Análisis de los índices de reprobación en la carrera de ITICS utilizando técnicas de inteligencia artificial y minería de datos en el tecnológico nacional de México campus Conkal

Analysis of failure rates in the ITICS career using artificial intelligence and data mining techniques at the tecnológico nacional de México campus Conkal

DOI: 10.46932/sfjdv3n6-069

Received in: November 23th, 2022
Accepted in: December 27th, 2022

Janet Guadalupe Pech de la Portilla
Doctora en Sistemas Computacionales
Institución: Universidad del Sur
Dirección: Avenida Tecnológico S/N Conkal, Yucatán, CP. 97345, Conkal - Yucatán, México
Correo electrónico: janet.pd@conkal.tecnm.mx

Eric Jesús Gamboa Várguez
Maestro en Ciencias en Planificación de Empresas y Desarrollo Regional
Institución: Instituto Tecnológico de Mérida
Dirección: Avenida Tecnológico, S/N, Conkal, Yucatán, CP. 97345, Conkal - Yucatán, México
Correo electrónico: eric.gv@conkal.tecnm.mx

Mario Rodolfo Chan Chí
Maestro en Maestría en Informática
Institución: Universidad Hispanoamericana Justo Sierra
Dirección: Avenida Tecnológico, S/N, Conkal, Yucatán, CP. 97345, Conkal - Yucatán, México
Correo electrónico: mario.cc@conkal.tecnm.mx

Carlos Humberto López May
Maestro en Ingeniería
Institución: Universidad de Guanajuato
Dirección: Avenida Tecnológico, S/N, Conkal, Yucatán, CP. 97345, Conkal - Yucatán, México
Correo electrónico: carlos.lm@conkal.tecnm.mx

Javier Antonio Martín Vela
Doctor en Ingeniería Eléctrica
Institución: Universidad de Guanajuato
Dirección: Avenida Tecnológico, S/N, Conkal, Yucatán, CP. 97345, Conkal - Yucatán, México
Correo electrónico: javier.mv@conkal.tecnm.mx

RESUMEN

Se analiza información académica identificando factores que influyen en los índices de reprobación y deserción de las y los estudiantes de la carrera de ITIC, utilizando técnicas de inteligencia artificial y minería de datos mediante el software WEKA. La fuente de datos contiene información de 4 semestres consecutivos realizando un análisis completo de las materias que conforman la carrera y de los docentes que participan. Se realiza la selección y depuración de datos, utilizando diferentes criterios de representación y aplicación de algoritmos de evaluación de atributos y de clasificación como árboles de decisión. Se identifican variables influyentes en los índices de reprobación y deserción, así como su relación con el desempeño académico, especialmente en los primeros años de la carrera. Entre los

resultados más destacados se observó que las materias de programación y electrónica son un alto referente en los índices de reprobación y deserción de las y los estudiantes.

Palabras clave: algoritmos de clasificación, índices de reprobación, inteligencia artificial, minería de datos, WEKA.

ABSTRACT
Academic information is analyzed to identify factors that influence the failure and dropout rates of ITIC students, using artificial intelligence techniques and data mining through WEKA software. The data source contains information from 4 consecutive semesters, performing a complete analysis of the subjects that make up the career and the teachers involved. Data selection and debugging is performed, using different representation criteria and the application of attribute evaluation and classification algorithms such as decision trees. Influential variables are identified in the failure and desertion rates, as well as their relationship with academic performance, especially in the first years of the career. Among the most outstanding results, it was observed that programming and electronics subjects are a high referent in the failure and dropout rates of students.

Keywords: classification algorithms, failure rates, artificial intelligence, data mining, WEKA.

