Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales

Prediction of academics marks of informatics students using neural networks

Jorge Álvarez Blanco¹ Rogelio Lau Fernández^{1*}
Sonia Pérez Lovelle¹ Exiquio C. Leyva Pérez¹

Recibido 3 de julio de 2015, aceptado 11 de enero de 2016
Received: July 3, 2015 Accepted: January 11, 2016

RESUMEN

En este artículo se muestra la aplicación de redes neuronales con el objetivo de predecir los resultados docentes que alcanzarán los estudiantes en las asignaturas Estructuras de Datos I y II, que forman parte del currículo de la carrera de Ingeniería Informática que se imparte en el Instituto Superior Politécnico José Antonio Echeverría (ISPJAE) en la República de Cuba. El presente trabajo fue motivado fundamentalmente por el hecho de que las asignaturas escogidas tienen un significativo grado de complejidad, lo que exige una importante dedicación y rigurosidad por parte de los estudiantes. Como consecuencia los resultados docentes que se obtienen actualmente en esas materias no son buenos. Con el objetivo indicado se desarrolló una aplicación, basada en el sistema MATLAB, que permite predecir resultados docentes en las materias de interés, partiendo de resultados en algunas asignaturas del año de estudio anterior y otros datos de interés. Se emplearon dos redes neuronales, ambas con la misma arquitectura, pero cada una entrenada con los datos específicos de cada asignatura (Estructuras de Datos I y II). Se realizó un grupo de experimentos para contrastar el comportamiento de las redes respecto a algunas estadísticas presentes en los datos de la muestra. En general se logró una efectividad en la predicción superior a un 78% para el caso de la primera asignatura, mientras que para la segunda se alcanzó una efectividad superior al 75%.

Palabras clave: Ingeniería informática, resultados docentes, rendimiento escolar, predicción, redes neuronales.

ABSTRACT

In this paper is shown the application of neural networks in order to predict academic marks that will be obtained for the students in the subjects of Data Structures I and II, included both in the Informatics Engineering curricula at Higher Polytechnic Institute José Antonio Echeverría in the Republic of Cuba. The main motivation for the present work is justified because selected subjects have a high level of complexity, demanding from the student to be rigorous and a permanent dedication. As a consequence the academic results obtained at the present time are not satisfactory. To reach the goal mentioned above a software based on MATLAB tool was developed and the marks obtained previously by students in some subjects and others data of interest were used. Two neural networks were employed, both with the same architecture, but each one trained with the specific data of each subject (Data Structures I and II). A group of experiments was carried out to contrast the behavior of the neural networks regarding some specific statistics in the data of the sample. An overall effectiveness in prediction superior to 78% for the case of the first subject was achieved, while for the second one effectiveness superior to 75% was reached.

Keywords: Informatics engineering, academic results, academic performance, prediction, neural networks.

¹ Facultad de Ingeniería Informática. Instituto Superior Politécnico José Antonio Echeverría. Ave 114 N° 11901, Marianao. La Habana, Cuba. E-mail: ralvarezg@yahoo.com; lau@ceis.cujae.edu.cu; sperezl@ceis.cujae.edu.cu; exiquio@ceis.cujae.edu.cu

* Autor de correspondencia

INTRODUCCIÓN

El pronóstico es inherente a la vida humana y acompaña al hombre desde tiempos remotos, tal vez lo más nombrado en este sentido, día a día, sea el parte referente al estado del tiempo y el comportamiento de las variables atmosféricas, esto por su incidencia determinante en el aseguramiento de la alimentación, seguridad y numerosas facetas de la vida misma. No obstante, las técnicas que permiten la realización de pronósticos se utilizan en un sinnúmero de distintas actividades del quehacer y el conocimiento [1-3]. Existe un alto número de técnicas y herramientas susceptibles de ser aplicadas, algunas de carácter específico y otras con un valor general [4-6].

En una institución educacional resulta de gran importancia el predecir los resultados docentes que podrían ser obtenidos por los estudiantes en las diferentes asignaturas y cursos donde se matriculan [7]. El disponer de esta información permite, por un lado, tomar determinadas decisiones organizativas ante la presencia de colectivos que se prevean puedan presentar dificultades por otro lado, brindar a los docentes, mediante informaciones más completas y oportunas sobre sus estudiantes, la capacidad para desarrollar un tratamiento diferenciado de sus educandos, lo que lleva a la corrección en tiempo de las limitaciones en la asimilación de cada uno de ellos, con vistas al logro de la excelencia académica [8]. Lo antes indicado se ve hoy favorecido por la existencia de los sistemas de información y almacenes de datos universitarios que contienen la información requerida para la aplicación de las técnicas correspondientes.

