Comparación de Algoritmos de Aprendizaje Supervisado para la obtención de perfiles de alumnos desertores

Ing. Osvaldo Mario Spositto[1] y Lic. Julio César Bossero[2].

[1] Decano del Departamento de Ingeniería e Investigaciones Tecnológicas. Prof. Titular Cátedra Base de Datos. UNLaM. sposito@unlam.edu.ar
[2] Prof. Cátedra Elementos de Programación y Base de Datos. UNLaM. jbossero@unlam.edu.ar

Resumen

En los últimos años, las técnicas de Minería de Datos han sido progresivamente incorporadas al ámbito universitario, como consecuencia del gran volumen de datos del que se dispone y del notable esfuerzo que se requiere para realizar análisis de los mismos. Este artículo presenta la comparación del rendimiento de dos algoritmos del tipo Aprendizaje Supervisado. Por un lado una arquitectura de Red Neuronal Artificial, del tipo Perceptrón Multicapa y por el otro, una Máquina de Vectores de Soporte, con el fin de determinar cuál tiene mayor porcentaje de aciertos en la obtención de perfiles de posibles alumnos que abandonen tempranamente sus estudios universitarios. Con base en la información extraída de un Almacén de Datos Institucional, se construyó un modelo que fue implementado, utilizando la herramienta WEKA. La exactitud de predicción se obtuvo a través de Matrices de Confusión y Curvas ROC. Los modelos predictivos obtuvieron resultados superiores al 90% de aciertos en todos los casos.

1. Introduction

La Minería de Datos Educacionales (MDE), es una rama de la Minería de Datos (MD o DM[1]) la cual se ha dedicado a aplicar diversas técnicas para analizar datos provenientes de ambientes relacionados a la educación y a extraer la mayor cantidad de conocimiento para tratar de entender mejor a los estudiantes, profesores y actores relacionados, con el fin de mejorar los procesos educativos [1]. Según [2] la Minería de Datos es *"…un conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de grandes conjuntos de datos, con objeto de predecir, de forma automatizada, tendencias o comportamientos y descubrir modelos previamente desconocidos…*"

Uno de los problemas en lo que se incursionó en varios trabajos de investigaciones, fue el empleo de diversas técnicas de MD, para intentar entender el problema del desgranamiento y deserción universitaria [3-6]. Las técnicas que componen la MD se encuadran dentro de una etapa dentro de un proceso más amplio llamado Descubrimiento de Conocimiento en Bases de Datos o KDD[2], que Fayyad en [7] la define como "… *proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos*".

Para poder llevar a cabo las diversas técnicas de Minería de Datos de una manera adecuada, se necesitó un sistema de adquisición, almacenamiento y manejo de la información eficiente. Un Almacenes de Datos (AD o DW[3]), que según W. H. Inmon "…*es un conjunto de datos integrado, orientados a una materia, que varían con el tiempo y que no son transitorios, los cuales soportan el proceso de toma de decisiones de una administración…"* [8], un DW es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta. La creación de un DW representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de MD, [11].

La Universidad Nacional de La Matanza (UNLaM), a través del DIIT[4] desarrolló en los últimos años dos proyectos de investigación en relación a este tema [9 y 10]. Cabe aclarar que esta tecnología es utilizada por las autoridades para la toma de decisiones. De este repositorio es de donde se tomaron los datos académicos y socioeconómicos de los alumnos para realizar este estudio. Según el objetivo del análisis de los datos, las técnicas de MD se clasifican en dos grandes categorías de algoritmos de aprendizaje: los ***No Supervisados*** y los ***Supervisados.*** [11]

- ***No Supervisados*:** Están compuestos por algoritmos que aspiran a descubrir patrones y tendencias sobre el conjunto de datos, sin tener ningún tipo de conocimiento previo de la situación a la cual se quiere llegar.

[1] por sus siglas en inglés *Data Mining.*

[2] por sus siglas en inglés *Knowledge Discovery in Databases.*
[3] por sus siglas en inglés *Data Warehouse.*
[4] Departamento de Ingeniería e Investigaciones Tecnológicas.

- ***Supervisados***: estos algoritmos predicen el valor de un atributo, llamado etiqueta, dentro de un conjunto de datos, a partir de otros atributos denominados atributos descriptivos. Esta técnica permite deducir una función a partir de un conjunto de datos de entrenamiento.

Para predecir específicamente los posibles perfiles de alumnos desertores se ha empleado diferentes técnicas de MD, siendo una de la más utilizada las Redes Neuronales Artificiales (RNA o ANN[5]). Estos sistemas conexionistas realizan el procesamiento de la información con una estructura y funcionamiento que está inspirado en las redes neuronales biológicas, su funcionamiento consiste en un conjunto de elementos simples de procesamiento llamados nodos o neuronas conectadas entre sí por conexiones que tienen un valor numérico modificable llamado peso [12].

