opción

Revista de Antropología, Ciencias de la Comunicación y de la Información, Filosofía, Lingüística y Semiótica, Problemas del Desarrollo, la Ciencia y la Tecnología

Año 35, 2019, Especial N° 25

Revista de Ciencias Humanas y Sociales
ISSN 1012-1587/ ISSNe: 2477-9385
Depósito Legal pp 198402ZU45

Universidad del Zulia
Facultad Experimental de Ciencias
Departamento de Ciencias Humanas
Maracaibo - Venezuela

opción
Revista de Ciencias Humanas y Sociales

ISSN 1012-1587/ ISSNe: 2477-9385
Depósito legal pp. 198402ZU45
Portada: De Cabimas a Maracaibo enamorado
Artista: Rodrigo Pirela
Medidas: 100 x 60 cm
Técnica: Mixta sobre tela
Año: 2010

Opción, Año 35, Especial No.25 (2019): 482-502
ISSN 1012-1587/ISSNe: 2477-9385

Minería de Datos: Una propuesta metodológica para educación superior

Jorge Díaz Ramírez
jdiazr@academicos.uta.cl

Ximena Badilla Torrico
xbadilla@academicos.uta.cl

José Luis Martí Lara
jmarti@inf.utfsm.cl

Universidad de Tarapacá – Sede Iquique - Chile

Resumen

El objetivo de este trabajo fue proponer una metodología de minería de datos en carreras de Ingeniería Civil en la Universidad de Tarapacá, sede de Iquique. Para lo cual se utilizó CRISP-DM. Con esto, es posible aplicar las diferentes etapas de la metodología a datos reales, dependiendo del problema a resolver, generando nuevo conocimiento y utilizando la herramienta Rapidminer, con diferentes mediciones y algoritmos. Luego, se concluye que esta metodología puede generar nuevos conocimientos basados en pasos establecidos, teniendo la posibilidad de aplicar prácticas innovadoras en la gestión de datos.

Palabras Claves: CRISP-DM; Metodología; Ingenierías Civiles; Universidad de Tarapacá.

Data Mining: A methodological proposal for higher education

Abstract

The objective of this work was to propose a methodology of data mining in civil engineering careers at the University of Tarapacá, Iquique headquarters. For which CRISP-DM was used. With this, it is possible to apply the different stages of the methodology to real data,

Recibido: 03-07-2019 •Aceptado: 15-11-2019

depending on the problem to be solved, generating new knowledge and using the Rapidminer tool, with different measurements and algorithms. Then, it is concluded that this methodology can generate new knowledge based on established steps, having the possibility of applying innovative practices in data management.

Keywords: CRISP-DM; Methodology; Civil Engineering; Universidad de Tarapacá.

1. INTRODUCCIÓN

Las organizaciones en la actualidad no solo cuentan con datos, sino también deben gestionar el conocimiento que se puede extraer de ellos, así, contar con prácticas innovadoras y tecnologías adecuadas, generan en las organizaciones una ventaja competitiva. Luego, las metodologías de Minería de Datos (MD) son una oportunidad para cualquier organización de entender los datos y obtener conocimiento nuevo de ellos. En la actualidad existen diferentes metodologías de MD que permiten buscar nuevas formas de trabajos sobre los datos o procesos para generar conocimiento, como por ejemplo, Descubrimiento de Conocimiento en Bases de Datos (KDD Knowledge Discovery in Databases) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), CRISP-DM (Chapman, y otros, 2000) y SEMMA (Azevedo & Santos, 2008).

Como toda organización que cuenta con datos, las Universidades cuentan con una gran variedad de ellos, de distinta índole, como personales, académicos, etc., dando la oportunidad de utilizar MD para generar conocimiento en los diferentes indicadores clave, como por ejemplo la retención, titulación oportuna, etc. Luego,

una de las principales actividades de las Universidades es la docencia, donde se deben implementar procesos de enseñanza-aprendizaje para dar las oportunidades a todos los estudiantes que ingresan en primer año y que pueden terminar su ciclo educativo, para finalmente obtener los títulos profesionales y/o grados académicos. Es en este sentido, los indicadores claves se vuelven fundamentales para cumplir lo antes señalado.

