

El origen de la Educación a Distancia en la Universidad Católica del Norte, en forma de programas conducentes a un título profesional, se remonta al año 1982. El Centro de Educación a Distancia (CED-UCN) fue fundada en el año 1996, la cual depende de la Vicerrectoría Académica de la Universidad y está ubicada en la II región. El modelo pedagógico del CED-UCN está en armonización con el diseño PE-UCN (modelo PEdagog en la Universidad Católica del Norte); es decir, un modelo pedagógico que da cuenta del sello institucional sustentado en la educación en valores, basado en la formación por competencias y tomando en cuenta los cambios constantes en la sociedad. Esta unidad académica, desarrolla espacios de formación continua, a través de programas 100% on-line, utilizando un modelo de educación centrado en las tecnologías de apoyo a la educación a distancia. Uno de los mayores problemas que enfrenta CED-UCN es el desconocimiento de las variables de éxito de los estudiantes que optan por estudiar a distancia.

Selección y comprensión de los datos

Para la comprobación del estudio, fue necesario disponer de los datos de los estudiantes matriculados en el periodo comprendido entre enero 2000 y diciembre 2018. Los datos en bruto de todos los estudiantes y los datos históricos se localizan en un servidor local (LICANCABUR) con acceso restringido por personal autorizado mediante autentificación de usuario y contraseña. Este servidor da servicios al sistema de base de datos Oracle Developer 2000 de nombre ANTEC, base de datos oficial del CED-UCN. Otros datos complementarios están almacenados en archivos e impresos manejados por los encargados de las distintas áreas. Finalmente, se obtuvieron los campos o atributos con 18.610 registros válidos correspondiente a los estudiantes de CED-UCN, datos exportados desde el sistema de base de datos ANTEC, mediante consultas SQL (Lenguaje Estructurado de Consultas) para la generación de un archivo Excel con los datos solicitados.

En el sistema de base de datos ANTEC, se utilizó el browser ANTEC para realizar las consultas SQL necesarias. Se seleccionaron las tablas *ALUMNO, DIRECCIÓN_ALUMNO, PROGRAMA y ALUMNO_PROGRAMA* para obtener los datos personales y del último estado académico que alcanzó un estudiante en un programa determinado. Los datos excluidos son indiferentes a la muestra ya que no contenían información fehaciente para la investigación. Las Figura 4 y 5 presentan un extracto del modelo relacional de sistema ANTEC.

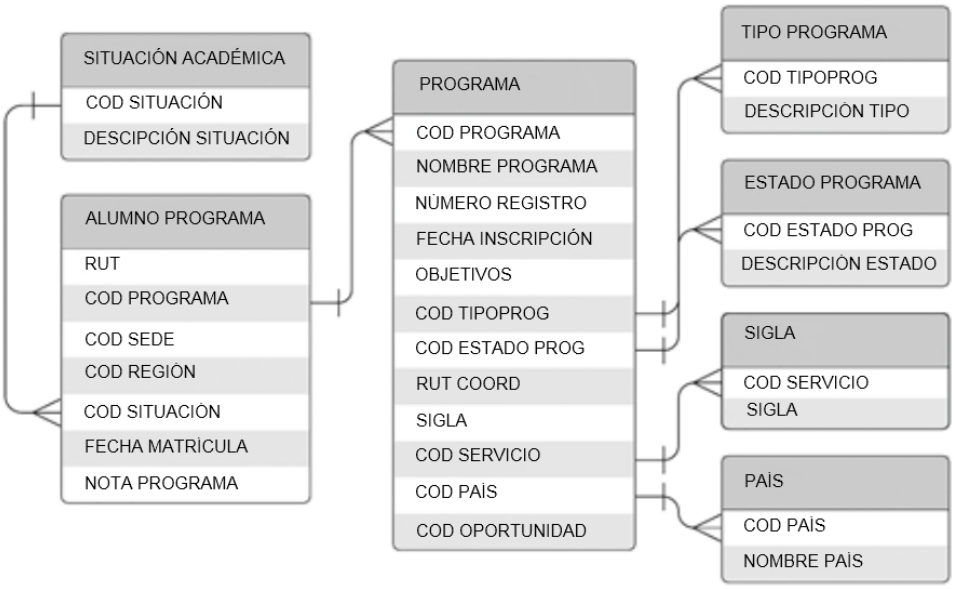

Fig. 4: Extracto del modelo relacional del sistema de base de datos ANTEC, parte 1