Por otra parte, a mediados de los años 90 se comenzó a utilizar el concepto de las Máquinas de Vectores de Soporte (MVS o SVM[6]), que pertenece a la familia de algoritmos de clasificación y regresión desarrollados por Vladimir Vapnik [11]. Las MVS han ganado un merecido reconocimiento gracias a los sólidos principios teóricos en los que se fundamenta su diseño, su alta capacidad de generalización y uso de funciones núcleo o kernel. Entre los kernels más comunes, se encuentran: la función lineal, polinomial y la RBF[7] entre otras. Las MVS son un método de clasificación supervisado, no paramétrico y aplicable a problemas de separación de clases lineal y no lineal [13]. Esta es una técnica ampliamente utilizada en el mundo, pero poco conocida o aplicada en este tipo de problemas.

A diferencia de las Redes Neuronales Artificiales que utilizan durante la fase de entrenamiento, el principio de Minimización del Riesgo Empírico (ERM[8]), las MVS se basan en el principio de Minimización del Riesgo Estructural (SRM[9]), la cual ha mostrado un mejor desempeño que el ERM, ya que las Máquinas de Vectores de Soporte minimizan un límite superior al riesgo esperado a diferencia del ERM que minimiza el error sobre los datos de entrenamiento [14].

Para la creación, ejecución y evaluación de los algoritmos se utilizó la herramienta WEKA[10] que permite de forma muy eficiente el procesamiento y clasificación de los datos. Este software permite evaluar la capacidad de los modelos a través de una Matriz de Confusión (MC), también llamada matriz de error o de contingencia. Esta es una matriz cuadrada de *n* x *n*, Figura 1, donde *n* es el número de clases. Las columnas de esta matriz indican las categorías clasificadas por el clasificador y las filas las categorías reales de los datos, por lo que los elementos en la diagonal principal se corresponden con las clasificaciones sin fallo y el resto de elementos son los errores de que el algoritmo ha cometido. [15]

Figura 1. Matriz de confusión diádica genérica.

Matriz de Confusión		*Clase Verdadera*	
		Positivos	*Negativos*
Clase Predicha	*positivos*	*VP*	*FP*
	negativos	*FN*	*VN*
Total columna		*P*	*N*

Valores

Verdaderos positivos (VP) Falsos positivos (FP)
Verdaderos negativos (VN) Falsos negativos (FN)

2. Metodología.

Este trabajo siguió la metodológica denomina "*Las fases del proceso de extracción de conocimiento*, que Hernández describe en su libro [11] las diferentes fases que componen un proceso KDD. Además señala que la MD es una de las fases más importantes del proceso. A continuación se describe brevemente cada fase y las tareas realizas en las distintas etapas para el desarrollo del presente trabajo:

2.1. Fase de integración y recopilación

En esta etapa se determinaron las fuentes de información que podrían ser útiles al proyecto y dónde conseguirlas. Como ya se mencionó, los datos se obtuvieron de un almacén de datos institucional. Se contó con una población de 718 alumnos, de la cohorte 2013, de los cuales 119 no presentaron actividad académica en el departamento en el año académico consecutivo (2014), considerándose así, como que los mismos, abandonaron sus estudios. Según Tinto [16] *"…un desertor es aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante tres semestres académicos consecutivos…"*. Como la UNLaM cuenta con un curso de verano, el requisito se cumple en un año civil.

2.2. Fase de selección, limpieza y transformación

Esta etapa consiste en la selección, limpieza y transformación de datos que puede aplicarse para remover el ruido y corregir inconsistencias de los

[5] por sus siglas en inglés *Artificial Neural Networks*.
[6] por sus siglas en inglés *Support Vector Machines*.
[7] por sus siglas en inglés *Radial Basis Function.*
[8] por sus siglas en inglés *Empirical Risk Minimization*
[9] por sus siglas en inglés Structural Risk Minimization.
[10] por sus siglas en inglés *Waikato Environment for Knowledge Analysis. http://www.cs.waikato.ac.nz/~ml/weka/*

mismos, que fueran extraídos desde sus distintas fuentes. En este trabajo, como ya se mencionó, se utilizó información proveniente de un DW departamental, por lo cual, la mayoría de las tareas descriptas anteriormente, se realizan como parte del proceso ETL[11] correspondiente a la carga del mismo. Este proceso está muy desarrollado en [17], donde Kimball explica su metodología, que es una de la más utilizada, para la creación de un DW y que el autor la denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle). La tarea que se llevó a cabo para este trabajo fue la confección de filtros para seleccionar sólo algunas carreras del DIIT y la selección de las materias comunes al primer año de cada carrera. Se transformaron, según los códigos correspondientes al tipo de nota y a la nota obtenida, al final de cada cuatrimestre, en valores discretos a las variables referida a las materias cursadas.