1 INTRODUCCIÓN

El estudio y uso de la IA (Inteligencia Artificial) y la MDE (Minería de Datos en la Educación) ha tomado mayor relevancia en los últimos años, la utilización de técnicas de Minería de Datos (MD) permite deducir fenómenos dentro del ámbito educativo; de esta forma, es posible determinar la probabilidad de que un estudiante se convierta o no en un posible desertor. Este trabajo se centra en el análisis de variables relacionadas con los resultados académicos que obtuvieron los docentes al impartir las materias asignadas, así como el análisis de sus funciones académicas dentro de la Institución. Se propone la utilización de la IA con técnicas de clasificación en MD para detectar, cuáles son las características y los factores de mayor incidencia en los estudiantes de la carrera de ITIC, con relación al índice de reprobación y lo suscitado ahora con la pandemia del COVID-19 cuando las clases fueron 100% en línea, impactando en el abandono o reprobación de los estudiantes. Para ello, se propone la utilización de algoritmos de clasificación para una mayor confiabilidad de los resultados, con el propósito de extraer conocimientos de los datos disponibles en el entorno institucional y generar modelos predictivos que ayuden a la identificación de los semestres más riesgosos, características comunes de estudiantes en riesgo, materias con mayor reiteración en reprobación y si existe alguna particularidad con los docentes que imparten esas asignaturas donde los índices de reprobación o deserción fueron altos. La herramienta de IA utilizada para la investigación es WEKA, la cual se caracteriza por utilizarse bajo licencia GNU, y además de que esta herramienta fue diseñada específicamente para ser utilizada en investigación y con fines educativos. El paquete WEKA contiene una colección de herramientas de visualización, algoritmos para el análisis de datos, modelado

predictivo y descriptivo, unido a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades (Witten y Eibe, 2005).

2 DESCRIPCIÓN DEL MÉTODO

La investigación utiliza un enfoque cuantitativo, utilizando como el conjunto de datos a medir, los datos académicos obtenidos de los resultados semestrales de las y los docentes en los años 2020 y 2021, para lograr determinar mediante la combinación de inteligencia artificial, minería de datos, medición numérica y análisis estadístico, patrones de comportamiento y la comprobación de teorías en contextos específicos, con un alcance descriptivo que permita seleccionar una serie de datos con el fin de recolectar información para poder entender y medir las variables de la investigación y determinar factores en los índices de deserción o reprobación, “los estudios descriptivos buscan especificar las propiedades, las características y los perfiles de personas, grupos, comunidades, procesos, objetos o cualquier otro fenómeno que se someta a un análisis” (Hernández et al, 2006).

2.1 RECOLECCIÓN DE DATOS

El período seleccionado para el estudio corresponde a las materias impartidas los semestres enero-junio y agosto-diciembre 2020, enero-junio y agosto-diciembre 2021; podemos observar que los últimos tres semestres los alumnos estuvieron totalmente en línea debido a la pandemia presentada por el COVID-19, y la mitad del semestre enero-junio 2020 la concluyeron de la misma manera. El atributo para determinar los casos de reprobación, es decir si existen estudiantes que debieron la materia es de tipo dicotómico, deudor (“Deu”) y no deudor (“NoDeu”).

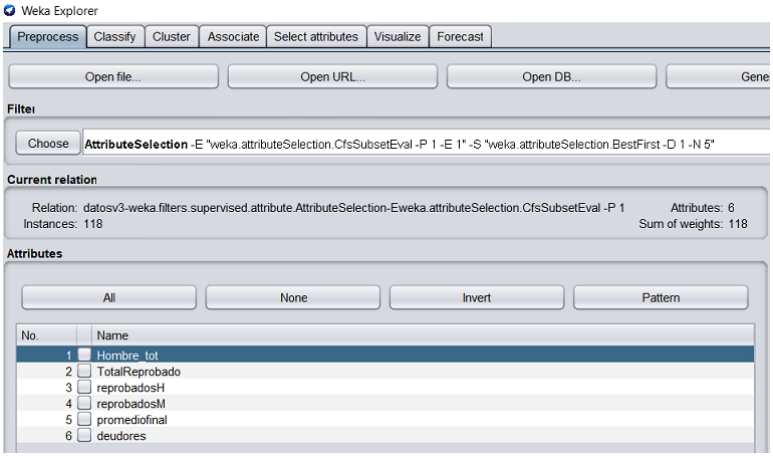
Los atributos recabados para utilizar en el presente estudio se muestran en el Apéndice, donde se incluye la descripción y la denominación estandarizada de cada atributo, así como el tipo de dato y valores posibles tanto nominales como numéricos.

