ISSN 1870-4069

Uso de datos abiertos y técnicas de minería de datos para la clasificación de estudiantes de instituciones privadas de educación superior de Belem-PA

Matheus Ferreira Vasconcelos[1], Bruno Kenji Hosoda Mineshita[1], Edian Franklin Franco de Los Santos[2], Alan Marcel Fernandes Souza[1]

[1] Universidade da Amazônia, Centro de Ciências Exatas, Belém, Pará, Brasil

[2] Universidade Federal do Pará, Centro Genômica e Biologia de Sistemas, Belém, Pará, Brasil

vasconcelosmf@outlook.com, hosoda_kenji1997@hotmail.com, edianfranklin@gmail.com, alan.souza@unama.br

Resumen. Las universidades privadas enfrentan el desafío de reducir la tasa de evasión de estudiantes de los cursos de grados. Si esta tasa es muy alta, el beneficio de estas universidades puede reducir drásticamente, alejando a los accionistas e inversores. Este trabajo tiene como objetivo aplicar técnicas de minería de datos clasificar entre estudiantes que abandonan y los que terminan los cursos de grados, para facilitar la identificación de los posibles factores de riesgos que influyen en la evasión universitaria. Para el análisis fueron utilizados datos abiertos extraídos del portal del gobierno y diversos algoritmos de minería de datos. En los experimentos realizados, fue posible seleccionar la técnica más eficiente de clasificación para este problema, mediante el análisis de métricas de confianza, consistencia y calidad. Además, fue realizado un levantamiento de los factores que tiene mayor incidencia en la evasión y abandono de los estudiantes de las universidades privadas de la ciudad de Belem do Pará (Brasil).

Palabras clave: Evasión de estudiantes, minería de datos, datos abiertos, análisis de métricas de calidad.

Usage of Open Data and Data Mining Techniques for the Classification of Students from Private High Education Institutions of Belém-PA

Abstract. The private universities face the challenge to reduce the evasion rate of the undergraduate students. If this rate is too high, the profit of these universities can drastically reduce, driving away shareholders and investors. This paper aims to apply data mining techniques to classify students who completed the course and those who dropped out, in order to facilitate the identification of possible risk factors that influence university dropout. The analysis used open data extracted from the government portal and various data mining algorithms.

In the experiments carried out, it was possible to select the most efficient classification technique for this problem, through the analysis of confidence, consistency and quality metrics. In addition, a survey was carried out of the factors that have a greater incidence in the evasion and abandonment of the students of the private universities of the city of Belem do Pará (Brazil).

Keywords: Student evasion, data mining, open data, analysis of quality metrics.

1. Introducción

La cantidad de estudiantes que ingresaron en una Institución de Enseñanza Superior (IES) y no llegan a graduarse es alarmante. Según el Ministerio de Educación [7], en la educación superior, hay más estudiantes de nuevo ingreso que concluyentes. En el año 2012 se registraron 7.037.688 alumnos matriculados, pero el total de alumnos que ingresó a las IES fue 2.747.089 y el número de concluyentes fue 1.050.413.

El número de matrículas en cursos de grado continuó creciendo en 2016, sin embargo, si se compara con años anteriores hubo una desaceleración. Según los estudios estadísticos, entre los años 2006 a 2016 hubo un aumento del 62,8% con un promedio anual del 5% en el número de matrículas, prevaleciendo el 75,3% en IES privadas. En 2016 se ofrecieron más de 10,6 millones de vacantes para estudios superiores, siendo el 73,8% nuevas y el 26% ya existentes, de las cuales solo el 33,5% de las nuevas vacantes y el 12% de las existentes fueron ocupadas por nuevos estudiantes. A partir de este estudio es posible diagnosticar que el número de vacantes ofertadas no implica la permanencia de los alumnos hasta la conclusión, lo que podemos considerar como evasión [8].