Con el fin de realizar el análisis comparativo del desempeño de un clasificador del tipo RNA frente a una MVS, se construyó una tabla ad hoc. En la tabla 1, se puede observar las variables utilizas, el tipo de dato y los posibles valores que pueden tomar. El objetivo de esta etapa es adecuar los datos para obtener buenos modelos para la aplicación a los algoritmos de minería de datos.

Tabla 1. Las variable utilizas y los posible valores que pueden tener.

Variable	Tipo de dato	Tipo Contenido	Valores
Cod_Sexo	Numérico	Discreta	1-2
EstadoCivil	Numérico	Discreta	1-2-3
Carrera	Numérico	Discreta	201-202-203-207
Edad	Numérico	Continuo	17-85
Alg_Geo_Analitica_I	Texto	Discreta	Cursada Promocionada Final Aprobado Cursada Aprobada Final Reprobado Final Ausente Cursada Reprobada Cursada Ausente No Cursada
Alg_Geo_Analitica_II	Texto	Discreta	
Analisis_Mat_I	Texto	Discreta	
Analisis_Mat_II	Texto	Discreta	
Computacion_Niv_I	Texto	Discreta	
Elementos_Prog_I	Texto	Discreta	
Mat_Discreta	Texto	Discreta	
Quimica_Gral	Texto	Discreta	
Tecn_Ing	Texto	Discreta	
Sist_Rep	Texto	Discreta	
Estado	Texto	Discreta	Abandono–No Abandono

A estos modelos se los conoce con el nombre de ***vista minable***, es decir, aquel conjunto de datos lo suficientemente libre de errores, ya filtrados, sin datos anómalos y cuyas variables sean lo más independientes entre sí.

2.3. Fase de minería de datos

[11] por sus siglas en inglés *Extract, Transform and Load.* es el proceso que se encarga de mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data Warehouse.

El objetivo del proceso de minería de datos es poder lograr crear conocimiento a partir de los datos recogidos con el fin de que este conocimiento sea utilizado por el usuario. Como se mencionó los algoritmos seleccionados para realizar este estudio fueron una Red Neuronal Artificial y una Máquinas de Vectores de Soporte.

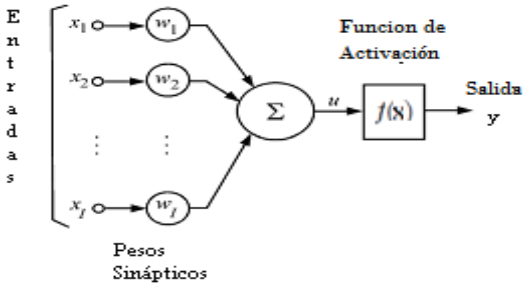
2.3.1. Redes Neuronales.

Las redes neuronales artificiales son sistemas de procesamiento de información, cuyo funcionamiento y estructura están basados en las redes neuronales biológicas de los animales. Se componen de un conjunto de elementos más simples denominados nodos o neuronas conectadas entre sí por un valor numérico modificable conocido como peso [10]. Debido a su fundamentación, las redes neuronales son capaces de aprender de la experiencia, generalizar a partir de casos anteriores y casos nuevos, y abstraer características relevantes a partir de un gran número de entradas que representan información irrelevante. Este algoritmo ya fue utilizado para este tipo de problema [18 y 19].

Modelo de Neurona Artificial

Warren McCulloch y Walter Pitts, en 1943 [11], concibieron un modelo abstracto y simple de una neurona artificial, este es el elemento básico de procesamiento en una red neuronal artificial, esta célula es considerada como un dispositivo con solo dos estados posibles: apagado (0) y encendido (1). En la figura 2 se muestra su esquema:

Figura 2. Estructura de una neurona tipo McCulloch-Pitts.