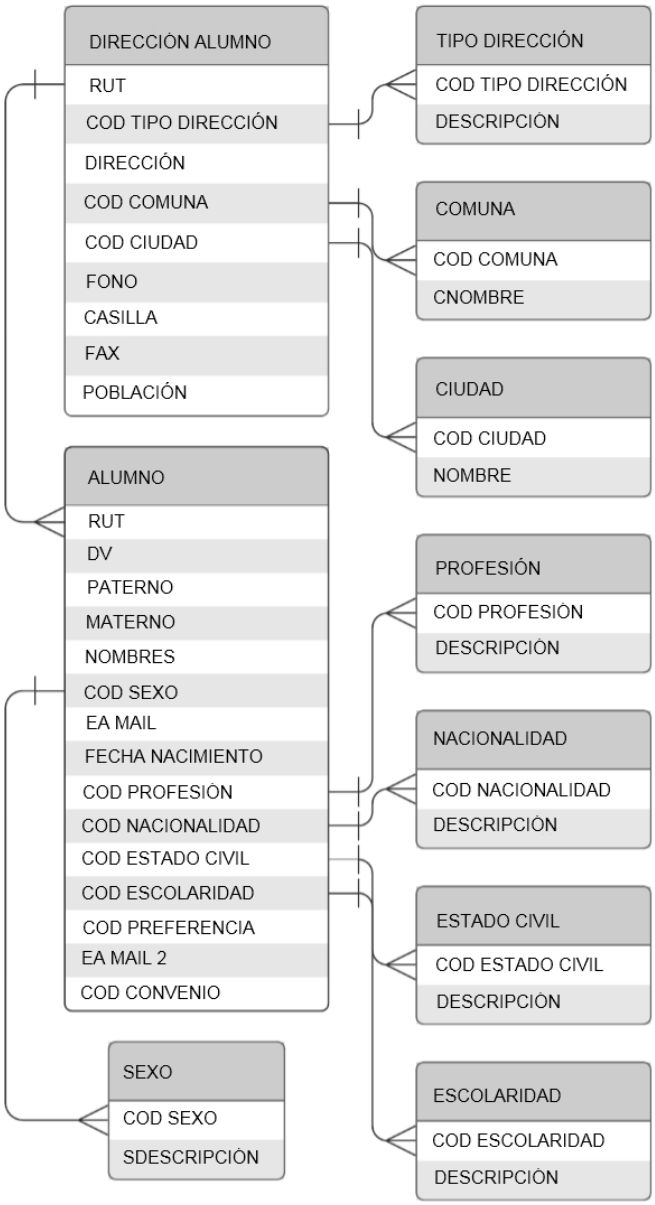

Fig. 5: Extracto del modelo relacional del sistema de base de datos ANTEC parte 1.

Preparación de los datos

Los cambios necesarios a los archivos se realizaron con la herramienta SPSS Statistics 22 debido a sus capacidades de presentación que hacen más comprensible el resultado al usuario final, estos pasos son: i) Selección de los datos y limpieza, primero son seleccionados los atributos que se utilizarán teniendo en cuenta el objetivo planteado, considerando los problemas de calidad de los datos. Para cada tabla se seleccionaron los siguientes atributos: 1) ALUMNO (RUT del estudiante, sexo, fecha de nacimiento, profesión, nacionalidad, estado civil, escolaridad); 2) DIRECCIÓN_ALUMNO (Rut del estudiante, comuna, ciudad); 3) PROGRAMA (nombre del programa, fecha de inscripción, tipo de programa); 4) ALUMNO_PROGAMA (RUT del estudiante, situación académica, nota final del programa). Para el análisis se planteó el siguiente filtro: Estudiantes que se inscribieron en los programas entre los años 2000 y 2018 ((fecha de inscripción $\geq$2000) y (fecha de inscripción $\leq$2018)).; ii) Calidad de los Datos, un problema que presentan los datos es la cantidad de datos faltantes en alguno de ellos, como el atributo nombre del programa, el cual se decidió conservar los registros con valores desconocidos debido a que al eliminar, se excluían también filas con valores válidos los cuales fueron consultados a la base de datos. Además, se encontró el problema de categorización de los datos, debido a que las técnicas que fueron utilizadas para el análisis (clasificación), utilizan fundamentalmente datos categóricos para facilitar su construcción e interpretación, por lo tanto, se fue necesario trabajar los atributos de esta forma, para facilitar la construcción del modelado; iii) Construcción de los Datos, para cumplir con el objetivo del proyecto se creó un atributo llamado región, que deriva del atributo comuna y ciudad. Región toma el valor correspondiente a la región que pertenece la comuna y ciudad. Cuando se terminó el proceso de selección y de construcción de datos, se guardaron los cambios a los archivos para que después pudiesen utilizarse en la etapa de modelamiento. Los nuevos archivos quedaron en formato SPSS.