2.2 TÉCNICAS DE MINERÍA DE DATOS

Para lograr una mejor integración, recopilación y filtrado de los datos se utilizaron dos técnicas de selección disponibles en la herramienta WEKA. La primera técnica utiliza algoritmos que se distinguen por su forma de evaluar los atributos, clasificándolos en filtros, donde se seleccionan y evalúan los atributos en forma independiente del algoritmo de aprendizaje, para determinar lo deseable de un subconjunto de datos, dentro de esta técnica aplicada utilizamos la denominada ‘Selección de Atributos’, la cual analiza que atributo o atributos en particular inciden sobre el atributo objeto (en este caso la condición de reprobación). Esto permite a su vez, optimizar posteriores pruebas y resultados a obtener con la técnica de clasificación, sobre todo para evitar clasificaciones muy complejas, como por ejemplo

árboles de decisión extensos y por ende difíciles de interpretar. Como resultado se obtuvieron 5 atributos relacionados con el índice de deserción: Hombre_tot, TotalReprobado, reprobadosH, reprobadosM y promediofinal, como se puede apreciar en la figura 1.

Figura 1: Técnica aplicada selección de atributos.



Se utilizaron otros métodos para corroborar los datos arrojados en la figura 1, se aplicó CfsSubsetEval y el de búsqueda BestFirst, que ofrecen una selección de subconjuntos de atributos de mayor calidad según (Witten y Eibe, 2005). Los resultados arrojados en lo referente a la deserción, fueron los mismos atributos.

Se realizaron otras pruebas siguiendo esta técnica, con la finalidad de evaluar diferentes factores que pueden incidir en la deserción, entre estas pruebas tenemos: el impacto del docente utilizándolo como valor nominal, los atributos arrojados fueron: materia, edad_docente, num_materias, grado_estudio, num_horas_base, puestoadministrativo, antiguedad_docente. El impacto de las materias como valor nominal, arrojando los siguientes atributos: docente, semestre y deudores. Finalmente se evaluó la antigüedad del docente como valor nominal, arrojando los siguientes atributos: cicloEscolar, edad_docente, Grado_estudio, num_horas_base y promediofinal.

Con todas las pruebas realizadas se puede distinguir cuales son los atributos que más preponderancia tienen en la información que se desea recabar. Con los atributos marcados en las pruebas anteriores dentro del programa WEKA se realizó una visualización de las relaciones obtenidas en cuanto a la reprobación por materia de las y los estudiantes en el 2020 y 2021. En estas pruebas se obtuvieron los siguientes resultados:

En la figura 2 correspondiente al 2020, se observa que la materia con mayor índice de reprobación es la de Circuitos Eléctricos y Electrónicos impartida en el cuarto semestre, con una población de 15 estudiantes y un índice de reprobación de 6, de los cuales 2 estudiantes ya no se inscribieron en el siguiente semestre.

En la figura 3 correspondiente al 2021, se observa que la materia con mayor índice de reprobación es la de Fundamentos de Programación impartida en el primer semestre, con una población de 20 alumnos y una reprobación de 14 alumnos de los cuales 10 alumnos, entre hombres y mujeres ya no se inscribieron en el siguiente semestre.

Figura 2. Materias reprobadas en el 2020.