Esta neurona elemental es conocida como una unidad de umbral lineal (linear threshold unito ltu) y representa una familia de funciones parametrizada por los pesos w_1 y w_2,..., w_n en particular, la función de salida será:

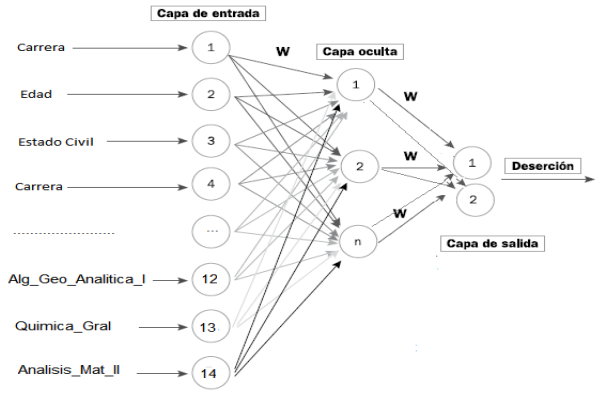
$$f(\underline{x}) = \begin{cases} 1 & \text{Si } \sum_{i=0}^{n} w_i x_i > 0 \\ 0 & \text{en otro caso} \end{cases} \qquad \textbf{(1)}$$

Una de las características que más interés despertó de este modelo fue su capacidad de aprender a reconocer patrones. El Perceptrón está constituido por conjuntos de sensores de entrada que reciben los patrones de entrada a reconocer o clasificar y una neurona de salida que se ocupa de clasificar a los patrones de entrada en dos clases, según que la salida de la misma sea **1** (activada) o **0** (desactivada). Sin embargo, dicho modelo tenía muchas limitaciones, como por ejemplo, no era capaz de aprender la función lógica XOR [13].

Si añadimos capas intermedias (ocultas) a un perceptron simple, obtendremos un Perceptron Multicapa (PM o **M**LP)[12]. El perceptrón multicapa es el modelo neuronal más conocido y empleado en la práctica y suele emplearse junto al aprendizaje de Retropropagación o (BP)[13] [13]. De forma simplificada, el funcionamiento de una red BP consiste de aprendizaje de un conjunto predefinido de pares de entradas-salidas dados como ejemplo, empleando un ciclo propagación-adaptación de dos fases: primero se aplica un patrón de entrada como estímulo para la primera capa de las neuronas de la red, se va propagando a través de todas las capas superiores hasta generar una salida que se desea obtener y se calcula un valor del error para cada neurona de salida. A continuación estos errores se transmiten hacia atrás, partiendo de la capa de salida hacia todas las neuronas de la capa intermedia en la salida original.

Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido un error que describa su aprobación relativa al error total. Basándose en el valor del error recibido, se reajustan los pesos de conexión de cada neurona, de manera que en la siguiente vez que se presente el mismo patrón, la salida esté más cercana a la deseada; es decir, el error disminuya. En la figura 3 se muestra el modelo propuesto para este trabajo.

Figura 3. Perceptron Multicapa.



[12] por sus siglas en inglés *Multi-Layer Perceptron.*
[13] *Backpropagation*

La red tiene la capacidad de auto-adaptar los pesos de las neuronas de las capas intermedias para aprender la relación que existe entre un conjunto de patrones dados como ejemplo y sus salidas correspondientes. Para poder aplicar esa misma relación, después del entrenamiento, a nuevos vectores de entrada con ruido o incompletas, dando una salida activa si la nueva entrada es parecida a las presentadas durante el aprendizaje. Su importancia se debe a su potencia y generalidad, pues se ha demostrado que constituye un **aproximador universal de funciones** [20], lo que hace de él uno de los modelos más útiles en la práctica.

2.3.2. Máquinas de Vectores de Soporte.

El método de las MVS está basado en los conceptos de aprendizaje estadístico propuestos por Vapnik y en la dimensión Vapnik-Chervonenkis [13]. Si bien los fundamentos datan de fines de los setenta el interés sobre este método ha crecido en los últimos años y hoy se encuentran varias buenas implementaciones y numerosos trabajos de investigación. Se ha resaltado la utilización de este método en volúmenes de datos muy grandes y particularmente en el análisis de imágenes [19-21]. La idea de esta técnica es lograr la separación lineal de los datos de aprendizaje elevando la dimensión del espacio vectorial y así eliminar las no linealidades originales del problema. Son similares a las redes neuronales salvo que utilizan el concepto de Kernel (núcleo o función) que se define según el tipo de problema. De hecho una MVS con un Kernel sigmoideo es equivalente a una red neuronal de dos capas con BP [20].

La idea que hay detrás de la clasificación binaria con MSV consiste en hallar una superficie de decisión que maximice la distancia *m* (*margen*) entre los patrones clasificados con el mínimo error de generalización. Por tanto, hallar el hiperplano de separación óptimo equivale a minimizar la norma euclídea de *w* que representa la distancia entre el hiperplano y las clases. Para encontrar una solución única al problema, Vapnik sugería recurrir a una formulación matemática que relacionara el hiperplano que maximiza el margen de separación con un determinado conjunto de entrenamiento.