Modelado

El modelamiento de los datos se realizó para el CED-UCN a nivel global, los datos están en el orden correcto, con la delimitación de parámetros requeridos, estos cambios sintácticos son realizados para satisfacer las exigencias de la herramienta de modelado. En este estudio, el modelo de clasificación predice “el perfil del estudiante” asociado al éxito en los programas con modalidad de aprendizaje en línea, y cuales con aquellos atributos que derivan tal perfil.

El modelo de clasificación toma como variable dependiente la variable “estado” la que es una variable categórica y que en la categoría “Titulado” es el mayor nivel de éxito de un estudiante en el modelo. El éxito académico de un estudiante en esta perspectiva mide como aquellos alumnos que se ubican en la categoría titulado de un programa que comenzaron.

Para la formulación el modelo se utilizará redes neuronales, las cuales desde un punto de vista lógico identifican relaciones entre las variables, y determinan su importancia respecto a la variable objetivo. Para la construcción de árboles de decisión se utilizará inicialmente el algoritmo C5.0 el cual tiene asociado adicionalmente un conjunto de reglas que permiten entender de forma más clara los particionamientos generadores y el algoritmo CHAID, el cual desde un punto de vista estadístico (basado en la significancia de la prueba chi-cuadrado) construye los árboles a través de la comparación de las categorías, contrayendo aquellas que no presenten diferencias en sus resultados. Posteriormente se selecciona un algoritmo de árbol de decisión basado en los resultados que se obtuvieron (predicción de casos) y el análisis de la construcción del árbol propiamente tal.

RESULTADOS

Los resultados obtenidos demuestran estadísticamente cuáles son los patrones de comportamiento que inciden en el éxito de los estudiantes que estudian en modalidad de aprendizaje en línea, como también se evidenciarán los fracasos de éstos. Los programas con mayor cantidad de estudiantes son Administración de Recursos Humanos, Gestión Ambiental, Medicación Familiar, Psicopedagogía, Gestión de Calidad Total, Gestión Integrada y Orientación Educacional (ver Tabla 1).

Tabla 1: Programas con mayor cantidad de estudiantes.

	Frecuencia	*Porcentaje*
Administración de Recursos Humanos	1996	10,7
Gestión Ambiental	1617	8,7
Mediación Familiar	1506	8,1
Psicopedagogía	1499	8,0
Gestión de Calidad Total	1169	6,3
Gestión Integrada: Calidad, Medio Ambiente y Seguridad	1009	5,4
Orientación Educacional	894	4,8

Tabla 1: continuación

	Frecuencia	*Porcentaje*
Educación Superior	648	3,5
Educación y Profesor de Educación Media Técnico Profesional	545	2,9
Profesor de Educación General Básica con Mención en NB1 Y NB2	509	2,7
Orientación Familiar	507	2,7
Técnicas de Manejo Conductual Aplicadas a Niños y Adolescentes	499	2,7
Educación y Profesor de Educación Básica	467	2,5
Derecho Procesal Penal: "Sistemas Acusatorio o Juicio Oral"	446	2,4
Trastorno de la Comunicación y el Lenguaje	418	2,2
Administración Educacional	402	2,2
Administración de Unidades Técnico-Pedagógicas	388	2,1
Preparación y Evaluación de Proyectos de Inversión	326	1,7
Mención en Lenguaje y Comunicación para Profesores del Segundo Ciclo de Lenguaje y Comunicación	292	1,6
Mención en Educación en Matemáticas para Profesores del Segundo Ciclo de Educación General Básica	247	1,3
Licenciado en Educación y Profesor de Educación Básica	237	1,3
Gestión en Comunicación Corporativa	179	1,0
Perfeccionamiento Continuo	173	0,9
Nivel Superior en Secretariado Ejecutivo	154	0,8
Educación Matemática para Profesores de Enseñanza Básica	145	0,8
Gestión Pedagógica para Formación Técnico de Nivel Superior	129	0,7
Formulación y Evaluación de Proyectos	98	0,5
Otros	2180	11,6

Fijándonos en el gráfico de árbol observamos como el primer nivel se clasifica el tipo de programa que considera CED-UCN; desde izquierda a derecha desde el nodo 1 al 6. El tipo de programa con mayor porcentaje de estudiantes titulados es Perfeccionamiento Continuo, de los cuales, los estudiantes pertenecientes a la región Metropolitana, Magallanes, Tarapacá y Bio Bio obtienen el mayor porcentaje de titulación con un 76,8%, seguido de la región de Aysén y Los Lagos con un 67,9%. (ver Figura 6).