Las redes neuronales [9], que han demostrado constituir una herramienta muy efectiva en la predicción en general [10-11], es uno de los varios enfoques más utilizados para el caso referente al rendimiento estudiantil [12-14]. Generalmente, estas técnicas se han aplicado para predecir simplemente los estados de “éxito” o “fracaso”, pero normalmente las calificaciones de las materias se emiten en una escala ordinal, y varían desde las más bajas categorías que equivaldrían a “fracaso”, hasta la más alta de “excelente”.

En este trabajo se aplican las redes neuronales con el objetivo de predecir los resultados docentes

que alcanzarán los estudiantes en las asignaturas Estructuras de Datos I y II, que forman parte del currículo o plan de estudio de la carrera de Ingeniería Informática que se imparte en el Instituto Superior Politécnico José Antonio Echeverría (ISPJAE) en la República de Cuba. Estas asignaturas, que resultan esenciales en la formación del especialista antes indicado, se ubican en el segundo año de la malla curricular de la carrera mencionada, precedidas por varias materias de formación general, pero también de otras con las que tienen una estrecha vinculación y que les resultan precedentes, estas son Introducción a la Programación, y Diseño y Programación Orientada a Objetos. Como se ha de suponer, los resultados docentes obtenidos por los estudiantes en estas asignaturas resultan datos esenciales en el pronóstico, considerándose también otros que serán indicados más adelante.

La motivación fundamental para el desarrollo del presente trabajo viene dado por el hecho de que las asignaturas escogidas tienen un significativo grado de complejidad, requiriendo una importante dedicación y rigurosidad por parte de los estudiantes y por ello los resultados docentes que se obtienen en la actualidad no son buenos, estando, junto a la Física y la Matemática, con los índices de promoción más bajos en los semestres en que se imparten. El colectivo de docentes de las asignaturas desea poder identificar con antelación los alumnos que pudieran tener dificultades, lo que facilitaría el desarrollo, con carácter anticipado, de acciones educativas diferenciadas.

La información acerca de los estudiantes que se requiere para el procesamiento, sobre todo las evaluaciones obtenidas por los alumnos con anterioridad, son tomadas de las bases de datos del sistema informativo de control docente existente en el centro de educación superior antes referido, lográndose de esta forma un uso más amplio de los datos que sobre los alumnos se acumulan. Aun cuando este trabajo perseguía el objetivo de disponer de un software que se pudiera aplicar en forma reiterativa en las dos asignaturas, no se pretendía construir la solución desde cero, sino utilizar lo ya disponible que fuera factible, por ello se hizo uso del MATLAB.

El presente artículo se inicia haciendo una breve descripción sobre el modelo de predicción utilizado y

la herramienta de software utilizada en el desarrollo del trabajo. A continuación se describen los elementos constituyentes del diseño de la red neuronal escogida para ser implantada, una vez que fueron realizadas pruebas con varios modelos. Después se explican los aspectos referentes al entrenamiento de las redes, el comportamiento de las pruebas y los resultados obtenidos en comparaciones con respecto a estadísticas presentes en la muestra. Por último, se explican las ideas sobre la generalización del trabajo ejecutado con vistas a su posible utilización en el conjunto de asignaturas de la carrera universitaria mencionada, por lo menos en las que pertenecen a sus disciplinas de la profesión.

MODELO DE PRONÓSTICO SELECCIONADO Y HERRAMIENTA UTILIZADA

En problemas de predicción, la situación típica que se presenta es la de estimar el comportamiento de una variable (salida) en función de una o más variables de entrada o de predicción. Tradicionalmente, la solución de estos problemas se ha abordado desde la óptica de modelos estadísticos de regresión. Sin embargo, para que estos sean aplicables, los datos deben cumplir algunas condiciones, entre ellas: independencia, linealidad y tamaño suficiente de la muestra, entre otros. Una dificultad surge cuando las premisas de los métodos estadísticos no se cumplen, o no se pueden demostrar. En estos casos, hacen falta métodos alternativos que permitan relajar los supuestos que necesitan cumplir los datos de predicción, al tiempo que puedan mantener una buena calidad en estimar un valor. En este escenario surgen técnicas del área de la Inteligencia Artificial, que incluyen los Sistemas de Inferencia Difusa y Algoritmos Evolutivos, pero con mayor protagonismo las Redes Neuronales Artificiales (RNA), como alternativa para resolver problemas de regresión en las condiciones indicadas [15].