El caso en particular que se utilizó como ejemplo fue el de la Universidad de Tarapacá (UTA) y sus carreras de Ingenierías Civiles presentes en la Sede Esmeralda en Iquique. Junto a esto, el Anuario Institucional 2017 entrega indicadores claves que son factibles de analizar, como por ejemplo Indicadores de Progresión (tasas de retención) e Indicadores de Resultados (Titulados pregrado, programas especiales, carreras técnicas y de postgrado), entre otros (Dirección de Calidad Institucional, 2018). Con estos datos, vinculados a carreras de ingenierías, se torna una oportunidad de análisis en base a una metodología de MD y se utilizará el Indicador de Progresión como ejemplo en este estudio.

Es por lo anterior, que el objetivo de este trabajo fue proponer una metodología de minería de datos en las carreras de Ingenierías Civiles en la Universidad de Tarapacá, sede Iquique. Y se decide utilizar CRISP-DM (Chapman, y otros, 2000), puesto que para Piatetsky (2014) sigue siendo la más utilizada en la industria, abarcando una visión general del problema a resolver, y en particular

la forma de North (2016) de su libro Data Mining for the Masses, second edition (North, 2016).

2. FUNDAMENTOS TEÓRICOS

Son varios los trabajos realizados en relación con la implantación de metodologías de MD en organizaciones, tanto públicas como privadas, y en la educación. Tal es el caso de Rosado (2017) donde aplica minería de datos a entornos educativos y extrae conocimiento para identificar el comportamiento de los estudiantes al interactuar con materiales y tutores (Rosado & Verjel, 2017). Además de Galán (2015) que aplica la metodología de CRISP-DM en detalle sobre datos académicos, con la finalidad de sacar conclusiones que ayuden a mejorar los servicios que ofrece la universidad a sus estudiantes (Galán Cortina, 2015) y Rodríguez (2017), donde utiliza un modelo predictivo para la permanencia de los estudiantes en base a variables de ingreso y medidas durante el primer semestre (Rodríguez, González Campos, & Patricio Aguilera, 2017).

Dependiendo del objetivo que se pretenda lograr, dentro de la metodología de MD existen las tareas y métodos, los cuales son descritos por Hernández (2004), donde la tarea es un problema de MD, las cuales son (1) Predictivas: se trata de problemas y tareas en los que hay que predecir uno o más valores para uno o más ejemplos y (2) Descriptivas: los ejemplos se presentan como un conjunto sin etiquetar ni ordenar de ninguna manera, por lo tanto, el objetivo no es predecir

nuevos datos, sino describir los existentes. Así mismo, para cada una de las tareas, como cualquier problema, requiere de métodos, técnicas o algoritmos para resolverlos (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004). Son estas tareas y métodos los que se deben utilizar para lograr resolver los problemas que emergen de los datos y a favor de los indicadores claves.

3. METODOLOGÍA COMPUTACIONAL

La metodología que se utilizó fue CRISP-DM, la que en base a Chapman (2000) y North (2016) se descompone en las siguientes etapas:

3.1 Comprensión Organizacional

Para Chapman (2000) esta etapa es inicial y se enfoca en comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición de problema de extracción de datos y un plan preliminar diseñado para alcanzar los objetivos (Chapman, y otros, 2000, pág. 10).

Al tomar esta definición, se debe definir cuál es el objetivo de la Universidad, con una visión empresarial, en base a su modelo

educativo. Así, obtener luego en objetivo de minería de datos, focalizando esto en indicadores claves.

3.2 Comprensión de los datos

Chapman (2000) define esta etapa como el comienzo de la recopilación de datos inicial y continúa con actividades que le permiten familiarizarse con los datos, identificar problemas de calidad de los datos, descubrir los primeros conocimientos sobre los datos y/o detectar subconjuntos interesantes para formar hipótesis con respecto a la información oculta (Chapman, y otros, 2000).