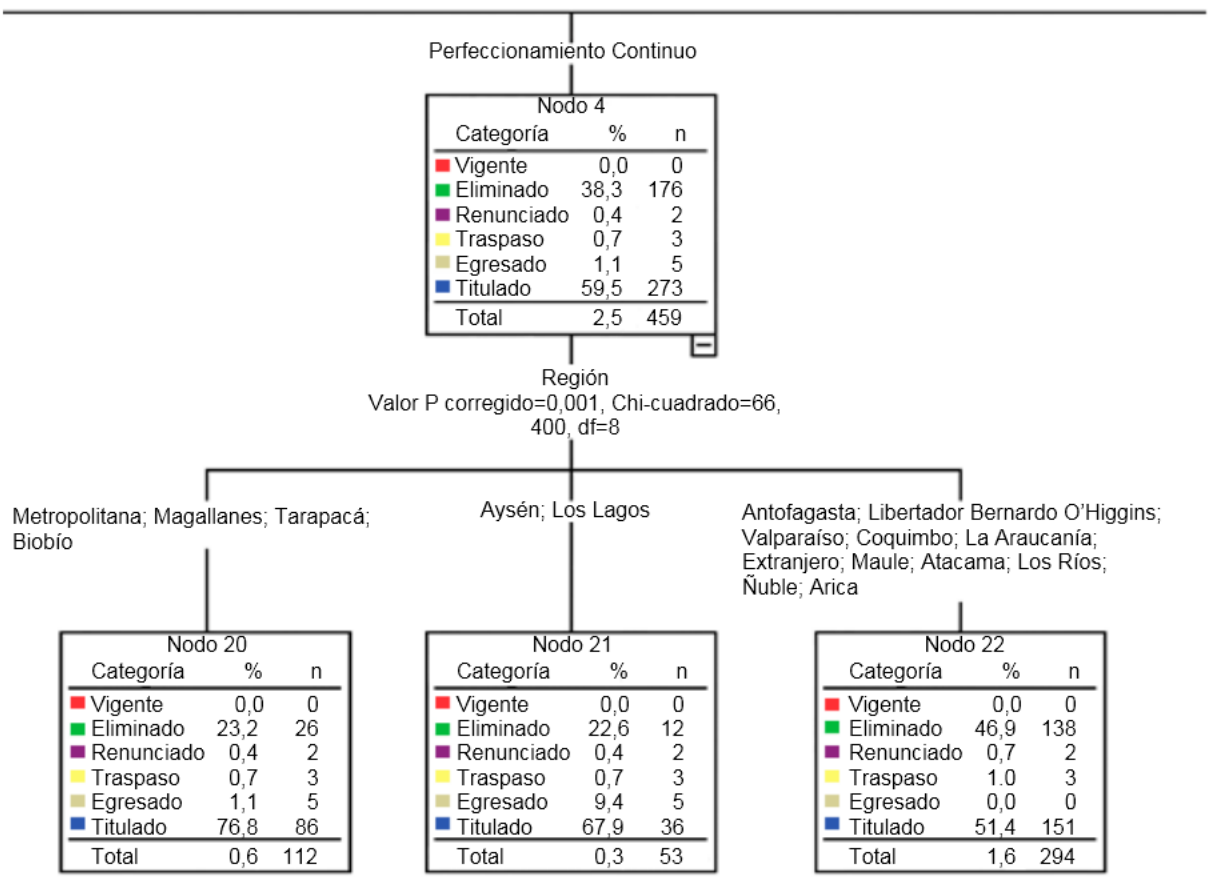

Fig. 6: Árbol de Decisión I: Programa con mayor porcentaje de estudiantes titulados

El segundo tipo de programa con mayor porcentaje de estudiantes titulados son los Cursos de Capacitación y Técnicos con un 53,8%, de los cuales los estudiantes con profesiones no universitarias tienen un gran porcentaje de titulación, en comparación de los estudiantes con profesiones universitarias del área de artes y ciencias de la salud. De los profesionales universitarios y no universitarios, ambos tienen una tendencia hacia los Cursos de Capacitación, debido a su porcentaje de titulación en comparación con los cursos técnicos (ver Figura 7).