Ha existido un interés en la literatura en analizar las relaciones entre los modelos estadísticos y las RNA en tareas de agrupamiento/clasificación. En cuanto a estas últimas, la atención ha estado centrada en las redes del tipo Perceptrón Multicapa (MLP) y de Función Base Radial (FBR), pero se han reportado resultados no concluyentes cuando se ha tratado de determinar cuáles de los modelos anteriores son más eficientes [16].

Se reconoce, de un lado, que las RNA proporcionan un modelo donde es difícil la interpretación de los parámetros y, de otro lado, se necesita de un cierto empirismo en su proceso de construcción puesto que no hay una fórmula clara y unívoca en cuanto a la arquitectura general para su diseño. Otro ingrediente de empirismo es requerido también para determinar la cantidad de datos para el entrenamiento y la validación.

No obstante, en este trabajo se decidió utilizar como modelo de predicción una red neuronal atendiendo a la habilidad para aprender y aproximar relaciones entre las variables de entrada y salida, independiente del tamaño y la complejidad del problema, y sobre todo, a que una de sus virtudes es el buen rendimiento ante problemas no lineales, así como su capacidad de admitir interacciones entre las variables de predicción, algo que es necesario incluir desde un inicio en algunas técnicas estadísticas [17].

Las RNA de Función de Base Radial (FBR) generalmente requieren de una mayor cantidad de neuronas en los nodos ocultos que en las redes MLP y comparativamente tienen un mejor desempeño solo cuando el volumen de datos de entrenamiento es mayor. Por ello, a pesar de que las FBR se han aplicado a una variedad de problemas, el tipo de red utilizada en el trabajo fue la MLP. Este tipo de red se encuentra entre las arquitecturas de red más poderosas y populares y, pueden aproximar hasta el nivel deseado cualquier función continua en un intervalo. En el tipo de problema que aquí se aborda se ha obtenido, en algunos casos, que los ajustes ofrecidos por las redes MLP son superiores a los obtenidos por las redes FBR, aplicados a los mismos conjuntos de datos [18-19].

Como fuera antes indicado, la herramienta que se utilizó para desarrollar la aplicación usada en este trabajo fue el MATLAB [20], aun cuando existían otras susceptibles de ser consideradas con tal objetivo, como Clementine [21] o KNIME (*Konstanz Information Miner*) [22]. La primera, creada por *Statistical Package for the Social Sciences* (SPSS), está enfocada hacia el desarrollo de modelos de predicción para la toma de decisiones, y utiliza diversas técnicas, entre las que están múltiples variantes de redes neuronales. Por otro lado, la segunda, desarrollada en la Universidad de Constanza, Alemania, es una plataforma de minería de datos

que permite el desarrollo de modelos en un entorno visual. La decisión de utilizar MATLAB en este trabajo tiene los siguientes fundamentos:

- Dispone de un *Toolbox* especializado en redes neuronales, totalmente integrado al resto de la herramienta, con diversidad de tipos de modelos y con la posibilidad del uso de cualquiera otra funcionalidad presente en ella [23].
- Utiliza un lenguaje de programación sencillo de aprender, aun cuando se esté trabajando con estructuras complicadas que realicen cálculos matemáticos de alta complejidad, en este caso las redes neuronales.
- Posee una vasta ayuda, contenida dentro del propio producto MATLAB, que contiene todas las características de la herramienta, existiendo una buena integración entre los temas.
- Permite diseñar aplicaciones visuales de forma muy sencilla y rápida debido a que contiene un GUI (*Graphic User Interface*) con componentes visuales de alta aplicación.
- Facilita la salva, en archivos con extensión .m, de los datos de interés, así como funciones que hayan sido instrumentadas y que podrán ser usadas cuando se requieran. Es decir, en el segundo caso estos ficheros se usan como bibliotecas.

DISEÑO DE LA RED NEURONAL

La siguiente decisión debía ser tomada en el sentido de usar una red con dos salidas, una para cada asignatura a ser pronosticada, o dos redes independientes, una para cada una de ellas. La decisión final fue utilizar dos redes neuronales independientes, aunque iguales, una con el objetivo de predecir resultados en Estructuras de Datos I (EDI), y la otra, para hacerlo en Estructuras de Datos II (EDII). Para arribar a esta conclusión se llevaron a cabo numerosas pruebas. Las realizadas a una única red neuronal no tuvieron resultados de predicción tan efectivos como los logrados con dos independientes. De esta forma cada una es entrenada para realizar predicciones correctas de un solo resultado.