Así, a modo de ejemplo, los datos en estudio fueron proporcionados por el Departamento de Análisis, Estudios y Calidad de la Universidad de Tarapacá desde las bases de datos centrales, las cuales son administradas por un sistema gestor de base de datos objeto-relacional ORACLE, versión 8i, el cual funciona en un servidor con una distribución de GNU/Linux (CentOS). Los datos obtenidos se almacenaron en copias locales en formato Excel y corresponden a datos de fichas académicas, ingreso a la universidad y promedios de asistencia, formando un total de 12.195 datos, con 304 registros (estudiantes) y 43 variables. Todos los datos son desde 2009 y hasta el primer semestre de 2018, de las tres carreras de Ingenierías civiles presente en la Sede Esmeralda de la Universidad, esto es Ingeniería Civil en Informática, Ingeniería Civil Industrial e Ingeniería Civil Eléctrica.

3.3 Preparación de los datos

Para Chapman (2000) esta etapa cubre todas las actividades necesarias para construir el conjunto de datos final a partir de los datos sin procesar iniciales. Es probable que las tareas de preparación de datos se realicen varias veces y no en cualquier orden prescrito. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de datos para herramientas de modelado (Chapman, y otros, 2000).

Luego, en esta etapa, y en base al ejemplo, se eliminaron manualmente algunas variables que no aportan al análisis, están repetidos y para la no identificación de estudiantes, como los siguientes: rut, digito verificador, fecha nacimiento, DEMRE, Plan, sistema de ingreso, código carrera, nombre carrera, 3era y 4ta oportunidad (con valores 0's), región y comuna (solo se utilizó ciudad), nombre de colegio, alumno, fecha matricula, celular y email. Además, se decide eliminar año última situación y situación actual ya que cuentan con 235 valores perdidos, esto es 77,3% de la nuestra total. Además, la variable Año de Egreso cuenta con cinco valores 0, Nota de Enseñanza Media cuenta con dos valores 0, Lugar cuenta con cinco valores 0, Puntaje PSU cuenta con ocho valores 0, Puntaje Ranking cuenta con catorce valores 0, Colegio cuenta con dos valores "Sin Datos", por lo que se decide eliminar esos registros (estudiantes), ya que no aportan al análisis. Finalmente, los datos que se utilizaron contienen 268 registros y se compone de 24 variables, descritos en la tabla 1.

Tabla 1. Descripción de datos

Dato	Significado
Año Ingreso (AI)	Año de ingreso a la universidad
Avance Curricular (AC)	Avance curricular del alumno
Promedio Final (PF)	Promedio final de notas del alumno
Cursadas (C)	Cantidad de asignaturas cursadas por el alumno
Aprobadas (Ap)	Cantidad de asignaturas aprobadas por el alumno
Reprobadas (Re)	Cantidad de asignaturas reprobadas por el alumno
1era Oportunidad (1era)	Cantidad de asignaturas aprobadas en primera oportunidad
2da Oportunidad (2da)	Cantidad de asignaturas aprobadas en segunda oportunidad
Edad (Ed)	Edad del alumno
Ciudad (Ciu)	Ciudad de procedencia del alumno
Sexo (Sx)	Género del alumno
Gratuidad (Gr)	Posee gratuidad el alumno
Año egreso (AE)	Año de egreso de la enseñanza media
Nota EM (NEM)	Notas de la enseñanza media
Preferencia (Pref)	Preferencia al postular a la carrera
Lugar (Lg)	Lugar en la posición de ingreso a la carrera
Puntaje Ponderado (PP)	Puntaje ponderado de la PSU
Puntaje PSU (PPSU)	Puntaje obtenido en la Prueba de Selección Universitaria
Puntaje Ranking (PR)	Ranking del alumno según promoción
Dependencia (Dp)	Tipo de dependencia del colegio de procedencia
Año Matrícula (AM)	Año de la matrícula en la carrera
Tutorado (Tt)	Ha realizado tutorías en la carrera
Porcentaje de Asistencia (PA)	Porcentaje de promedio asistencia hasta el 2do semestre de 2017
Matricula (Mt)	Matricula vigente (SI/NO), la que para este ejemplo fue la clase

Fuente: Elaboración propia

3.4 Imputación de datos perdidos

Useche (2006) en base a Lohr (1999), indica que la imputación no solo reduce el sesgo, sino también produce datos limpios (Useche & Mesa, 2006, pág. 130). Además, Hernández (2004) destaca tres razones por las cuales se pueden reemplazar los valores perdidos, la primera razón, menciona que el método de minería de datos no trata bien los valores perdidos, la segunda razón, menciona que la agregación de datos numéricos para generar vistas minables y finalmente puede que el método trate los valores faltantes, pero produzca sesgo al ignorar el ejemplo completo (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004, pág. 74).