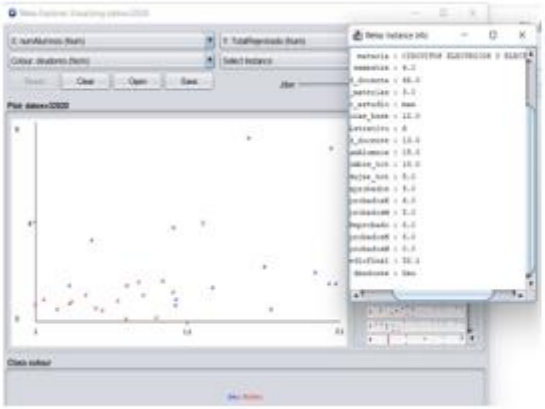
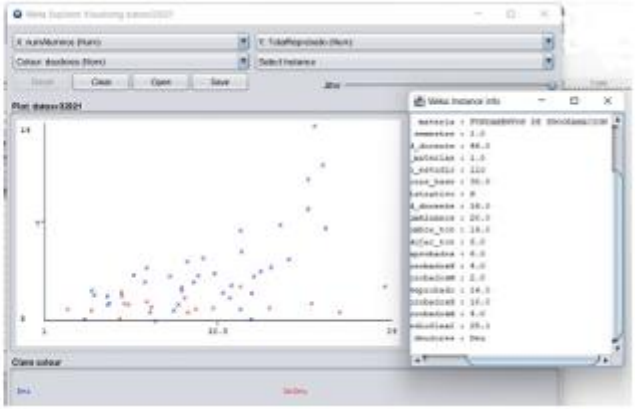


Figura 3. Materias reprobadas en el 2021.



Bajo este mismo esquema se realizaron otras visualizaciones sobre el análisis de reprobación, se obtuvo que las materias impartidas en el 2020 tuvieron mayor índice de reprobación comparadas con aquellas materias impartidas en el año 2021, donde las clases ya fueron 100% virtuales. También se analizó que en el año 2020 la mayor incidencia de los estudiantes que reprueban se da en los primeros 5 semestres, esto se observa en la figura 4. Sin embargo, en el año 2021, se observa que en todos los semestres hubo bastante incidencia de reprobación, algo muy notorio con la desestabilización que se dio con la pandemia, muchos estudiantes ya no continuaron, independientemente de estar en los últimos semestres de la carrera, esto se observa en la figura 5.

Figura 4. Reprobación del 1° al 5° semestre del 2020

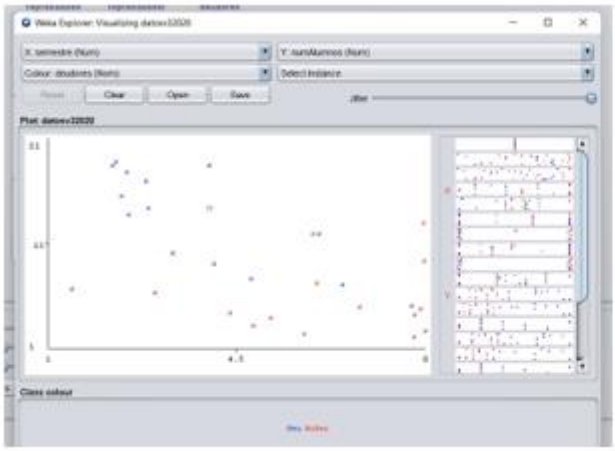
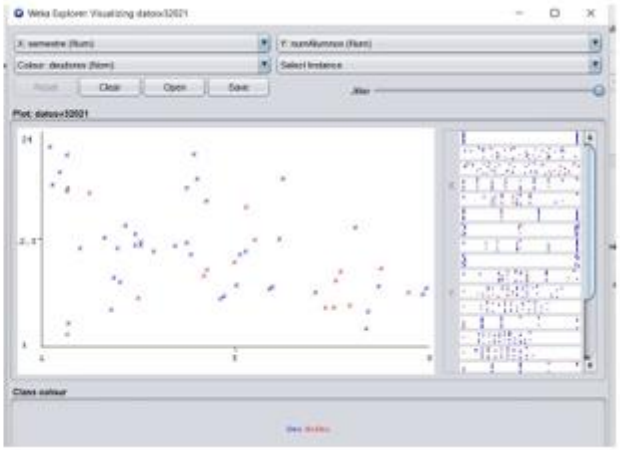