Lo señalado se puede considerar lógico si se tiene en cuenta que al disponer de una sola red para pronosticar dos resultados que son relativamente independientes entre sí, el entrenamiento forzará a una dependencia que solo es posible sacrificando precisión en ambos pronósticos. Con dos redes

neuronales independientes, cada una es entrenada para realizar predicciones correctas de un solo resultado.

A continuación se discute la topología o arquitectura que finalmente presentan las dos redes neuronales utilizadas del tipo MLP. Esta arquitectura de RNA es, dentro de las redes de aprendizaje supervisado unidireccionales, muy popular gracias a su capacidad de simular prácticamente cualquier comportamiento no lineal. Una adecuada selección de esta es fundamental a la hora de obtener resultados adecuados en la predicción.

La arquitectura típica de este tipo de red está constituida por varias capas de nodos con interconexión completa entre ellos. Se realizaron diferentes pruebas con más de una capa intermedia, pero los resultados no fueron satisfactorios, oscilando la efectividad (expresado en el por ciento de aciertos) entre un 55% y un 63%, para ambas redes, valores inferiores a los logrados finalmente, como se mostrará más adelante. Por todo ello, se tendrá:

- Una capa de entrada.
- Una capa intermedia u oculta.
- Una capa de salida.

Las funciones de activación utilizadas en las neuronas de cada capa fueron las siguientes: para las capas de entrada y la oculta, la sigmoideal tangente hiperbólica; para la capa de salida, la lineal, ya que se deseaba la mayor amplitud en el rango del intervalo de salida y facilitar la interpretación.

El uso de las funciones de activación de tipo sigmoideal, es muy común en el trabajo con las redes neuronales de alimentación hacia delante que utilizan el algoritmo de retropropagación (*backpropagation*) para el entrenamiento, como es en el presente caso.

Para la determinación del número de neuronas en la capa de entrada se puede seguir el criterio de establecer una neurona por cada variable de entrada o un número menor, lo que implica que en varias neuronas existirá más de una entrada. En el presente caso se realizaron numerosas pruebas, comprobándose un mejor rendimiento de la red neuronal cuando existía una neurona por cada variable de entrada que cuando se dejó un número más reducido, como se muestra en la Tabla 1.

Tabla 1. Comparación del error mínimo cuadrado con diferentes números de neuronas.

	Entrenamiento con una neurona por cada variable de entrada	Entrenamiento con menos neuronas
Error mínimo cuadrado	Por debajo de 0,03	Entre 0,2 y 0,65

Por lo antes indicado, se decidió tener 21 neuronas en la capa de entrada, que es exactamente el número de variables o factores de entrada que se había decidido considerar en el estudio a realizar, aunque existen otros diversos factores endógenos y exógenos que influyen en el rendimiento de un estudiante [24]. Con esta decisión, se coincide con otros trabajos realizados anteriormente, entre los cuales se encuentran los de Collantes [16] y Oladokun [14]. Las 21 variables o factores se describen a continuación agrupados en varios conjuntos:

a) Notas en materias de primer año (2): Introducción a la Programación (IP) y Diseño y Programación Orientada a Objetos (DPOO). Se consideraron los resultados de estas asignaturas, debido a las siguientes razones:

- Ellas brindan a los estudiantes conocimientos y habilidades que le son indispensables, como prerrequisitos, al momento de cursar las asignaturas de Estructuras de Datos.
- Ellas le exigen al estudiante un amplio uso del pensamiento lógico, heurístico y algorítmico, lo que es una habilidad que, además de ayudar al estudiante durante toda la carrera, es un pilar fundamental para enfrentar la complejidad de las materias de Estructuras de Datos.
- Los estudiantes que obtienen buenas notas en estas asignaturas normalmente muestran tener las capacidades para enfrentar con éxito las materias EDI y EDII. No ocurre lo mismo con aquellos que no obtienen buenas notas en esas materias.

El rango de valores de estas dos variables de entrada es 2 (desaprobado), 3 (aprobado), 4 (bien) o 5 (excelente), correspondiendo con las posibles calificaciones establecidas en la educación universitaria de la República de Cuba.

b) Notas en materias de matemática (3): Matemática I (Mat I), Matemática II (Mat II) y Álgebra Lineal

(AL). También se cursan en el primer año. Se consideraron los resultados de estas asignaturas debido a lo siguiente:

- En esas materias se desarrollan los fundamentos de la formación de un ingeniero informático, al tiempo que lo capacita para modelar y analizar representaciones de procesos y objetos en diferentes escenarios, por lo que contribuyen al desarrollo del pensamiento lógico, heurístico y algorítmico. También se reconoce en varias fuentes, que estudian los factores que influyen en el rendimiento académico en carreras de ingeniería, el impacto de los resultados obtenidos por los estudiantes en las asignaturas de matemáticas [14, 25].