Luego, para tener una claridad en términos de las variables que contienen valores perdidos, en los datos de ejemplo, se observó que Ciudad cuenta con 11, Año de egreso cuenta con 31, Notas de enseñanza media cuenta con 32, Preferencia y Lugar cuentan con 35, Puntaje ponderado cuenta con 25, puntaje PSU y Ranking cuentan con 26 y Dependencia cuenta con 31 valores perdidos. Para este caso de imputación de valores perdidos, se consideran los criterios mostrados por Useche (2006), en particular el tipo de variable a imputar y los parámetros que se desean estimar (p. 145). Luego, para Goicoechea (2002) y Useche (2006) existen varios algoritmos de imputación de datos, las cuales se contrastaron con la herramienta RapidMiner (2018) para llevar a cabo esta etapa, así después comparar las varianzas de las variables antes y después de la imputación. Tales resultados se observan en tabla 2.

Tabla 2. Varianzas de algoritmos de imputación

Variable	Original	K-NN	Decision Tree	Random Forest	Neural Net	Deep Learning
AE	6,3	6,0	6,4	6,1	5,7	6,0
NEM	0,2	0,2	0,2	0,2	0,2	0,2
Pref	0,8	0,7	0,7	0,9	0,7	0,7
Lg	207,4	188,5	187,4	189,9	187,7	185,0
PP	5098,0	4783,1	4662,2	4622,6	4747,2	4619,9
PPSU	2611,2	2448,3	2366,8	2359,0	2490,0	2360,0
PR	14161,0	13156,1	12927,7	12814,2	12996,0	12814,2

Fuente: Elaboración propia

Cabe destacar que algunos algoritmos usados en la imputación solo aceptan variables numéricas, por lo mismo las variables Ciudad y Dependencia no están presentes en tabla 2. Caso distinto es el algoritmo K-NN (K Nearest Neighbor o Vecinos más cercanos), el cual tiene resultados más eficientes, debido al trabajo con datos polinomiales, enteros y reales, ya que se adapta mejor a esas variables (Zhang, 2012). Así, se entrega un acercamiento más real de las variables imputadas, ya que en Rapidminer se puede seleccionar el tipo de medida para encontrar los vecinos más cercanos, siendo para este caso Mixed Measures, puesto que existen variables con valores nominales y numéricas en conjunto. Además, se observa en tabla 2, que el algoritmo de Deep Learning entrega más varianzas pequeñas, siendo esto un factor importante para la decisión de una técnica u otra (Goicoechea, 2002). Finalmente, en consideración a estos tipos de variables y las diferencias entre las varianzas de Deep Learning y K-NN, se define que la técnica final a utilizar para la imputación de

valores perdidos será K-NN, por su mejor adaptabilidad a las variables.

3.5 Modelamiento

Para Chapman (2000) en esta etapa se seleccionan y aplican diversas técnicas (algoritmos) de modelado, y sus parámetros se calibran a valores óptimos. Normalmente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos en la forma de datos. Por lo tanto, a menudo es necesario volver a la fase de preparación de datos.

Para este estudio, se realizaron 3 fases experimentales, las cuales son en base a 3 fuentes de datos, todas ellas obtenidas desde la fuente original. La primera fuente fue con los datos sin las filas (registros) que contenían valores perdidos, quedando con un total de 225 registros, esto es 83,96% de los datos originales, lo cual está dentro de los porcentajes aceptados (Useche & Mesa, 2006, pág. 130). La segunda fuente de datos fue con todos los registros (268), pero con valores perdidos imputados. Finalmente, la tercera fuente será la fuente original incluyendo los datos perdidos (268). Todo esto con el objetivo de encontrar el algoritmo que se adapte mejor a los datos en las diferentes fases, según las fuentes de datos creadas y proponerlo dentro de la metodología.