Figura 5. Reprobación todos los semestres en el 2021



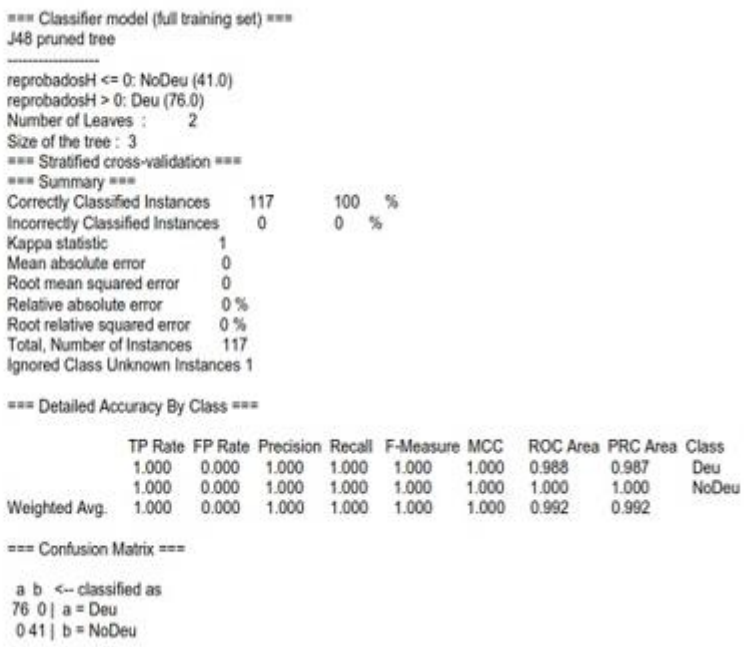
La segunda técnica utilizada es la de clasificación en minería de datos, esta es una técnica supervisada, donde generalmente se tiene un atributo llamado clase y se busca determinar si los atributos pertenecen o no a un determinado concepto. La clasificación, es la habilidad para adquirir una función que mapee (clasifique) un elemento de dato a una de entre varias clases predefinidas. Un objeto se describe a través de un conjunto de características (variables o atributos) $X \rightarrow \{X_1, X_2, \ldots, X_n\}$. El objetivo es clasificar el objeto dentro de una de las categorías de la clase $C = \{C_1, \ldots, C_k\}$, la función obtenida es f: $X_1 * X_2 * \ldots * X_n \rightarrow C$. (Segrera et al. 2005).

Para la presente investigación utilizando esta técnica de clasificación se trabajó con el Árbol de decisión. Técnica de clasificación supervisada, que permite determinar la decisión que se debe tomar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas. El árbol de decisión se construye, partiendo el conjunto de datos en dos o más subconjuntos de observaciones, después estos subconjuntos se vuelven a particionar empleando el mismo algoritmo. La raíz del árbol es el conjunto de datos inicial, los subconjuntos y sus subconjuntos conforman las ramas del árbol. El conjunto en el que se realiza una partición se llama nodo y permite bifurcar en función de los atributos y sus valores. Las hojas del árbol proporcionan predicciones. (Robles y Sotolongo, 2013). El algoritmo utilizado es el Clasificador J48.

Clasificador J48: Es un algoritmo de clasificación del árbol de decisiones de aprendizaje automático, es generado por C4.5 (una extensión de ID3). También se conoce como clasificador estadístico. Para estas pruebas se utilizó la integración de los años 2020 y 2021. Ejecutando este algoritmo en el programa WEKA (figura 6), se obtuvo el siguiente resultado mostrado en la figura 7, donde se aprecia que el valor más significativo de reprobación es en estudiantes del género masculino. En la matriz de confusión se observa una clara correlación de los datos. Obteniendo 76 instancias deudoras y 41 no deudoras del total del conjunto de datos utilizado.

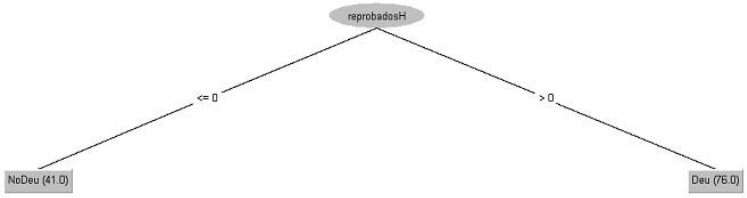
Figura 6. Ejecución del algoritmo J48

Figura 7. Resultados obtenidos.