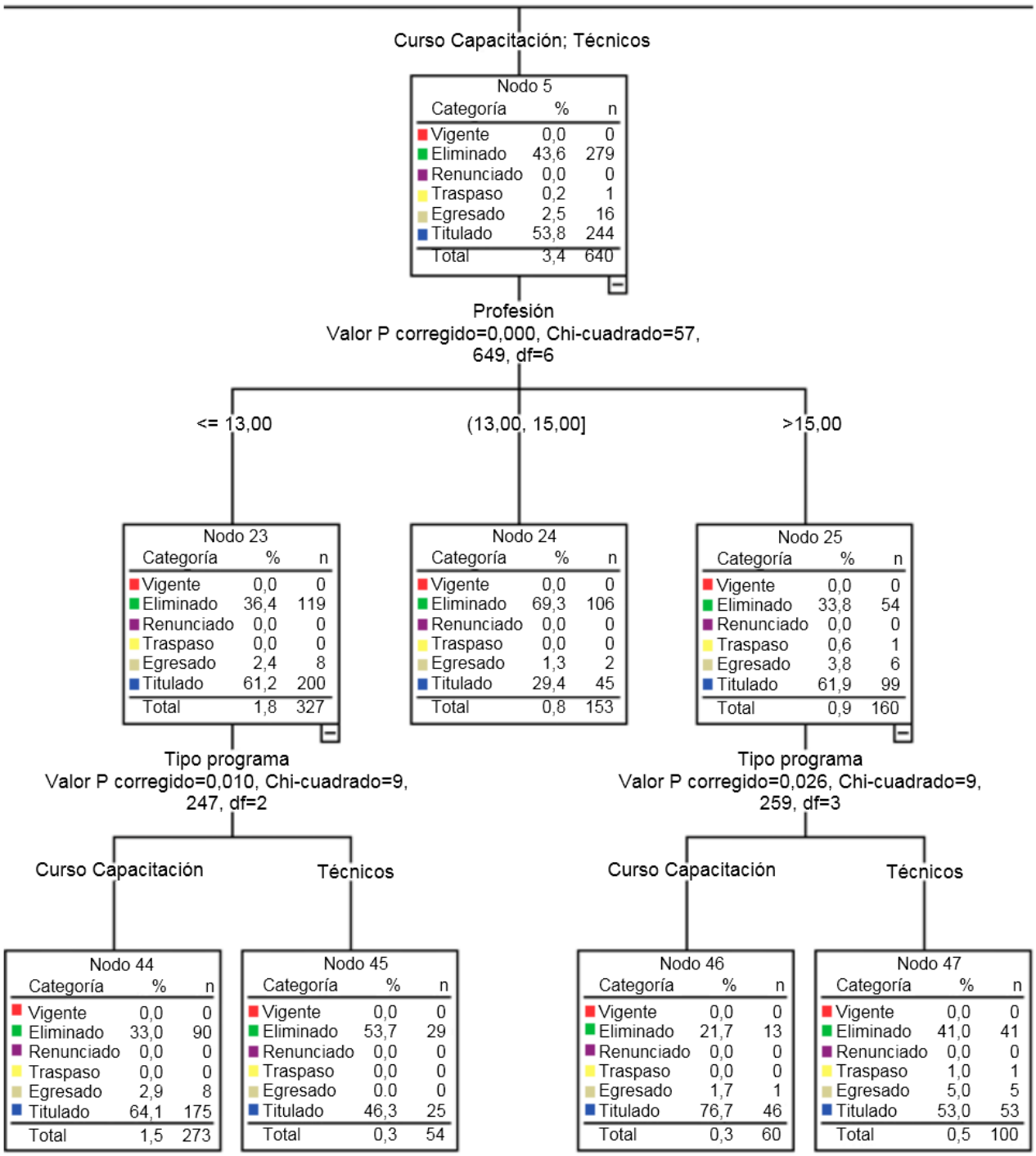

Fig. 7: Árbol de Decisión I: Segundo programa con mayor porcentaje de estudiantes titulados.

Los estudiantes con mayor porcentaje de eliminación pertenecen al tipo de programa Título de Pregrado, de los cuales, los estudiantes con enseñanza básica, enseñanza media completa, universitaria completa y técnicos de nivel superior tienen una mayor tendencia de eliminación con un porcentaje del 79,7%, en comparación de los estudiantes con estudios superiores, como universitaria completa e instituto profesional con un 66,2% (ver Figura 8).