El rango de valores de estas tres variables de entrada es igual al establecido en el punto anterior.

c) Cantidad de convocatorias requeridas para aprobar cada una de las asignaturas de IP, DPOO, Mat I, Mat II y AL (5). Estos valores se toman como variable de entrada porque brindan información sobre el posible éxito posterior de este alumno en aprobar las asignaturas. Una tendencia a aprobar en primera convocatoria puede sugerir que el estudiante tiene aplicación y capacidades para enfrentar las asignaturas. Si necesita múltiples convocatorias para aprobar, puede significar falta de dedicación o que no tiene suficientes capacidades, o ambas, lo que podría repetirse, en particular, en las asignaturas de Estructuras de Datos. El rango de valores posibles, para estas cinco variables, se muestra en la Tabla 2.

d) Cantidad de veces que el estudiante ha cursado y suspendido las asignaturas de IP, DPOO, Mat I, Mat II y AL (5). Significa que suspendió todas las convocatorias y se vio en la obligación de cursarla de nuevo al siguiente curso. Estos factores se tomaron en cuenta por las mismas causas que en el conjunto anterior.

Existe diferencia entre el estudiante que aprueba una asignatura la primera vez que la cursa respecto

Tabla 2. Posibles valores de las variables referidas a las convocatorias.

Valor	Significado
1	El estudiante alcanzó el aprobado en la primera convocatoria.
2	El estudiante alcanzó el aprobado en la segunda convocatoria.
3	El estudiante alcanzó el aprobado en la tercera convocatoria.
4	El estudiante no logró aprobar la asignatura antes de cursar las Estructuras de Datos.

a otro que la suspendió una vez, o incluso dos veces, y termina pasándola, incluso, con la misma nota que el primero. El rango de valores de entrada, para estas cinco variables, se muestra en la Tabla 3.

Tabla 3. Posibles valores de la variable cantidad de veces.

Valor	Significado
0	El estudiante nunca ha suspendido la asignatura.
1	El estudiante suspendió en una ocasión la asignatura.
2	El estudiante suspendió en dos ocasiones la asignatura.

El valor máximo de esta variable de entrada es dos porque por reglamento un alumno no puede cursar más de tres veces una asignatura.

e) Género (1). Esta característica fue considerada como variable de entrada a las redes neuronales debido a que en el ámbito escolar se manifiestan diferencias entre el género masculino y el femenino, entre las que se encuentran: la inclinación hacia la teoría o la práctica, el gusto por las letras o las ciencias técnicas, la dedicación al estudio y la perseverancia ante el reto de nuevos aprendizajes. Este factor sociodemográfico se emplea casi sin excepción en todos los trabajos consultados que tratan de predecir resultados académicos [12-14, 19, 24] a nivel universitario. El rango de valores de entrada se muestra en la Tabla 4:

Tabla 4. Posibles valores de la variable género.

Valor	Significado
0	El estudiante pertenece al sexo masculino.
1	El estudiante pertenece al sexo femenino.

f) Origen académico y fuente de ingreso (2). Estas dos informaciones fueron consideradas como variables de entrada a las redes neuronales debido a que existen diferencias entre los orígenes

y fuentes de ingreso (la forma en que obtienen el ingreso a la universidad) de los estudiantes y esto tiene influencia en la preparación previa, motivación y otros factores de los que dependen los resultados académicos. Se debe aclarar, de manera sintética, que en Cuba la cantidad de plazas por cada carrera universitaria está limitada a cifras que se corresponden con las necesidades del país y, por tanto, con las posibilidades de un empleo ulterior. En la mayoría de los casos el acceso a cada carrera particular se logra cuando el alumno obtiene una cierta puntuación que le permite clasificar dentro de la cuota asignada a una de las diez carreras que él solicitó previamente en orden de prioridad. La puntuación anterior es el promedio entre las calificaciones obtenidas en todas las materias del nivel preuniversitario y la media de las calificaciones de los exámenes de ingreso a la universidad. Los estudiantes que provienen de un instituto preuniversitario son mayoría y se ajustan al procedimiento anterior, pero hay otras cinco fuentes de ingreso. Tres de ellas son de diferentes instituciones armadas que, por acuerdos ministeriales, pueden otorgar derechos de ingreso a la universidad a jóvenes seleccionados que han servido en sus filas. También de forma diferenciada se otorgan plazas a estudiantes con buenos resultados que provienen de institutos politécnicos. Una última fuente de ingreso, para un número más reducido de plazas de una carrera particular, se genera de exámenes de concurso, que pueden realizar todos aquellos que no han podido obtener plaza en la carrera que desean, por alguna causa. El indicador origen académico tiene relación con la fuente de ingreso, pero se trata de distinguir si el alumno, independiente de su fuente de ingreso, tuvo una formación tradicional de preuniversitario o si su formación se produjo en un instituto de enseñanza técnica y profesional. El rango considera valores enteros, entre 0 y 1 para la variable origen académico y entre 0 y 5 para la variable fuente de ingreso.