Para formalizar esta idea matemáticamente y encontrar una solución única, se puede recurrir a una formulación que relacione el hiperplano que maximiza el margen de separación con un determinado conjunto de entrenamiento. Supongamos que tenemos dos posibles etiquetas por Y = {−1, 1} y un conjunto de vectores S = $\{(x_1,y_1);\ldots;(x_n,y_n)\}$ donde $x_i \in \Re^d$ e $y_i \in \{+1,-1\}$, para $i=1;\ldots;n$ se dice separable si existe algún hiperplano en

$\Re^d$ que separa[14] los vectores X = $\{x_1, \ldots x_n\}$ con etiqueta $y_1 = 1$ de aquellos con etiqueta $y_i = -1$. Es decir, aquellos puntos **x** que se encuentran en el hiperplano de separación que satisfagan la siguiente relación:

$$\mathbf{w} \cdot \mathbf{x} + b = 0 \qquad \textbf{(2)}$$

Donde:
w = *Vector normal y perpendicular al hiperplano de separación.*
x = *Vector de entrada.*
x * w = *Producto punto entre los dos vectores.*

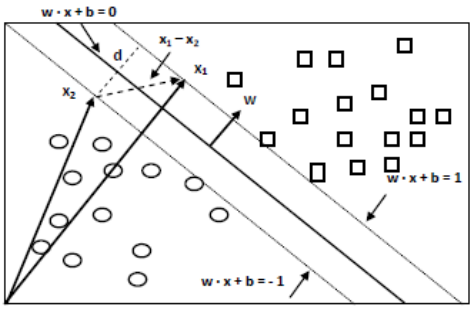
El margen, por lo tanto, se puede ver como la distancia entre las proyecciones perpendiculares del punto a la izquierda y el de la derecha más cercanos al hiperplano de separación. Para formular lo anterior, es posible suponer que los datos que pertenecen al conjunto de entrenamiento satisfacen las siguientes restricciones:

$$< w, x_i > + b \geq 0 \quad \text{si} \quad y_i = +1 \qquad \textbf{(3)}$$
$$< w, x_i > + b \leq 0 \quad \text{si} \quad y_i = -1 \qquad \textbf{(4)}$$

El planteamiento completo del problema considera, adicionalmente, los puntos para los que se cumple la igualdad en (3) y se encuentran en el hiperplano $H1$: $x_k \cdot w = 1$. También, se toman los puntos en los que se cumple la igualdad en (4) que se encuentran en el hiperplano $H2$: $x_k \cdot w + b = -1$. [21]

Con base en los conceptos anteriores se puede plantear el problema que encuentra una solución óptima y única para w y b, de tal forma que se maximiza el margen. En el caso bidimensional, la siguiente figura ilustra estos conceptos.

Figura 4. Problema de clasificación - Hiperplano.



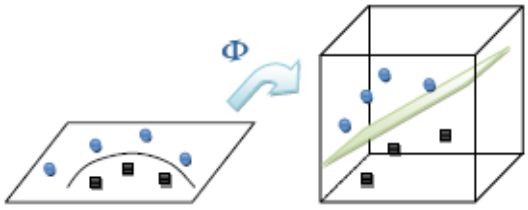
Cuando las ecuaciones (3) y (4) no son funciones lineales (no es posible separar linealmente los datos en su espacio de entrada), es necesario emplear un método que permita generalizar el uso de las MVS para resolver este tipo de problemas. Las MVS poseen una metodología para aplicar conjuntos de datos que tienen un límite de decisión no lineal, esta se basa en un procedimiento denominado kernel trick [25], que consiste en asignar a los datos otro espacio euclidiano ***H*** en el que aproximadamente se tiene una estructura lineal. Sea **Φ** una función tal que:

$$\Phi: \Re^n \rightarrow H \qquad \textbf{(5)}$$

Ahora el algoritmo depende de los datos por medio del producto punto en el espacio *H*; es decir, de las funciones de la forma $\boldsymbol{K(x_k \cdot x) = \Phi(x_k) \cdot \Phi(x).}$

Si se utiliza esta función ***K***, denominada función kernel, no será necesario especificar a **Φ**. Por lo tanto, el caso no lineal es equivalente a aplicar una función a los datos en el espacio de entrada. Posteriormente, el algoritmo de aprendizaje es utilizado en el espacio de llegada de la función (Figura 5).

Figura 5. Transformación de los datos por medio de un kernel.