La evasión puede ser causada por diversos factores y variables que afectan la vida de un estudiante, basados en el trabajo de Couto y Santana [2], es evidente que las IES aún no cuentan con las informaciones o conocimientos de los motivos y factores que llevan a la evasión en la educación superior. Martins [6], en su trabajo estudió los diversos factores que pueden llevan a la evasión y clasificó la evasión en tres tipos: i. Evasión de carrera, la interrupción de la carrera por diversas situaciones (abandono, desistimiento, transferencia, re-opción, bloqueos, exclusión por normas institucionales), ii. Evasión de la institución (abandono o desistimiento de la IES en la cual está matriculado) y iii. Evasión del sistema (abandono de forma definitiva o temporaria de la educación superior). En este estudio serán considerados como estudiantes evasores, todos aquellos que abandonaron la institución o carrera de forma temporal o definitiva.

Este artículo utiliza datos abiertos, haciendo uso de cuatro técnicas de minería de datos para identificar las variables que más impactan en la evasión de los alumnos de IES privadas de Belém do Pará:

- Árboles de decisión (CART);
- Naïve Bayes;
- K-NN;
- Support Vector Machine (SVM).

Para la evaluación del desempeño de los modelos fueron utilizadas las siguientes metricas: precisión, sensibilidad, especificidad, valor predictivo positivo (PPV), valor predictivo negativo (VPN) y Matthew Coefficient Calculator (MCC).

Este trabajo está organizado de la siguiente manera: la sección 2 presenta los trabajos relacionados. En la sección 3, se muestran las etapas del proceso de descubrimiento de conocimiento en la base de datos. En la sección 4, los experimentos desarrollados utilizando las técnicas de minería de datos, siguiendo el análisis de los resultados. Por último, en la sección 5, se presentará la conclusión y los trabajos futuros.

2. Trabajos relacionados

Los trabajos correlacionados en esta sección se refieren a enfoques de métodos de minería de datos con el uso de diversos algoritmos de clasificación, buscando analizar el desempeño a través de métricas, para la identificación del mejor clasificador para el problema en cuestión: evasión de estudiantes de las IES.

En [3] realizaron un estudio de caso sobre la predicción de evasión de alumnos, usando datos del Electrical Engineering department of Eindhoven University of Technology junto a técnicas de Minería de datos. Para identificar a los estudiantes con riesgos de evasión en el estado inicial, se realizaron pruebas con 8 algoritmos, (CART (simple Cart), C4.5 (J48), redes bayesianas, modelos de regresión logísticas (Simple Logistic), reglas basadas en aprendizaje (JRip), Random Forest y OneR), en las pruebas todos mantuvieron una exactitud entre 75% - 80%.

En [14] presentaron un estudio utilizando registros de alumnos registrados en el Online Information Technologies Certificate Program en el periodo 2007-2009, con el objetivo de prevenir la evasión de estudiantes, a través de la utilización de métodos de minería de datos. Se realizaron pruebas con cuatro algoritmos de clasificación (K-NN (k = 3), Árboles de decisión, Naïve Bayes y Redes Neuronales). Las sensibilidades de detección de los algoritmos fueron, respectivamente, el 87%; 79,7%; el 76,8% y el 73,9%. Los modelos fueron entrenados, probados y validados con utilizando validación cruzada (10-fold).

El trabajo de Couto e Santana [2] presenta una serie de pruebas realizadas con datos proporcionados por el SIGAA (Sistema de Gestión de Actividades Académicas) referentes a los estudiantes que cursos de grado que ingresaron hasta el año 2016, con el objetivo de creación de subsidios para auxiliar las IES e identificar a los alumnos con riesgo de evasión utilizando Knowledge Discovery Databases (KDD).

Se realizaron pruebas con nueve algoritmos en la etapa de minería de datos, con la utilización de métricas para análisis de desempeño de los clasificadores, con el objetivo de identificar el mejor modelo para cumplir al objetivo del trabajo. Fue seleccionado un modelo basado en redes Bayesianas, con el consiguieron obtener una precisión global del 85%.

3. Etapas de la Minería de Datos

La minería de datos puede ser definida como un conjunto de técnicas que permiten la asociación y correlación de información de forma automatizada en grandes bases de datos. Esto hace posible el descubrimiento de patrones de manera más eficiente. [12]

La aplicación de técnicas de minería de datos forma parte del proceso de KDD. Sin embargo, para ello, se hace necesario pre-procesar, organizar y filtrar los datos brutos para que sea posible obtener un patrón, que permita el análisis e interpretación de los resultados obtenidos. De esta forma, es necesario seguir algunas etapas del proceso de KDD que se describen a continuación.