g) Nivel académico de los padres (2). Esta información fue considerada como variable de entrada a las redes neuronales debido a:

- La posible influencia que sobre el estudiante puede tener la experiencia académica de los padres al momento de orientar a sus hijos y la ayuda que estos puedan recibir en lo motivacional y afectivo.

El rango de valores de entrada, para estos dos parámetros, se muestra en la Tabla 5.

Tabla 5. Posibles valores de las variables nivel académico del padre y de la madre.

Valor	Significado
0	El progenitor tiene enseñanza primaria o media.
1	El progenitor tiene enseñanza desconocida.
2	El progenitor tiene enseñanza media superior.
3	El progenitor tiene enseñanza superior.

h) Zona de residencia (1). Con esta variable se trata de reflejar la posible influencia que tiene el lugar donde se vive (se tomó el municipio) con los resultados académicos en la universidad. Se trata de reflejar aquí los costos del traslado en términos de tiempo, de desgaste físico y dinero en alimentación; los que pueden reducir la capacidad de aprendizaje en la medida en que el individuo necesite utilizar varias y diversas conexiones para llegar al establecimiento educativo y regresar.

El rango de valores de entrada se muestra en la Tabla 6.

Tabla 6. Posibles valores de la variable zona de residencia.

Valor	Significado
0	El estudiante reside en un municipio de difícil acceso a la universidad.
1	El estudiante reside en un municipio de fácil acceso a la universidad.

Cada RNA tiene en la capa de salida una neurona que emitirá valores numéricos reales que la aplicación transformará a valores cualitativos más fáciles de interpretar por el usuario. La aplicación muestra dos tipos de resultados para cada red, uno general que indica si el alumno suspende o aprueba, y otro específico donde se indica la nota aproximada

del estudiante, como muestran las Tablas 7 y 8, respectivamente. La variable *respuesta* que aparece en dichas tablas representa el valor de salida brindado por la red.

Tabla 7. Resultados generales brindados al usuario.

Resultados obtenidos por las redes	Resultados brindados al usuario
respuesta < 1	suspenso
1 ≤ respuesta < 3	entre 2 y 3 puntos (frontera)
respuesta ≥ 3	aprobado

Tabla 8. Resultados específicos brindados al usuario.

Resultado obtenidos por las redes	Resultados brindados al usuario
respuesta < 1	2 puntos
1 ≤ respuesta < 3	entre 2 y 3 puntos
3 ≤ respuesta < 4	entre 3 y 4 puntos
4 ≤ respuesta < 6	entre 4 y 5 puntos
respuesta ≥ 6	5 puntos

En cuanto a la determinación del número de nodos de la capa oculta se ha seguido la regla que sugiere tomar la cantidad promedio entre las neuronas de la capa inicial y de la final [17]. Luego el valor fue 11,5. Este resultado se redondeó por defecto a 11 neuronas, con lo que se obtuvieron valores de predicción aceptables y mejores que los que se lograron con una cantidad mayor o menor de nodos en esa capa.

ENTRENAMIENTO DE LAS REDES NEURONALES

El algoritmo básico para el entrenamiento de redes MLP es conocido como retropropagación (*backpropagation*), del cual hay diferentes variantes que apuntan a lograr rapidez o a lograr una mejor utilización de la memoria. El sistema MATLAB también ofrece opciones para seleccionar variantes del algoritmo básico. Después de varios ensayos se lograron los mejores resultados en las predicciones de la asignatura EDI con el algoritmo *Resilient Backpropagation*, además de mostrar una convergencia bastante rápida. Para la predicción de calificaciones en la asignatura EDII se adoptó el algoritmo del gradiente conjugado en la variante Powell-Beale.