Para la selección del algoritmo, se utilizó la herramienta RapidMiner, que cuenta con el proceso denominado “Auto Model”, el

cual se aplicó a las 3 fases. Este proceso tiene las siguientes secciones: Selección de los datos, Selección de tareas (acá se puede utilizar "Predict", "Cluster" o "Outliers" dependiendo del objetivo a resolver) y que para el ejemplo se utilizó "Predict", luego preparación del objetivo, selección de entradas, tipos de modelos y finalmente resultado.

3.6 Evaluación

Para Chapman (2000) en esta etapa se ha construido un modelo (o modelos) que parece tener una alta calidad desde una perspectiva de análisis de datos. Antes de continuar con el despliegue final del modelo, es importante evaluarlo exhaustivamente y revisar los pasos ejecutados para crearlo, para asegurarse de que el modelo logre los objetivos comerciales correctamente. Al final de esta fase, se debe tomar una decisión sobre el uso de los resultados de la extracción de datos (Chapman, y otros, 2000).

Para aplicar lo anterior, en esta etapa se evalúan los diferentes resultados obtenidos en la etapa previa y mediante las siguientes medidas: Accuracy, Sensitivity, Recall, Specificity, Precision, Classification Error (Kotu & Deshpande, 2015, pág. 260). Además de, AUC (Area Under the Curve) (Yang, Zhang, Lu, Zhang, & Kalui, 2017, pág. 74) y F Measure (Jiawei, Micheline, & Jian, 2012, pág. 369). Estas medidas entregan resultados en porcentajes, de donde se puede evaluar cada algoritmo y generar una matriz comparativa.

3.7 Despliegue

Para Chapman (2000) esta etapa a pesar de la creación del modelo, generalmente no es el final del proyecto. Incluso si el propósito del modelo es aumentar el conocimiento de los datos, el conocimiento adquirido deberá organizarse y presentarse de manera que el cliente pueda utilizarlo. A menudo, implica la aplicación de modelos "en vivo" dentro de los procesos de toma de decisiones de una organización. Dependiendo de los requisitos, la etapa de implementación puede ser tan simple como generar un informe o tan compleja como implementar un proceso de minería de datos repetible en toda la empresa. Además, para North (2016), es en esta etapa donde se realizan las acciones de lo que se aprendió del modelo o algoritmo gracias a las evaluaciones obtenidas (North, 2016, pág. 74). También gracias a Rapidminer y su simulador, se pueden utilizar los modelos usados y analizar posibles acciones a realizar, enfocadas, por ejemplo, en la retención, titulación oportuna u otro indicador que se estime conveniente.

4. DISCUSIÓN DE RESULTADOS

La herramienta utilizada Rapidminer, gracias a su simulador, entrega diferentes resultados, en base a las tareas seleccionadas, ya sea Predictiva (Predict) o Descriptiva (Cluster o Outliers) y se obtienen diferentes algoritmos, por ejemplo para la tarea Predict se tiene NaiveBayes, Generalized, Linear Model, LogisticRegression,

DeepLearning, DesiciónTree, RandomForest y GradientBoostedTrees; para la tarea de Cluster se tiene K-Means Clustering y X-Means Clustering y finalmente para la tarea Outliers se tiene Distance-based Outlier Detection y Local Outlier Factors. Además, cada uno de estos algoritmos tiene diferentes mediciones, por ejemplo Accuracy permite ver el rendimiento del algoritmo, Sensitivity selecciona lo que debe seleccionar, Recall entrega la proporción de todos los casos relevantes, Specificity tiene la capacidad de rechazar lo que se debe rechazar, Precision proporciona casos encontrados que fueron relevantes, AUC mide el rendimiento en función de la frecuencia de los pares de instancias erróneas, Classification Error muestra el complemento de Acurracy y F Measures es una medida ponderada entre Precisión y Recall. Con esta información se puede generar una matriz comparativa, antes mencionada, como se observa en tabla 3.