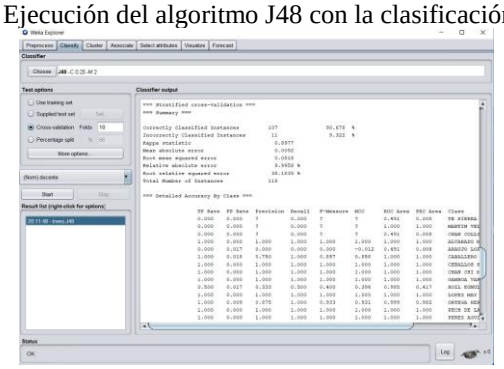
En la figura 8 se muestra el árbol generado, corroborando el mayor índice de reprobación esta dado en los hombres con clasificación Deu (deudores) y el número de instancias encontradas en el set de entrenamiento fue de 76.

Figura 8. Árbol de deudores y no deudores.



Si generamos un árbol para la distribución de los docentes obtenemos la información mostrada en la figura 9:

Figura 9. Ejecución del algoritmo J48 con la clasificación docente.



Este árbol nos arroja los siguientes parámetros:

Correctly Classified Instances	107	90.678 %
Incorrectly Classified Instances	11	9.322 %
Kappa statistic	0.8977	
Mean absolute error	0.0092	
Root mean squared error	0.0818	
Relative absolute error	9.9958 %	
Root relative squared error	38.1835 %	
Total Number of Instances	118	

Lo relevante que se observa es que se tienen 2 maestros que independientemente de tener más de 30 horas base no participan activamente en la carrera de ITIC, también haciendo un análisis más detallado se puede observar que maestros son los que tienen un mayor número de materias o menor número dentro de la carrera de ITIC.

3 COMENTARIOS FINALES

3.1 RESUMEN DE RESULTADOS

Este trabajo de investigación analizó información académica proporcionada por el departamento de Ciencias Económico-Administrativas de los años 2020 y 2021, identificando factores o situaciones que pueden influir en los índices de reprobación y deserción de los estudiantes de la carrera de Ingeniería en Tecnologías de la Información y Comunicación (ITIC) en el Tecnológico Nacional de México Campus Conkal, utilizando técnicas de inteligencia artificial y minería de datos por medio del software WEKA. Se identificaron como variables influyentes en la deserción, asignaturas aprobadas y reprobadas, cantidad de estudiantes aprobados o reprobados por semestre y divididos por sexo, promedio de asignaturas cursadas, antigüedad del docente, grado máximo de estudios del docente y número de horas contratadas del docente. Mediante este proceso fue posible identificar los atributos que caracterizan a los casos de deserción y su relación con el desempeño académico, especialmente en los primeros años de la carrera.

4 CONCLUSIONES

En cuanto al software utilizado podemos concluir que WEKA ofrece una amplia gama de conjuntos de datos de muestra para aplicar algoritmos de aprendizaje automático. Los usuarios pueden realizar tareas de aprendizaje automático como la clasificación, la regresión, la selección de atributos y la asociación en estos conjuntos de datos de muestra. En cuanto a la información relevante obtenida que ha caracterizado el índice de deserción, se pudo observar que la incidencia recae en los estudiantes varones, que los cinco primeros semestres son los más caóticos o donde mayor índice de deserción se ha obtenido, pero un dato muy interesante fue observar que ahora con la pandemia del COVID-19 y al estar 100% en línea, los alumnos independientemente del semestre en el que se encontraban presentaron muchos problemas el cual hizo que abandonaran sus estudios o debieran en la mayoría de las materias que se encontraban cursando en ese semestre. Otro dato por considerar es que los docentes que cuentan con un puesto administrativo, probablemente por sus funciones tienen un alto índice de reprobación y esto es algo que se debe de atender porque se sabe que la mayoría de estos docentes no dan asesorías fuera de la clase.