Algunas especificaciones para el funcionamiento del algoritmo fueron las siguientes: el error esperado para ambas redes se estableció en 0,01. La tasa de

aprendizaje para predecir las notas en la asignatura EDI fue 1,05, mientras en la otra red fue de 0,7. En la primera red el gradiente mínimo fue de 10^{-5} , mientras en la segunda red fue de 10^{-20} . La cantidad máxima de iteraciones se fijó en 5.000 para ambas redes.

El paso posterior fue definir los conjuntos de valores de entrada que serían utilizados, tanto para el entrenamiento como para las pruebas. Los estudiantes seleccionados fueron aquellos que cursaron estas asignaturas a partir del curso 2010-2011, puesto que anteriormente los contenidos de ambas estaban unidos en una sola materia. En total la muestra que se alcanzó fue de 350.

La muestra anterior se dividió en dos conjuntos: conjunto de entrenamiento y conjunto de prueba. Se determinó utilizar alrededor de un 80% de los datos totales para el primero y el resto para la prueba, siendo este el criterio que usualmente se aplica. Las cantidades finales fueron:

- Cantidad de estudiantes utilizados en el entrenamiento: 290.
- Cantidad de estudiantes utilizados en la prueba: 60.

Esto significa que los estudiantes utilizados en la fase de entrenamiento representan un 82,8% del total, mientras los correspondientes a la fase de prueba representan el 17,2% restante, considerándose las diferencias cualitativas que tienen que existir entre los datos en cada uno de los conjuntos. Así, para cada conjunto, en el presente trabajo se tuvo en cuenta:

- Necesidad de que las RNA fueran entrenadas con un conjunto de estudiantes que abarcaran todas las notas posibles de las asignaturas EDI y EDII (2, 3, 4 o 5 puntos).
- Necesidad de que las RNA fueran entrenadas con un conjunto de datos de estudiantes que presentaran valores de los parámetros de entrada representativos del universo de los posibles.
- Necesidad de que el conjunto de prueba contuviera estudiantes que abarcaran todas las notas posibles de EDI y EDII. Así, se podría comprobar si la red es capaz de realizar predicciones precisas con cada una de ellas.
- Necesidad de que el conjunto de prueba contuviera estudiantes que poseyeran valores de los parámetros de entrada representativos del

universo de los posibles. Esta razón es similar a la segunda expuesta antes, solo que en este caso permitiría comprobar si las implicaciones que le asignaron las redes a cada uno de los parámetros fueron las correctas.

El entrenamiento de las redes neuronales brindó los siguientes resultados: mínimo error cuadrado para la red de Estructuras de Datos I = 0,0077853, y mínimo error cuadrado para la red de Estructuras de Datos II = 0,0063383.

PRUEBA Y COMPROBACIONES ESTADÍSTICAS DE LAS REDES NEURONALES

Se realizaron numerosos entrenamientos y pruebas para buscar la mayor efectividad posible, incluso se modifica varias veces la composición de los conjuntos de entrenamiento y prueba.

Debido a que los valores de las predicciones se brindan en dos modos diferentes, específicos y generales, cada red neuronal tiene dos porcentajes de efectividad. El mejor valor se logra en los pronósticos generales. En la red neuronal para predecir las notas en la asignatura EDI, la efectividad de resultados específicos fue de 78,3%, mientras que la de resultados generales ascendió a 86,7%. En la red para predecir las notas en la materia EDII, la efectividad de resultados específicos alcanzó 75%, mientras que en la de resultados generales llegó a 78,3%. Estos resultados son muy satisfactorios dado que un poder de predicción de una red entre un 68% y un 72% ya la hace efectiva.

Resultó muy interesante realizar un grupo de experimentos para contrastar el comportamiento de estas con respecto a algunas estadísticas de interés presentes en los datos de la muestra. De hecho, si las predicciones de las redes neuronales se comportan de acuerdo a las estadísticas existentes, entonces se reafirma que ellas se encuentran correctamente entrenadas. Aquí no se mostrarán todos los casos analizados, sino solamente algunos interesantes.

a) Estudiantes que habían obtenido 4 y 5 puntos en las asignaturas IP y DPOO.

De los 350 estudiantes existentes en la muestra, 114 cumplen con la condición indicada. De ellos,

106 aprobaron la asignatura EDI para un 93% de promoción, mientras 107 lo hacían en EDII para un 93,8%. De paso, se puede observar en estos datos que si un estudiante obtiene en las dos asignaturas en cuestión 4 o 5 puntos, tiene una altísima probabilidad de aprobar finalmente EDI y EDII.