Tabla 3. Propuesta de Matriz Comparativa

Algoritmo	**Medición 1**	**Medición 2**	**…**	**Medición N**
Algoritmo 1	%	%	…	%
Algoritmo 2	%	%	…	%
…	…	…	…	…
Algoritmo N	%	%	…	%

Fuente: Elaboración propia

La matriz comparativa variará según la tarea y los métodos (algoritmos) utilizados, así se tiene un panorama general de qué algoritmo es el mejor para la tarea seleccionada y con esto resolver algún objetivo, ya que al proponer una metodología de MD se debe tener claridad en términos de estas tareas y métodos.

Luego, y en base al ejemplo planteado, se presentan 3 matrices comparativas en torno a las 3 fases planteadas en la etapa de modelamiento.

Tabla 4. Mediciones de los modelos en RapidMiner, fase 1

Modelo	**Accuracy**	**Classification Error**	**AUC**	**Precision**	**Recall**	**F Measure**	**Specificity**
Naive Bayes	77,8%	22,2%	0,81	97,1%	79,1%	87,2%	50,0%
Generalized Linear Model	95,6%	4,4%	0,83	95,6%	100,0%	97,7%	0,0%
Logistic Regression	91,1%	8,9%	0,44	95,3%	95,3%	95,3%	0,0%
Deep Learning	93,3%	6,7%	0,63	95,5%	97,7%	96,6%	0,0%
Desición Tree	93,3%	6,7%	0,49	95,5%	97,7%	96,6%	0,0%
Random Forest	95,6%	4,4%	0,73	95,6%	100,0%	97,7%	0,0%
Gradient Boosted Trees	95,6%	4,4%	0,68	95,6%	100,0%	97,7%	0,0%

Tabla 5. Mediciones de los modelos en RapidMiner, fase 2

Modelo	**Accuracy**	**Classification Error**	**AUC**	**Precision**	**Recall**	**F Measure**	**Specificity**
Naive Bayes	83,0%	17,0%	0,63	93,6%	88,0%	90,7%	0,0%
Generalized Linear Model	94,3%	5,7%	0,62	94,3%	100,0%	97,1%	0,0%
Logistic Regression	90,6%	9,4%	0,73	95,9%	94,0%	94,9%	33,3%
Deep Learning	94,3%	5,7%	0,71	94,3%	100,0%	97,1%	0,0%
Desición Tree	94,3%	5,7%	0,50	94,3%	100,0%	97,1%	0,0%
Random Forest	94,3%	5,7%	0,47	94,3%	100,0%	97,1%	0,0%
Gradient Boosted Trees	94,3%	5,7%	0,63	94,3%	100,0%	97,1%	0,0%

Tabla 6. Mediciones de los modelos en RapidMiner, fase 3.

Modelo	Accuracy	Classification Error	AUC	Precision	Recall	F Measure	Specificity
Naive Bayes	79,2%	20,8%	0,71	93,3%	84,0 %	88,4%	0,0%
Generalized Linear Model	92,5%	7,5%	0,70	94,2%	98,0 %	96,1%	0,0%
Logistic Regression	75,5%	24,5%	0,0	93,0%	80,0 %	86,0%	0, 0%
Deep Learning	92,5%	7,5%	0,68	94,2%	98,0 %	96,1%	0,0%
Desición Tree	94,3%	5,7%	0,50	94,3%	100, 0%	97,1%	0,0%
Random Forest	94,3%	5,7%	0,43	94,3%	100, 0%	97,1%	0,0%
Gradient Boosted Trees	94,3%	5,7%	0,59	94,3%	100, 0%	97,1%	0,0%

Según lo observado en la Tabla 4, Tabla 5 y Tabla 6, para las tres fases, los modelos basados en árboles y el modelo lineal generalizado tuvieron una mejor Acurracy en términos de predicciones correctas, esto es, sobre el 95%. Además, los menores errores de clasificación para las fases se mantienen las técnicas basadas en árboles y modelo lineal generalizado. Para AUC, se debe analizar el resultado que este más cercano a 1, para este caso, en las tres fases, fue el modelo lineal generalizado con 0,83. En términos de Presicion el modelo que se comportó mejor para la fase 1 fue Naives Bayes con un 97,1%. Para Recall, nuevamente se comportan mejor los modelos basados en arboles con 100% de predicciones de verdaderos positivos, para todas las fases, y en particular Random Forest y Gradient Boosted Trees. La F measure indica entre más alto Recall, más alta Presicion, lo que significa que identifica a la mayoría que pueden tener matrícula SI, para el caso de ejemplo, donde la clase fue matrícula. Lo anterior sucede con los modelos basados en árboles, por sobre el 97% en las