Las funciones kernel utilizadas frecuentemente en el reconocimiento de patrones por medio de MVS son:

- Kernel lineal: Este es un clasificador lineal. Generalmente es utilizado para probar la existencia de no linealidades en el conjunto de entrenamiento. Además, el uso de este kernel se recomienda cuando hay vectores de datos dispersos:
$$K(x_i, x_j) = x_i * x_j \qquad \textbf{(6)}$$
- Kernel polinomial: Este es un método eficiente y simple para modelar relaciones no lineales. Tiene como desventaja que conforme aumenta el grado del polinomio (d) la superficie de clasificación se hace más compleja:
$$K(x_i, x_j) = (1 + x_i x_j)^d \qquad \textbf{(7)}$$
- Kernel RBF: Es uno de los kernels más utilizados en la literatura.
$$K(x, x_i) = e^{-\frac{1}{2}\left[(x^T x_i)^T \Sigma^{-1}(x - x_i)\right]} \qquad \textbf{(8)}$$

2.4. Fase de evaluación e interpretación

[14] En el sentido de dejar en dos regiones del espacio diferentes.

Existen varias herramientas confiables y de libre distribución que tienen ya implementado los algoritmos MLP y MVS [10]. Dichas bibliotecas permiten, entre otras cosas, definir los parámetros del algoritmo y/o elegir, como en el caso de las MVS, un tipo de kernel determinado. El software elegido para probar estos algoritmos es Weka, que fue desarrollada por el grupo de Aprendizaje Automático de la Universidad de Waikato en Nueva Zelanda.

La información recopilada sobre los valores de los indicadores para cada estudiante fue introducida en un archivo para su gestión. Los archivos que utiliza Weka tienen un formato estándar de datos denominado ***arff***[15]**,** para el procesamiento de la información [13]. En la opción Clasificación (***Classify),*** el software presenta una amplia variedad de algoritmos. Como ya se adelantó se compararon los resultados de una RNA del tipo MLP con 7 y 3 neuronas en la capa oculta y una MVS con 2 kernels diferentes, función de Base Radial (RBF) y otra función Polinomial. El resultado de aplicar el algoritmo de clasificación se efectúa comparando la clase predicha con la clase real de las instancias. Existen diversos técnicas para realizar la evaluación de los algoritmos, para el presente trabajo se emplearon las siguientes:

Cross-validation: evaluación con ***validación cruzada***. Se dividen las instancias en tantas carpetas como indica el parámetro ***Folds*** (10 en este ensayo), y en cada evaluación se toman las instancias de cada carpeta como datos de test, y el resto como datos de entrenamiento para construir el modelo. Los errores calculados serán el promedio de todas las ejecuciones.

Percentage split: evalúa la calidad del clasificador según lo bien que clasifique un porcentaje de los datos que se reserva para test.

3. Resultados

Para la discusión de los resultados, se comprobaron los mismos conjuntos de datos. Para validar los resultados de los experimentos con los métodos ***Percentage split*** y ***Cross-validation*** se utilizan medidas que cuantifican el desempeño del pronóstico para los dos modelos. Para ello se tiene en cuenta el concepto de matriz de confusión [10] figura 1. En las Tablas 2 y 3 se pueden observar los valores obtenidos en cada una de las matrices elaboradas por WEKA. Los verdaderos positivos (VP), corresponden a la cantidad de alumnos que ***no abandonan*** y que fueron clasificados de esa manera; los falsos positivos (FP), corresponden a la cantidad de alumnos que ***abandonan*** y fueron clasificados como que ***no abandonan;*** los verdaderos negativos (VN) atañen a la cantidad de alumnos que ***abandonan*** sus estudios y que clasificaron como que ***no abandonan***, y los falsos negativos (FN), a la cantidad de personas que ***abandona*** y se clasificaron en forma contraria.

Tabla 2. Resumen de WEKA para Percentage split

Matriz de confusión para el método: Percentage split	RNA MLP 7 Capas ocultas	RNA MLP 3 Capas ocultas	MVS kernel Polinomial	MVS kernel RBF
VP / FP	186 / 14	191 / 9	193 / 7	197 / 3
FN / VN	10 / 34	12 / 32	11 / 33	16 / 28
Exactitud:	90.16%	91.39%	92.62%	92.21%
Sensibilidad:	0.949	0.941	0.946	0.924
Especificidad:	0.708	0.780	0.825	0.903
Tiempo en seg:	9.39"	9.02 "	0.31"	0.44"
ROC Area:	0.946	0.953	0.858	0.811