3.1. Selección de los datos

La etapa de selección de datos nos permite agrupar todas las fuentes de datos en una base de datos. Para Weiss [13], entender los datos que se utilizarán es crucial para el éxito en el desarrollo de las aplicaciones.

Los datos proporcionados por el Instituto de Estudios e Investigaciones Educativas Anisio Teixeira (INEP), fueron obtenidos de forma estructurada en el formato CSV (Comma-separated values) junto con sus metadatos, lo que facilito el entendimiento de los atributos y sus valores. La herramienta computacional utilizada para procesar los datos fue el RStudio versión 1.0.143, a través del lenguaje de programación R versión 3.4.1 (30/06/2017). Para el agrupamiento de datos, se creó un banco SQLite conteniendo los datos de las Instituciones de Educación Superior, los cursos y los estudiantes, obteniendo una base de datos con 265 atributos y 11.449.222 registros.

3.2. Preprocesamiento

Según Han e Kamber [5], con los datos organizados en una base de datos es necesaria la realización de etapas como: limpieza, integración, transformación y reducción del conjunto de datos.

La limpieza de los datos consiste en eliminar las inconsistencias encontradas: atributos incompletos, valores con errores, valores nulos, etc. En este caso, se optó por eliminar los registros con atributos incompletos, también fueron eliminados de varios conjuntos de datos, los atributos con pocas cantidad de instancia.

La integración de los datos consiste en un análisis en profundidad de los datos observando redundancia, categorías diferentes para los mismos atributos, atributos divergentes, datos repetidos, entre otros.

En el conjunto de datos utilizado gran parte de los atributos no fueron considerados por ser redundantes, debido a que presentaban el mismo valor o la misma información. Por ejemplo, el conjunto de datos utilizados contenía los atributos “sexo” y “cod_sexo”, los dos con la misma información, en este conjunto de datos solo fue preservado el primer atributo referente al sexo y el segundo fue eliminado por ser redundante.

La transformación de los datos se realiza de manera diferente de acuerdo con la técnica a ser aplicada, pues algunos algoritmos sólo funcionan con valores numéricos y otros con valores categóricos. Por lo tanto, se hace necesario la utilización de técnicas de: suavización, agrupamientos, generalización, normalización o la generación de

nuevos atributos (generados a partir de otros ya existentes). En R, las bibliotecas utilizadas para la creación de los modelos ya aplican algunas técnicas de transformación de los datos de forma intrínseca, de acuerdo con el algoritmo a ser utilizado. Sin embargo, se hicieron algunas conversiones de atributos con valores numéricos (ejemplo: 0 o 1) en datos cualitativos y la creación de nuevos atributos (por ejemplo: se sumaron las cantidades de empleados con maestría por sexo (femenina y masculino) y se creó un solo atributo: cantidad de empleados con maestría).

La reducción del conjunto de datos es importante, pues en algunos casos el volumen de datos utilizado en el proceso de minería de datos se considera muy grande hasta el punto de hacer inviable la aplicación de las técnicas minería de datos e incluso el análisis de los datos [6]. Este proceso de reducción de los datos permitió reducir el conjunto de datos original que tenía 256 atributos y 11,449,222 registros, para un conjunto de datos más manejable con 41 atributos y 25,468 registros, conteniendo solo las informaciones de las IES privadas de la ciudad de Belém-PA

La etapa de preprocesamiento es crucial para obtener un buen resultado y es la que demanda más tiempo. Según Olson e Delen [10], dicha fase puede comprender más del 50% del tiempo en proyectos de minería de datos.

3.3. Algoritmos de clasificación

Los algoritmos de clasificación crean un modelo que consigue determinar a cuál categoría pertenece una instancia específica. La clasificación generalmente es un proceso de aprendizaje supervisado, donde los datos se dividen en dos conjuntos: entrenamiento y pruebas [1]. En la fase de entrenamiento, los modelos son construidos utilizando el conjunto de datos de entrenamiento. Para evaluar el desempeño los modelos son testados utilizando del conjunto de prueba y diferentes métricas específicas.

Los algoritmos de clasificación obtienen resultados diferentes dependiendo del conjunto de datos utilizado, existen algoritmos que se adaptan mejor a un conjunto de datos que otros.