En cuanto al tipo de materia donde se observa alto índice de reprobación no son las materias aplicadas a las ciencias básicas entendiendo por estas las de cálculo, álgebra o probabilidad, sino más bien están impactando en las materias de programación y en las de electrónica, por lo que se debe poner atención en este tipo de materias y crear programas de asesorías como se ha desarrollado con las materias de ciencias básicas.

RECOMENDACIONES

La línea de investigación a trabajar será el desarrollo de herramientas didácticas que ayuden a fortalecer la educación. Determinar factores de deserción desde la perspectiva de los expertos. Es importante conocer la percepción que tiene los especialistas en temas de deserción en las universidades, con la finalidad de establecer nuevos factores que influyan negativamente en la decisión que toman los estudiantes de abandonar sus estudios. La aplicación de nuevas técnicas de Machine Learning como el Deep Learning podrían ser consideradas como una alternativa de mejora que permita la comparación de los algoritmos tradicionales de minería de datos. La propuesta de una metodología de predicción de la deserción estudiantil universitaria orientada a estudiantes universitarios con necesidades especiales.
Plantear estrategias para mitigar los efectos negativos de la deserción y potenciar la permanencia estudiantil en las universidades.

REFERENCIAS

Hernández, R., Fernández, C. y Baptista, P. *Metodología de la investigación*. (2006). http://dim.pangea.org/revistaDIM27/docs/AR27inglespreescolargemagutierrez2.pdf

Robles Y, Sotolongo A. Integración de los algoritmos de minería de datos 1R, PRISM E ID3 A POSTGRESQL. Gestión de Tecnología y Sistemas de Información. 2013; p. 389-406.

Segrera S, Moreno M, Miguel, L. Aplicación de la minería de datos en la evaluación de la aptitud física de las tierras para el cultivo de la caña de azúcar. III Taller Nacional de Minería de Datos y Aprendizaje. 2005; p. 349-358

Witten I. y F. Eibe. Data Mining: Practical Machine Learning Tools and Techniques, 2005. https://www.researchgate.net/publication/335572298_Data_mining_para_evaluar_el_riesgo_operativo_en_procesos_tecnologicos

APÉNDICE

Atributos seleccionados y estandarizados

Atributos Seleccionados	**Estandarización**	**Tipo de dato**
Ciclo Escolar	cicloEscolar	Numérico {1,2,3,4}
Nombre del Docente	docente	Alfanumérico
Materia	materia	Alfanumérico
Semestre de la Materia	semestre	Numérico {1,2,3,4,5,6,7,8,9}
Edad del Docente	edad_docente	Numérico
Número de materias impartidas en la carrera de TIC's en el semestre	num_materias	Numérico
Grado de estudios del docente	Grado_estudio	Nominal (lic, mae, doc).
Número de Horas Base del docente	num_horas_base	Numérico
Cargo Administrativo del docente.	PuestoAdministrativo	Nominal (N, S).
Antigüedad del Docente en el Instituto	antiguedad_docente	Numérico
Número de Alumnos Inscritos a la Materia	numAlumnos	Numérico
Desglose por sexo de Hombres Alumnos Inscritos a la Materia	Hombre_tot	Numérico
Desglose por sexo de Mujer Alumnos Inscritos a la Materia	Mujer_tot	Numérico
Total de alumnos aprobados	Totalaprobados	Numérico
Total de alumnos aprobados Desglose por sexo Hombres	aprobadosH	Numérico
Total de alumnos aprobados Desglose por sexo Mujer	aprobadosM	Numérico
Total de alumnos reprobados	TotalReprobado	Numérico
Total de alumnos reprobados Hombres	reprobadosH	Numérico
Total de alumnos reprobados Mujeres	reprobadosM	Numérico
Promedio Final de la materia obtenido por el grupo de estudiantes.	promediofinal	Numérico
Materia con deudores	deudores	Nominal (Deu, NoDeu).