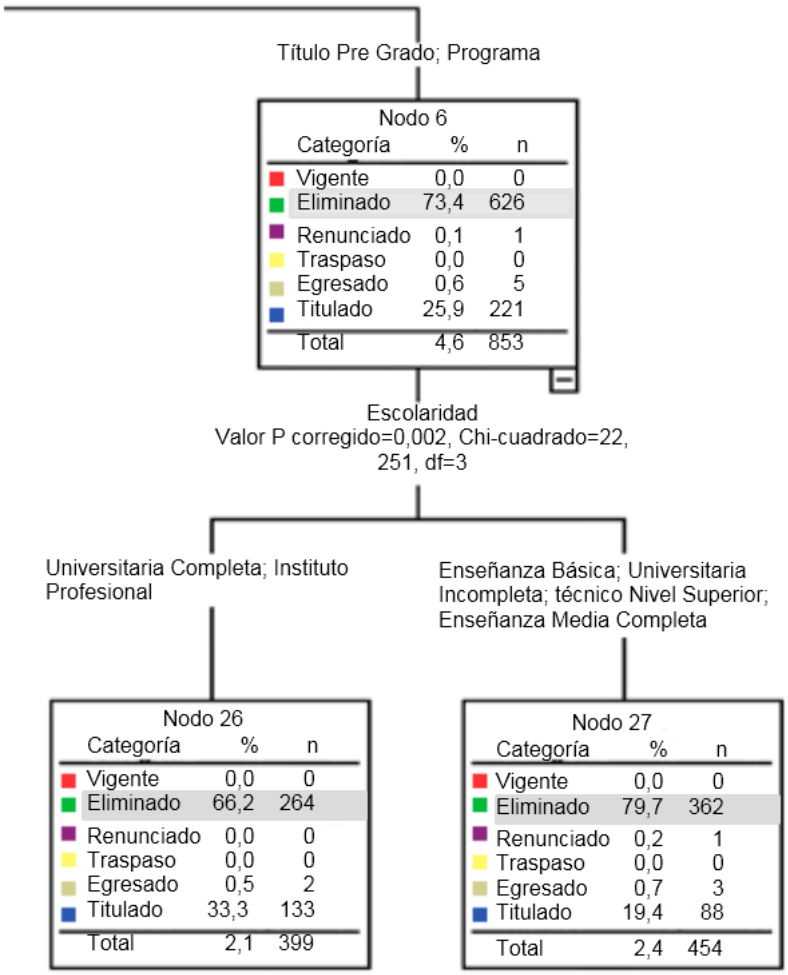

Fig. 8: Árbol de Decisión I: Programa con mayor porcentaje de estudiantes eliminados

El resultado obtenido con redes neuronales clasifica un 60,8% de predicciones correctas (ver Tabla 2) y el modelo de importancia de las variables (ver Figura 9), el cual permite tener una aproximación hacia la identificación de los factores determinantes en el éxito académico en cada tipo de programa, identifica el año el cual se dicta un programa, el programa y la profesión, como las variables más importantes según la situación académica final del estudiante en los programas con mayor éxito, lo cual otorga una primera aproximación razonables respecto del tema.

Tabla 2. Resultado clasificación global con redes neuronales

Ejemplo		*Pronosticado*						
		Vigente	*Eliminado*	*Renunciado*	*Traspaso*	*Egresado*	*Titulado*	*Porcentaje correcto*
Entrenamiento	Vigente	0	2	0	0	0	34	0,0%
	Eliminado	0	6102	0	0	0	1140	84,3%
	Renunciado	0	426	0	0	0	20	0,0%
	Traspaso	0	17	0	0	0	3	0,0%
	Egresado	0	302	0	0	0	201	0,0%
	Titulado	0	2906	0	0	0	1720	37,2%
	Porcentaje global	0,0%	75,8%	0,0%	0,0%	0,0%	24,2%	60,8%
Pruebas	Vigente	0	0	0	0	0	16	0,0%
	Eliminado	0	2565	0	0	0	479	84,3%
	Renunciado	0	181	0	0	0	8	0,0%
	Traspaso	0	13	0	0	0	0	0,0%
	Egresado	0	103	0	0	0	97	0,0%
	Titulado	0	1300	0	0	0	753	36,7%
	Porcentaje global	0,0%	75,5%	0,0%	0,0%	0,0%	24,5%	60,2%