En este trabajo fueron utilizados los algoritmos de: arboles de decisión (CART), Naïve Bayes, K-NN y Support Vector Machine (SVM), debido a que presentaron un mejor desempeño en las pruebas realizadas con los conjuntos de datos en la fase de pre-procesamiento, además fueron seleccionadas algoritmos con distintas metodologías de clasificación, lo que permitió comparar y seleccionar el mejor modelo para los datos.

4. Clasificación de los alumnos

Esta sección presenta los experimentos desarrollados con los microdatos de censo de Educación Superior. Los experimentos comparan las métricas de los algoritmos utilizados, identificando si un estudiante concluyó o evadió la carrera. El objetivo del estudio es generar resultados para análisis y llegar a las variables que más impactan en la permanencia o no del alumno en una IES.

4.1. Experimentos

Para los experimentos se utilizaron más de 25 mil instancias con dos clases objetivo - Graduado y Evasor - siendo 17.921 alumnos graduados y 7.547 evasores.

Durante los experimentos, los datos fueron sometidos a la ejecución de los algoritmos de clasificación (Árboles de decisión (CART), Naïve Bayes, K-NN y Support Vector Machine (SVM)) con la finalidad de identificar cuál genera un mejor modelo de clasificación para los dos tipos de alumnos.

Los datos fueron divididos en un 70% para entrenamiento y un 30% para la prueba manteniendo la proporción de cantidades de registros en cada clase (Graduados y Evasores).

Para evitar la pérdida de patrones y tendencias importantes que pudieran aumentar el margen de error en la validación del modelo [4], fue utilizada la técnica de validación cruzada, que divide el conjunto en k partes iguales, donde k-1 partes se utilizan para entrenamiento y la parte restante se utiliza para la prueba, repitiendo hasta que todas las k partes se hayan utilizado para la prueba. En cada prueba se calcula la exactitud para obtener el mejor modelo que será utilizado con los datos de prueba. En los experimentos se utilizó diez para el valor de k.

4.2. Resultados

Los resultados presentados a continuación son referentes a la aplicación de los cuatro algoritmos de clasificación, utilizando los datos abiertos obtenidos del INEP. Fueron calculados los resultados de las pruebas realizadas, utilizando métricas distintas obtenidas a partir de la matriz de confusión resultante de cada modelo. A partir de las comparaciones, el modelo que obtuvo mayor desempeño fue utilizado para identificar los factores con mayor influencian o peso en la evasión de un alumno de grado, estos factores fueron clasificado en una escala 0 a 100.

Se aplicaron los cuatro algoritmos de clasificación y se generaron sus respectivas matrices de confusión a partir del conjunto de datos de prueba. Los modelos fueron evaluados a de la siguiente forma: en primer lugar, se generaron tres métricas: Exactitud (total de aciertos / total de datos en el conjunto), Sensibilidad (aciertos positivos / total de positivos) y Especificidad (aciertos negativos / total de negativos). La Fig. 1 muestra los resultados en porcentaje.

CA partir del experimento los mejores resultados fueron: para exactitud y sensibilidad - Árbol de Decisión con 82,65% y 92,87%, respectivamente, y para la especificidad el K-NN (k = 7) con 66,67%.

Después de las comparaciones con las métricas que calculan la tasa de confiabilidad del algoritmo en relación a su aplicación en la minería de datos. La Fig. 2 presenta los valores predictivos positivos (VPP) y los valores predictivos negativos (VPN). En el caso de que el VPP es representado por las fórmulas: VPP = aciertos positivos / total de predicción positivas y VPN = aciertos negativos / totales de aciertos negativos. [9, 11].

Según Oliveira e Kaestner [9] cuanto más similares los valores de VPP y VPN más consistentes son los resultados de la clasificación.