tres fases, pero no así en el resto de los modelos, variando en una porción menor. En términos de Specifity, Jiawei la define como "tasa negativa verdadera" (Jiawei et al., 2012, p. 367), esto quiere decir la proporción de tuplas negativas que están identificadas correctamente, así se pued2085 e decir que en la fase 1 el mejor clasificador fue Naives Bayes con un 50,0%, en fase 2 fue Logistic Regression con un 33,0% y finalmente en fase 3 no identifica tuplas negativas correctamente, lo que se debe a la baja la cantidad de matrículas "NO" que existen en el conjunto de datos analizado.

Junto a lo anterior, en la etapa de resultados de Auto Model de Rapidminer, entrega más información para la toma de decisiones, como por ejemplo para Predict existe una visión general de la mediciones y tiempos requeridos por cada algoritmo, una comparación de ROC (Receiver Operating Characteristic) que provee herramientas que permiten seleccionar el subconjunto de algoritmos que tienen un comportamiento óptimo general (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004). Para las tareas de Cluster y Outlier existe un análisis de los datos utilizados, las correlaciones existentes y los resultados mismos de los métodos (algoritmos). Además, para cada método se cuenta con el Modelo, Simulador, Rendimiento (performance) y Lift Chart (gráfico de elevación) que representa la mejora que proporciona un modelo de minería de datos en comparación con una estimación aleatoria.

Finalmente, con esta información, se pueden tomar decisiones, tal como lo hizo Marcano Aular (2007), donde en base a indicadores se

puede aplicar minería de datos y además, los autores plantean que "una organización que reflexiona, documenta y aprende, está en condiciones de innovar y obtener ventajas competitivas" (Marcano Aular & Talavera Pereira, 2007, pág. 111). Así, para el ejemplo planteado, la Universidad puede identificar a nuevos o antiguos estudiantes que estén en riesgo de irse de la institución, con lo cual se pueda aplicar acciones remediales, como por ejemplo, acompañamiento, tutorías, etc, pero de manera personalizada por estudiante.

5. CONCLUSIONES

La cantidad de datos existentes en cualquier organización es enorme. Aun así, no es suficiente si lo que se desea es obtener conocimiento de los datos. Es por ello, que las metodologías, tareas, métodos (algoritmos) de minería de datos entregan mecanismos para comprender los datos, con el fin de mejorar la toma de decisiones en las organizaciones. Además, una metodología no solo aporta desde la formalidad de acciones a realizar, sino también puede generar nuevos procesos dentro de las organizaciones en favor de los datos.

Si se utilizan de forma correcta los algoritmos que proporciona la minería de datos en base a las mediciones, se comprenden mejor los datos y sus posibles acciones a realizar. En este sentido, Rapidminer nos proporciona una manera sencilla, pero de alta calidad técnica, nuevo conocimiento en base a diferente información dentro del software, para que tanto personal no calificado como personal

involucrado con el análisis de datos, pueda utilizar todos los recursos que entrega esta herramienta

En base al objetivo de este trabajo, se logra utilizar una metodología de MD y datos reales de las Ingenierías Civiles de la Sede Iquique, así con todo el nuevo conocimiento generado se pueden realizar acciones para mejorar los indicadores claves, en base a pasos formales y establecidos.

6. REFERENCIAS BIBLIOGRÁFICAS

Azevedo, A., & Santos, M. 2008. KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conference Data Mining, 182-185. Disponible en: http://recipp.ipp.pt/handle/10400.22/136. Consultado el: 26/01/2018.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. 2000. CRISP-DM 1.0 Step-by-step data mining guide. The CRISP-DM consortium.

Consejo Nacional de Educación. 2006. Consejo Nacional de Educación. Disponible en: https://www.cned.cl/file/1866/download?token=I9R8EP2L. Consultado el: 05/03/2018.