Tabla 3. Resumen WEKA para Cross-validation

Matriz de confusión para el método: Cross validation	RNA MLP 7 Capas ocultas	RNA MLP 3 Capas ocultas	MVS kernel Polinomial	MVS kernel RBF
VP / FP	568 / 31	561 / 38	577 / 22	585 / 14
FN / VN	29 / 90	30 / 89	30 / 89	39 / 80
Exactitud:	91.64%	90.52%	92.75%	92.61%
Sensibilidad:	0.951	0.949	0.950	0.937
Especificidad:	0.743	0.700	0.801	0.851
Tiempo en seg:	9.14 "	9.39 "	0.29 "	0.48 "
ROC Area:	0.954	0.94	0.856	0.824

Para los experimentos de RNA y MVS se tienen en cuenta principalmente los siguientes indicadores de precisión de la clasificación total del modelo:

Exactitud: Se calcula como el número de unidades clasificadas correctamente, sobre el número total de unidades consideradas. Se obtiene sumando los elementos de la diagonal divididos por el Total de observaciones. Este índice tiende a sobrestimar la bondad de la clasificación. Sus valores se encuentran en el intervalo [0, 1], siendo la clasificación mejor cuanto más se acerque a la unidad.

$$exactitud = \frac{VP + VN}{P + N} \quad \textbf{(9)}$$

Alcance: Mide la probabilidad de que si un individuo pertenece a cierta categoría, el sistema lo asigne a la categoría. Hay uno para cada clase:

$$rVP = \frac{VP}{P} = sensibilidad \quad \textbf{(10)}$$

$$rVN = \frac{VN}{N} = especificidad \quad \textbf{(11)}$$

El Análisis ROC Area es una metodología desarrollada para analizar un sistema de decisión.

[15] acrónimo del inglés de Attribute-Relation File Format.

Trabaja con las nociones de Sensibilidad y Especificidad. Cuando se utiliza una prueba dicotómica (una cuyos resultados se puedan interpretar directamente como positivos o negativos). La sensibilidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como positivo respecto a la condición que estudia la prueba, razón por la que también es denominada también Razón de Verdaderos Positivos (*rVP*) o también razón de éxitos.

La especificidad o Razón de Verdaderos Negativos (*rVN*), es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea el definido como negativo. Es igual al resultado de restar a uno la fracción de Falsos positivos (FP).

Se usa el área bajo esta curva ROC o AUC[16], como un indicador de la calidad del clasificador. En tanto dicha área esté más cercana a 1, el comportamiento del clasificador está más cercano al clasificador perfecto (aquel que lograría 100% de VP con un 0% de FP). Dado que VP es equivalente a sensibilidad y FP es igual a 1-especificidad, el gráfico ROC también es conocido como la representación de sensibilidad frente a (1-especificidad). Cada resultado de predicción o instancia de la matriz de confusión representa un punto en el espacio ROC [15].

Por último, se mide también el tiempo que tarde el clasificador en construir el modelo, el mismo se expresa en segundos.

Conclusiones

No existe un modelo clasificador mejor que otro de manera general. Es por esto que han surgido varias medidas para evaluar la clasificación y comparar los modelos empleados para un problema determinado. A continuación se enuncian algunos de los resultados arrojados:

- El empleo de ambas técnicas de clasificación en la realización de pruebas a estudiantes permite obtener una clasificación con altos porcentajes de Verdaderos positivos.
- A través de esta comparación, tanto en el modo testeo como en el de evaluación, fueron obtenidos los parámetros propios de funcionamiento (tipo kernel y cantidad de neuronas ocultas) con los cuales los algoritmos muestran los mejores resultados en la clasificación de este tipo de problemática. El algoritmo MVS mostró ser, en ambas comparaciones, el de mayor exactitud y el más rápido, cuando hace uso del kernel Polinomial.
- Respecto a la Sensibilidad, las RNA's resultaron tener mejor porcentaje que la MVS, si bien la diferencia es muy pequeña.
- En la Especificidad, la medición de la MVS resultó tener mejor probabilidad que las RNA's.
- Por último, en el análisis ROC, las RNA resultaron tener mejores valores que las MVS.

La herramienta de clasificación desarrollada como resultado de la presente investigación, si bien no es una herramienta de diagnóstico, sirve como ayuda

4. Referencias

[1] The 8th International Conference on Educational Data Mining Edm 2015. [En Línea] http://www.educationaldatamining .org/EDM2015/

[2] Frawley, W., Piatetski-Shapiro, G.; Matheus, C.J., *"Knowledge Discovery in Databases: an Overview"*. AI Magazine, 1992.

[3] Spositto, Osvaldo, et al. "Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil." *Novena Conferencia Iberoamericana en Sistemas, Cibernética e Informática, CISCI*. Vol. 29. 2010.