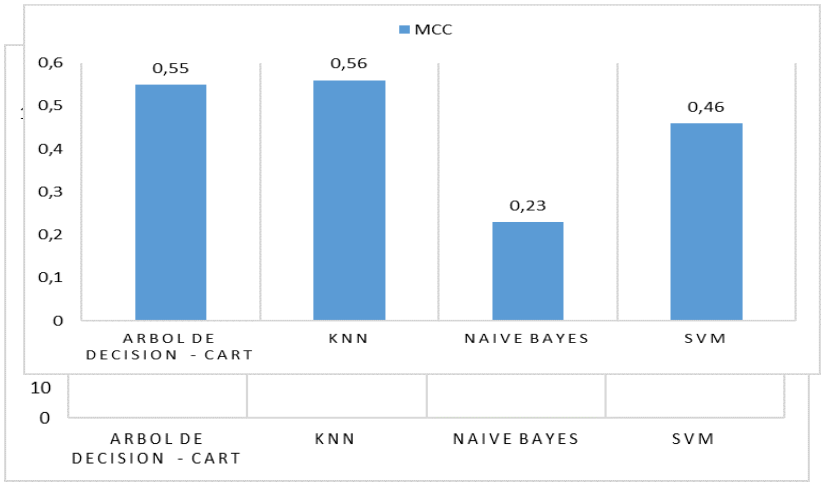

Fig. 1. Exactitud, Sensibilidad, Especificidad de los resultados.

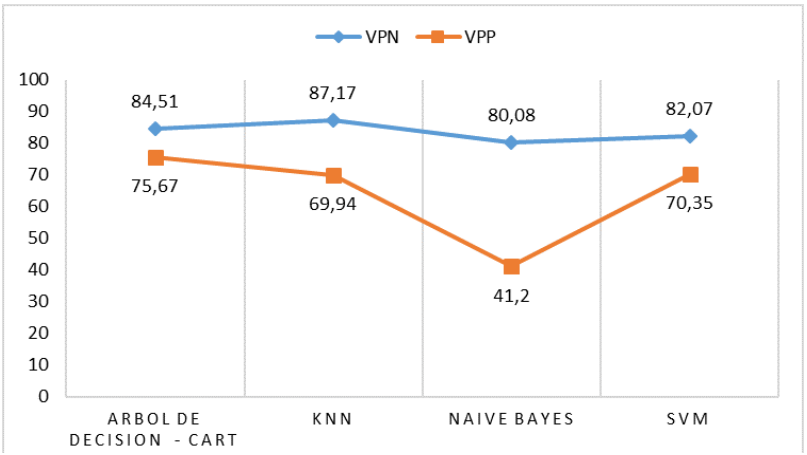

Fig. 2. VPP x VPN de cada clasificador.

Al analizar las dos métricas, que también se basan en la matriz de confusión, la que obtuvo los resultados más discrepantes fue Naïve Bayes, teniendo la mayor variación entre los cuatro modelos probados. El Árbol de Decisión (CART) y SVM fueron los que presentaron los mejores desempeños, obteniendo valores menos discrepantes.

Después del análisis de la consistencia, se calculó el Matthews Correlation Coefficient (MCC), también conocido como coeficiente de PHI, para cada modelo. El MCC es responsable de clasificar la calidad de los modelos. La métrica devuelve valores entre -1 y +1, siendo -1 para clasificación inapropiada, 0 para una clasificación aleatoria y +1 para clasificador correcto [11]. La Fig. 3 resume los valores de MCC encontrados para cada modelo.

El K-NN (k = 7) obtuvo el mejor resultado (0,56) seguido del árbol de decisión (0,55), siendo los dos mejores resultados comparado al SVM y NB. Este último no obtuvo un resultado muy satisfactorio.

En la comparación de los resultados presentados a través de las métricas que fueron utilizadas para clasificar el mejor algoritmo para solución del problema, fue identificado que dentro de los experimentos realizados dos algoritmos, obtuvieron los mejores resultados con valores similares, quedando por cuenta del Árbol de Decisión y K-NN, que obtuvieron valores superiores en cuestión de confiabilidad. Sin embargo, las matrices de confusión que presentaron los mejores resultados en términos de consistencia fueron el Árbol de Decisión y el SVM, con valores similares entre VPP y VPN. La diferencia de las métricas ocurre en el gráfico (Fig. 3) referente al MCC, donde el K-NN y el Árbol de Decisión obtuvieron los mejores resultados, 0,56 y 0,55 respectivamente.