Dirección de Calidad Institucional. 2018. Universidad de Tarapacá. Disponible en: https://www.uta.cl/web/site/artic/20180629/asocfile/20180629155848/anuar2017.pdf. Consultado el: 17/03/2018.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. 1996. "From data mining to knowledge discovery in databases". AI Magazine, 17(3), 37-53. doi:https://doi.org/10.1609/aimag.v17i3.1230.

Galán Cortina, V. 2015. Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario (Tesis de Pregrado). Universidad Carlos III. Madrid.

Goicoechea, A. 2002. Imputación basada en árboles de clasificación. Disponible en: http://www.eustat.eus/document/datos/ct_04_c.pdf. Consultado el: 28/09/2018.

Hernández Orallo, J., Ramírez Quintana, M., & Ferri Ramírez, C. 2004. **Introducción a la Minería de Datos**. Madrid: Pearson Educación .

Himmel K., E. 2002. "Modelos de análisis de la deserción estudiantil en la educación superior - Retención y movilidad estudiantil". Revista Calidad En La Educación, 91-108. Disponible en: http://www.alfaguia.org/alfaguia/files/1318955602Modelo de analisis de la desercion estudiantil en la educacion superior.pdf. Consultado el: 17/04/2018.

Jiawei, H., Micheline, K., & Jian, P. 2012. **Data Mining: Concepts and Techniques**. Waltham, MA: Morgan Kaufmann.

Kotu, V., & Deshpande, B. 2015. **Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner**. Waltham, MA: Morgan Kaufmann. doi:https://doi.org/10.1016/C2014-0-00329-2

Lohr, S. 1999. **Sampling: Design and Analysis**. New York: Cengage Learning.

Marcano Aular, Y. J., & Talavera Pereira, R. 2007. "Minería de Datos como soporte a la toma de decisiones empresariales". Opción, 23(52), 104-118.

North, M. 2016. **Data Mining for the masses, second edition, with implementations in Rapidminer and R**. Middletown: CreateSpace Independent Publishing Platform.

Piatetsky, B. 2014. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Disponible en: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html. Consultado el :15/08/2018.

Rapidminer. 2018. Lightning Fast Data Science for Teams. Disponible en: https://rapidminer.com/. Consultado el: 18/03/2018.

Rodríguez, V., González Campos, J., & Patricio Aguilera, J. 2017. Modelo Predictivo para la Permanencia en la Educación Superior. Disponible en: http://revistas.utp.ac.pa/index.php/clabes/article/view/1588. Consultado el: 20/05/2018.

Rosado, A., & Verjel, A. 2017. "APLICACIÓN DE LA MINERIA DE DATOS EN LA EDUCACION EN LINEA". Revista Colombiana de Tecnologías de Avanzada, 1(29). doi:https://doi.org/10.24054/16927257.v29.n29.2017.2491.

Universidad de Tarapacá. 2012. Modelo Educativo. Disponible en: https://www.uta.cl/adjunto/mei.pdf. Consultado el: 21/03/2018.

Useche, L., & Mesa, D. 2006. "Una introducción a la imputación de valores perdidos". Terra, XXII(31), 127-152.

Yang, Z., Zhang, T., Lu, J., Zhang, D., & Kalui, D. 2017. "Optimizing area under the ROC curve via extreme learning machines". Knowledge-Based Systems, 130, 74–89. doi:https://doi.org/10.1016/j.knosys.2017.05.013.

Zhang, S. 2012. "Nearest neighbor selection for iteratively kNN imputation". Journal of Systems and Software, 85(11), 2541–2552. doi:https://doi.org/10.1016/j.jss.2012.05.073.

UNIVERSIDAD DEL ZULIA

opción

Revista de Ciencias Humanas y Sociales
Año 35, Especial No. 25 (2019)

Esta revista fue editada en formato digital por el personal de la Oficina de Publicaciones Científicas de la Facultad Experimental de Ciencias, Universidad del Zulia.
Maracaibo - Venezuela

www.luz.edu.ve

www.serbi.luz.edu.ve

produccioncientifica.luz.edu.ve