[4] Timarán R, Pereira J., *Detección de Patrones de Deserción Estudiantil en Programas de Pregrado de Instituciones de Educación Superior con CRISP-DM*. Congreso Iberoamericano de Ciencia, Tecnología, Innovación y Educación. ISBN: 978-84-7666-210-6. 2014

[5] Formia, S., *Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitario*. Universidad Nacional de La Plata, Tesis, Facultad de Informática, 2012.

[6] Pautsch J., *Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación*. Universidad Nacional de Misiones, Tesis de grado, 2009.

[7] Fayyad, U.M., Piatetsky-Shapiro and P. Smyth, 1996. "*From Data Mining to Knowledge Discovery: An Overview. … in knowledge discovery and data mining table of contents.* pp: 1-34. Press 1996.

[8] Inmon, W.H., *Building the Data Warehouse*. 3rd Edition, John Wiley, Chichester, 2002.

[9] Ryckeboer H., *"Utilización de técnicas de Data Warehouse para la toma de decisiones en el Área Académica"*. Anuario Resúmenes Extendidos. UNLaM. 2010, ISBN:978-987-1635-55-9

[10] Ryckeboer H., "*Implementación de un Data Warehouse para la toma de decisiones en el Área Académica*". Anuario Resúmenes Extendidos. UNLaM. 2013 ISBN:978-987-3806-30-8

[11] Hernández Orallo, J., *Introducción a la minería de datos.* Pearson Education, Valencia. 2004. *Editorial Pearson, 2004. ISBN: 84-205-4091-9.*

[16] por sus siglas en ingles Area Under the Curve.

[12] Montaño Moreno J. J., *Redes Neuronales Artificiales aplicadas al Análisis de dato*. U. Les Illes Balears. Tesis Doctoral Palma De Mallorca, 2002. [En línea] http://www.tdx.cat/bitstream/handle /10803/9441/ tjjmm1de1.pdf?sequence=1
[13] Bonifacio Martin Del Brio; Alfredo Sanz Molina; Redes *Neuronales y Sistemas Borrosos* (3ª ed.). Ed.Ra-Ma, 2007 ISBN 978-970-15-1250-0
[14] Witten, H., and Eibe F., *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.
[15] Fawcett, T., *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers.* HP Laboratories. 2004. [En línea] http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf
[16] Tinto, V., *Definir la Deserción: Una Cuestión de Perspectiva.* Revista de Educación Superior. 71 (1989): 33-51.
[17] Kimball, R., *The Data Warehouse lifecycle toolkit: expert methods for designing, developing, and deploying Data Warehouses.* John Wiley & Sons, New York.1998.
[18] del Alcázar León D., *Sistema inteligente para perfilar la deserción en estudiantes universitarios de carreras* técnicas. [En línea] http://contratosocialecuador.org/images/publicaciones/cuadermos/10.pdf
[19] Fischer Angulo E.S., *Modelo para la automatización del proceso de determinación de riesgo de deserción en estudiantes universitarios.* Santiago De Chile 2012. U. de Chile. [En línea] http://repositorio.uchile.cl/bitstream/handle/2250/111188/cf-fischer_ea.pdf?sequence=1
[20] Niño Sandoval T,, *Uso de Técnicas de Aprendizaje Automatizado para Predicción de Morfología Mandibular en Clase I, II y III Esquelética*. Tesis Doctoral. Universidad Nacional de Colombia. 2012

[21] Debandi N., *Reconocimiento y clasificación de hongos dermatofitos usando Máquinas de Soporte Vectorial (SVM).* UBA. Facultad de Ciencias Exactas y Naturales. Departamento de computación. [En línea] https://www.dc.uba.ar/inv/tesis/licenciatura/2007/debandi.pdf
[22] Pérez Ortiz J., *Clasificación con discriminantes: un enfoque neuronal. Departamento de* Lenguajes y Sistemas Informáticos Universidad de Alicante. Julio 1999. [En línea] ttp://www.dlsi.ua.es/~japerez/pub/pdf/cden1999.pdf
[23] Hilera González, J.R., *Redes neuronales artificiales: fundamentos, modelos y aplicaciones.* Editor RA-MA, 1994.
[24] Carreras, X., Márquez, L. & Romero, *E. Máquinas de Vectores Soporte, en Hernández, J., Ramírez, M. y Ferri, C., Introducción a la Minería de Dato.* Editorial Pearson, España, 2004.
[25] J. Moreno Gutiérrez, L. Melo Velandia. *Pronóstico De Incumplimientos De Pago Mediante Máquinas De Vectores De Soporte: Una Aproximación inicial a la gestión del riesgo de crédito*. Banco de la República. Colombia.