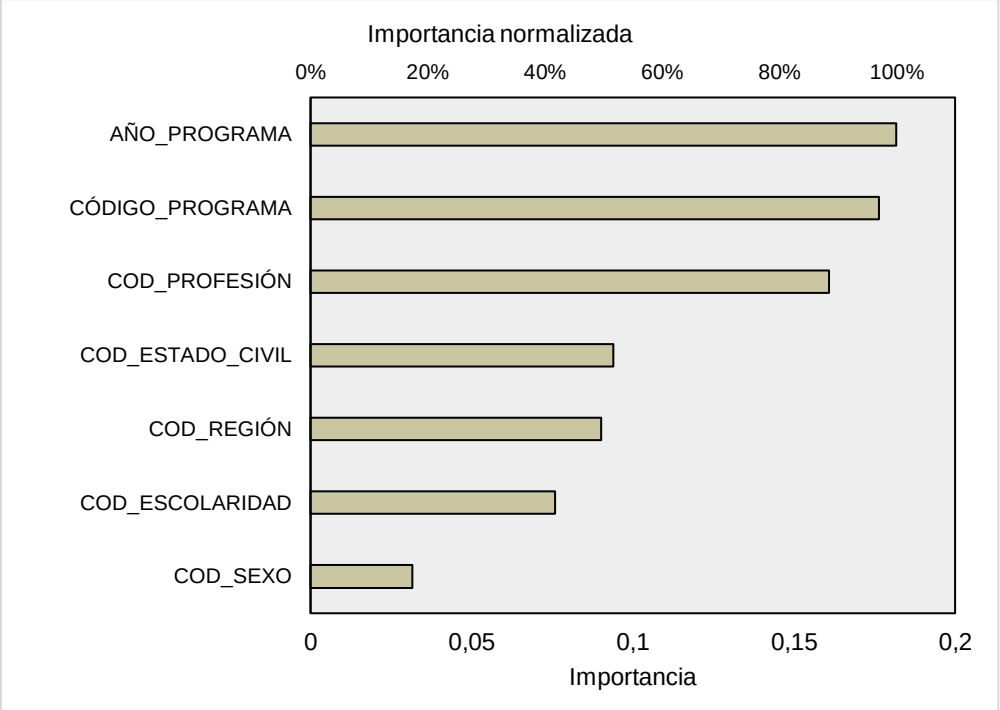

Fig. 9: Importancia de variables en el tipo de programa global con redes neuronales.

DISCUSIÓN

La aplicabilidad de técnicas y herramientas de Minería de Datos es una realidad hoy en día por los avances en tecnología. Este proyecto, el cual tenía por objetivo el identificar las variables de éxito de los estudiantes en los programas de modalidad e-learning en CED-UCN, fue construido utilizando la implementación de la herramienta de minería de datos para apoyar el proceso de toma de decisión, mediante la metodología CRISP-DM que es una de las herramientas más utilizadas en este ámbito de la investigación. A través del uso de técnicas de árboles de decisión y redes neuronales, se logró identificar aquellos factores que determinaban el éxito de los estudiantes que cursan programas de forma on-line, además de contribuir a un mayor entendimiento de los factores asociados con este tema contingente en la educación a distancia en Chile. De acuerdo con esto se identificó según la situación académica final del estudiante, los tipos de programas con mayor éxito en términos de titulación del estudiante y por el contrario, los programas con mayor fracaso. En el caso de los programas con mayor fracaso son ambos Título de Pregrado y Licenciatura los cuales son programas de mayor duración y el tipo de programa con mayor porcentaje de titulación es perfeccionamiento continuo que es por sus características de corta duración.

Este estudio tiene una gran relevancia para los programas de e-learning ya que utilizó una base de datos obtenida de uno de los programas en esta modalidad más antiguos de Chile con registros que comienzan el año 2000 hasta el 2018 inclusive, lo que implica contar con 18.610 registros que equivalen a estudiantes de estos programas en estos 19 años. Uno de los hallazgos más importantes está relacionado con el rol de las variables en el tipo de programa y que el éxito o fracaso de los estudiantes. Se estableció que el éxito o fracaso va a depender en gran medida de variables como edad, sexo, profesión, nivel de escolaridad y región. La comprensión de los factores relevantes en el éxito académico en cada tipo de programa es determinante a la hora de seleccionar a los estudiantes y en los procesos de difusión de los programas. Estos resultados permitirán apoyar al know-how de la organización respecto al establecimiento de sus políticas de difusión y mantenimiento de estudiantes en modalidad de aprendizaje en línea.

CONCLUSIONES

De acuerdo al trabajo presentado y a los resultados obtenidos, se pueden plantear las siguientes conclusiones principales:

La utilización de la minería de datos educacional y particularmente la metodología CRISP-DM es un gran aporte a la sistematización y eficiencia en la identificación de patrones en los datos en la educación a distancia. El estudio permitió no solo sistematizar los datos que existían en diversas fuentes y formatos en las plataformas de la educación a distancia en la institución en estudio, UCN, sino que también permitió aportar valiosa información para futuros análisis en este contexto.

Las herramientas de Minería de Datos pueden presentar mayores ventajas que el uso de herramientas puramente estadísticas ya que ellas son de carácter básicamente exploratorio por lo que permiten trabajar con diversas dimensiones de un mismo problema,. También es importante destacar la posibilidad y flexibilidad de estas herramientas de análisis al permitir el trabajo con diferentes tipos de variables, categóricas y numéricas, en un mismo análisis.

Los árboles de decisión mostraron ser una herramienta de gran ayuda al momento de encontrar relaciones entre variables que no se identificaron como probables en análisis con otras herramientas previamente ya que las técnicas utilizadas son menos restrictivas que las estadísticas, dado que no requieren por ejemplo condiciones de normalidad de datos y son tolerantes a ruidos en los datos.

Los resultados del estudio en el caso de la Universidad Católica del Norte permitirán a la organización focalizar los esfuerzos de admisión así como de retención de estudiantes potencialmente más expuestos y proclives a la deserción.