Una mejor forma de observar el desempeño fue trazar una línea referente a la media de cada métrica y los algoritmos que posee la mayor cantidad de valores por encima de la media, presentan los mejores resultados, como muestra la Fig. 4. A través de estos análisis y pruebas se concluyó que el mejor algoritmo para la solución del problema es Árbol de Decisión, basados en las seis métricas utilizadas, este presentó un desempeño superior con relación a los demás algoritmos en cinco de las métricas.

Como el árbol de decisión fue el modelo que obtuvo los mejores resultados, se utilizó su modelo para identificar las variables que más impactan en la evasión de los estudiantes de las IES privadas. Para ello, se utilizó el valor de importancia que el algoritmo CART, que es calculado durante la construcción del Árbol de decisión. Esto permitió la construcción de un ranking de las diez variables que más impactan en la clasificación del alumno como evasor o graduado (Fig. 5).

A partir de análisis de la figura 5, donde son presentadas las 10 variables más influyentes en la evasión del estudiante de una IES, objetivo principal de esta investigación, fue posible identificar que la edad es la variable con más peso. Esta situación que puede ser comprobada cuando se hace un análisis de las reglas obtenida en el modelo de árbol de decisión, en el cual fueron obtenidas las siguientes reglas:

```
22) NU_EDAD_ALUNO < 21.5 Evadió (0.74 0.26) 1%

7573) NU_EDAD_ALUNO >= 34.5 Graduado (0.29 0.71) 1%
```

En estas es posible percibir que cuando la edad de un estudiante es inferior a 22 años, la probabilidad de evasión es un 74%, en cambio cuando un estudiante posee una edad superior a 35 años la probabilidad de graduarse es de un 71%, basados en estas informaciones podemos inferir que estudiantes con edades más elevadas presentan una mayor probabilidad de graduarse en los cursos de grados.

Es importante esclarecer que a pesar de esta variable poseer el mayor peso, no podemos llegar a la conclusión que solamente esta variable es decisiva para predecir la evasión o conclusión de los estudios superiores de un estudiante, pues, existen otras variables que tienes incidencia en el resultado del modelo. El hecho de poseer mayor

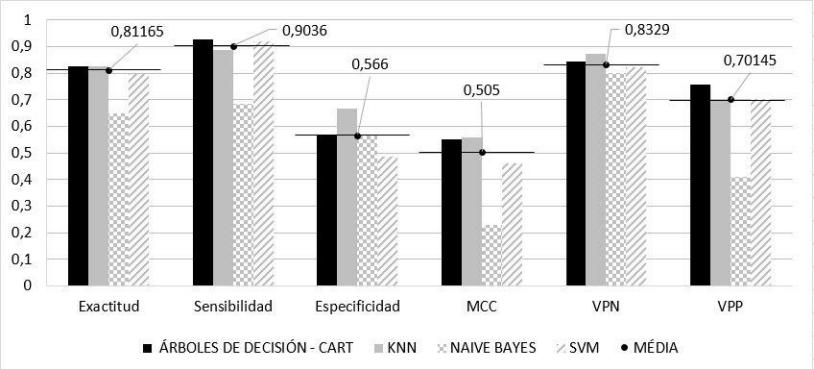

Fig. 4. Análisis de desempeño con todas las métricas calculadas.

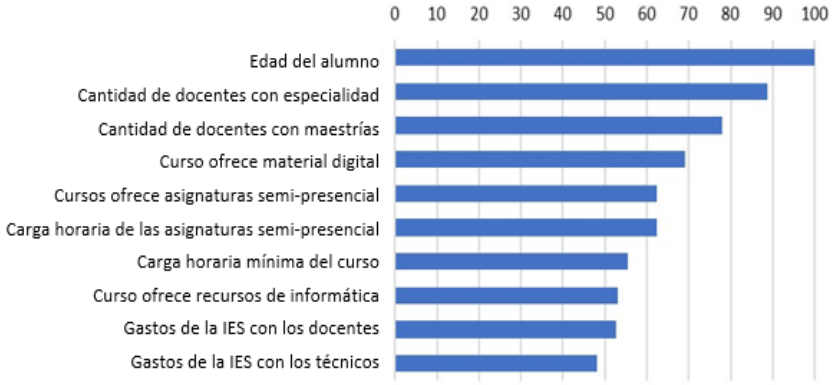

Fig. 5. Análisis de desempeño con todas las métricas calculadas.

peso significa que esa variable tuvo un mayor poder decisión con relación a las demás variables en el modelo.