REFERENCIAS

Alsabawy, A., Cater-Steel, A. y Soar, J., *A Model to Measure E-Learning Systems Success*, In Measuring organizational information systems success: New technologies and practices, doi: 10.4018/978-1-4666-0170-3.ch015, IGI Global (2012)

Ball-Rokeach, S. J. y Hoyt, E. G., *Communication Technology and Community*, doi: 10.1177/009365001028004001, Communication Research, 28(4), 355–357 (2001)

Torras, E. y Bellot, A., *Informe e-learning 2018* - OBS Business School (en línea), http://www.obs-edu.com. Acceso: 10 de Enero (2020)

Cidral, W. A., Oliveira, T., Di Felice M. y Aparicio, M., *E-learning success determinants: Brazilian empirical study*, doi: 10.1016/j.compedu.2017.12.001, Computers & Education, 122, 273–290 (2018)

Cummins, M. R., *Nonhypothesis-Driven Research: Data Mining and Knowledge Discovery*, doi: 10.1007/978-3-319-98779-8_16, Clinical research informatics, 341–356 (2019)

Daderman, A. y Rosander, S., *Evaluating Frameworks for Implementing Machine Learning in Signal Processing and KDD,* KTH Skolan För Elektroteknik Och Datavetenskap, Estcolmo, Suecia (2018)

Delone, W. H. y McLean, E. R., *The DeLone and McLean Model of Information Systems Success: A Ten-Year Update, doi:* 10.1080/07421222.2003.11045748, Journal of Management Information Systems, 19(4), 9–30 (2003)

Hand, D. J., Mannila, H. y Smyth, P., *Principles of Data Mining*, MIT press, Cambridge, Massachusetts (2001)

Hendrickx, T., Cule, B., Meysman, P. y otros tres autores, *Mining association rules in graphs based on frequent cohesive item sets,* doi: 10.1007/978-3-319-18032-8_50*,* Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9078(3), 637–648 (2015)

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B., A systematic review of deep learning approaches to educational data mining, doi: 10.1155/2019/1306039, Complexity 1, 1 – 22, Mayo (2019)

Herrera, M., Ruiz, S., Romagnano, M. y otros tres autores, *Aplicando métodos y técnicas de la ciencia de los datos a datos universitarios*, XXI Workshop de Investigadores en Ciencias de la Computación, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Argentina, Abril (2019)

Huber, S., Wiemer, H., Schneider, D. y Ihlenfeldt, S., *DMME: Data mining methodology for engineering applications a holistic extension the CRISP-DM model*, doi: 10.1016/j.procir.2019.02.106, en Actas 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 79, 403–408 (2018)

IBM, *Manual CRISP-DM de IBM SPSS Modeler* - IBM Corporation, (en línea), ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf. Acceso: 10 de Enero (2020)

Fayyad, U., Piatetsky-Shapiro, G. y Smyth, P., *From Data Mining to Knowledge Discovery in Databases*, doi: 10.1609/aimag.v17i3.1230, AI Magazine, 17(3), 37 (1996)

Moro, S., Laureano, R. M. S. y Cortez, P., *Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology*, 25th European Simulation and Modelling Conference - ESM'2011, 117–121 (2011)

Nájera, A y De la Calleja, J., *Brief Review of Educational Applications Using Data Mining and Machine Learning*, Redie Revista Electrónica de Investigación Educativa, 19(4), 84–96 (2017)

Ronteltap, F. y Eurelings, A., *Activity and Interaction of Students in an Electronic Learning Environment for Problem-Based Learning*, doi: 10.1080/01587910220123955, In Distance Education, Vol. 23 (2002)

Sánchez, A., Boix, J. y. Jurado, P, *La Sociedad del Conocimiento y las TICs: Una Inmejorable Oportunidad para el Cambio Docente*, Pixel-Bit, Revista de Medios y Educación, 34, 179–204 (2009)

Scheuer, O. y Mclaren, B. M., *Educational Data Mining*, doi: 10.1007/978-1-4419-1428-6_618, Encyclopedia of the Sciences of Learning, 1075–1079 (2011)

Sugiyarti, E., Jasmi, K. A., Basiron B. y otros dos autores, *Decision Support System of Scholarship Grantee Selection using Data Mining*, International Journal of Pure and Applied Mathematics, 119 (15), 2239–2249 (2018)

Ustrov, Y., *e-Learning 2019* - OBS Business School, (en línea), https://www.obs-edu.com/es/informe-de-investigacion/informe-obs-e-learning-2019. Acceso: 15 de Enero (2020)

Wani, H., *The Relevance of E-Learning in Higher Education, ATIKAN* Journal Kajian Pendidikan, 3 (2), (2013)

Witten, I. H., Frank, E. y Hall, M. A., *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Cambridge, MA, United States (2005)

Página en blanco