5. Conclusiones y trabajo a futuro

Los experimentos realizados en esta investigación buscaron identificar al mejor clasificador, a partir de los resultados obtenidos por la minería de datos, referente a los datos abiertos del gobierno sobre evasión de estudiantes de grado de las IES privadas de Belém-PA.

Por lo tanto, a través de las pruebas realizadas con las técnicas de minería de datos, es posible concluir que el uso de diversas métricas y diversos algoritmos para analizar el conjunto de datos es viable y eficiente. Siendo posible identificar a través de experimentos el mejor algoritmo para cada proceso de minería, dependiendo del conjunto de datos obtenidos en la fase de preprocesamiento. Entre los experimentos realizados en este trabajo, el mejor resultado fue obtenido por el Árbol de Decisión (CART). A través de los experimentos el modelo generó una exactitud del 82,65% y

posibilitó identificar los factores de peso en la evasión de estudiantes de las IES de Belém do Pará. Entre los factores generados fue separado los diez más importantes (Fig. 5), posibilitando el análisis de cada uno individualmente por un especialista del área, con su grado de importancia entre 0-100.

Como trabajos futuros, se anhela:

- Consultar a un especialista para realizar una validación del conocimiento descubierto;
- Ampliar el conjunto de datos utilizados, considerando otras IES públicas y / o privadas de otras regiones;
- Crear un sistema web, con el modelo implementado para simulación de escenarios escogidos por el usuario.

Referencias

1. Camilo, C.O., Silva, J.C.: Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos /RT-INF_001-09.pdf, last accessed 2017/08/15 (2009)
2. Couto, D., Santana, A.: Mineração de Dados Educacionais Aplicada à Identificação de Variáveis Associadas à Evasão e Retenção. In: Araújo A.; Rebouças A.; Souza F.; Aguiar Y. II Congresso sobre Tecnologia na Educação, 1877, pp. 333–344 (2017)
3. Dekker, G., Pechenizkiy, M., Vleeshouwers, J.: Predicting Students Drop Out: A case Study. In: Barnes, T.; Desmarais, M.; Romero, C.; Ventura, S. 2nd International Conference On Educational Data Mining (EDM), 2, pp. 41–50 (2009)
4. Towards Data Science: Cross-Validation in Machine Learning. https://medium. com/towards-data-science/cross-validation-in-machine-learning-72924a69872f (2017)
5. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Cap. 2, pp. 47–103 (2006)
6. Martins, C.: Evasão dos alunos nos cursos de graduação em uma instituição superior.116f. Dissertação (Mestrado Profissional em Administração), Fundação Cultural Dr. Pedro Leopoldo, Minas Gerais (2007)
7. Ministério da Educação: http://portal.mec.gov.br/component/tags/tag/32123? limitstart=0 (2012)
8. Ministério da Educação: http://portal.mec.gov.br/index.php?option=com_docman &view=download&alias=71221-notas-sobre-censo-educacao-superior-2016-pdf& category_slug=agosto-2017-pdf&Itemid=30192 (2016)
9. Oliveira D., Kaestner, C.: Classificação Automática das Reclamações de Clientes de uma Empresa de Telecomunicações. In: Computer on the Beach, pp. 230–238 (2017)
10. Olson, D., Delen, D.: Advanced Data Mining Techniques. Springer (2008)
11. Santos, C.: Avaliação do uso de classificadores para verificação de atendimento a critérios de seleção em programas sociais, 87f. Dissertação (Programa de Pós Graduação em Modelagem computacional) (2017)
12. Santos, R.: Conceitos de Mineração de Dados na Web. http://www.lac. http:.br/~rafael.santos/Docs/WebMedia/2009/webmedia2009.pdf (2009)
13. Weiss, G. M.: Data Mining in Telecommunications. In: Data Mining and Knowledge Discovery Handbook, pp. 1189–1201 (2005)
14. Yukselturk, E., Ozekes, S., Turel Y.: Predicting Dropout Student: an Application of Data Mining Methods in an Online Education Program. European Journal of open, Distance and e-Learning, 17, pp. 119–133 (2014)