Para comprobar el posible ajuste de las redes a estas estadísticas se procedió al uso de ambas para simular estudiantes ficticios que obtenían entre 4 y 5 puntos en las asignaturas mencionadas, asignándole al resto de los 19 parámetros valores siempre diferentes. En el experimento se utilizaron 40 estudiantes ficticios en la forma que se muestra en la Tabla 9, donde también se recogen los pronósticos obtenidos.

Como se podrá notar, la RNA asociada a la asignatura EDI brindó un 87,5% que es un porcentaje muy cercano al que da la estadística, sobre todo si se tiene en cuenta que la red tiene una probabilidad de fallo de un 13,3%, para la predicción general, que fue la que se utilizó en este caso. Además cabe destacar que los cinco estudiantes que no fueron predichos como aprobados, se ubicaron en la frontera (entre 2 y 3 puntos), lo que es una previsión de que algunos de ellos aprobarán también.

En el caso de la RNA para predecir notas en la materia EDII, el resultado también se asemeja al

estadístico. Comparado con la red anterior, existe más diferencia entre la estadística y la predicción de la red, lo que se corresponde con la efectividad exhibida durante la prueba. Sucede también que los seis estudiantes que no fueron pronosticados de aprobados en la asignatura, clasificaron en la zona de frontera, por lo que también se puede asumir el pronóstico de que algunos podrían terminar aprobando.

b) Estudiantes que aprobaron en la primera convocatoria las asignaturas de IP, DPOO, Mat I, Mat II y AL.

Un total de 101 estudiantes cumplen con la condición establecida. De ellos, 94 aprobaron la asignatura EDI para un 93,1% de promoción, mientras 97 lo hacían en EDII para un 96%. De paso, aquí también se puede observar que si un estudiante aprueba todas las asignaturas en cuestión en la primera convocatoria, tiene una altísima probabilidad de aprobar finalmente EDI y EDII.

Para la comprobación del ajuste de las redes a esta estadística se utilizaron 40 estudiantes ficticios que obtenían todas sus notas en la primera convocatoria, para el resto de los parámetros se asignaron diferentes valores. Los resultados de esta comprobación se muestran en la Tabla 10.

Tabla 9. Comparación con la estadística de los estudiantes que obtienen 4 y 5 puntos en las asignaturas IP y DPOO.

Resultados en IP y DPOO	Total de estudiantes simulados	Pronóstico de aprobados en EDI	Pronóstico de aprobados en EDII
Con 5 en ambas asignaturas	10	10	10
Con 4 en ambas asignaturas	10	8	7
Con 5 en IP y 4 en DPOO	10	9	8
Con 4 en IP y 5 en DPOO	10	8	9
Totales	40	35	34
% de aprobados según pronóstico		87,5	85
% de aprobados según estadística		93	93,8

Tabla 10. Comparación con la estadística de los estudiantes que aprobaron en primera convocatoria las cinco asignaturas.

	Total de estudiantes simulados	Pronóstico EDI	Pronóstico EDII
Estudiantes que aprueban en la primera convocatoria	40	37	35
% de aprobados según pronóstico		92,5	87,5
% de aprobados según estadística		93,1	96

Como se puede apreciar en la Tabla 10, la red neuronal que predice las calificaciones de la asignatura EDI estuvo bien cerca del resultado contenido en la muestra, la otra red también brindó una gran aproximación, aunque no estuvo tan cerca de los resultados acorde con la estadística. Hasta aquí se comprueba que la efectividad de predicción de esta segunda red es menor, lo que deberá ser objeto de revisión posteriormente. Los estudiantes que no fueron pronosticados como aprobados por ambas redes fueron clasificados entre 2 y 3 (frontera).

c) Estudiantes con padres que tienen nivel académico superior.

En la muestra, un total de 89 estudiantes tienen ambos padres con nivel académico superior, de ellos, 76 aprobaron la asignatura EDI para un 85,3% de promoción, mientras 73 lo lograban en EDII para un 82%. Por otro lado, un total de 71 estudiantes tienen un padre con nivel académico superior, de ellos, 51 aprobaron la asignatura EDI para un 71,8% de promoción, mientras 59 lo lograban en EDII para un 83%. Por último, un total de 190 estudiantes no tienen ningún padre con nivel académico superior, de ellos, 117 aprobaron la asignatura EDI para un 61,6% de promoción, mientras 132 lo lograban en EDII para un 69,5%.

En esta estadística se observa que los estudiantes que poseen los dos padres con nivel académico superior tienden a ser más exitosos que los demás, luego le siguen los que tienen un solo padre con nivel superior, y por último se encuentran los que no poseen ninguno que tenga nivel superior. De lo anterior se concluye que tener padres que

constituyan guías intelectuales para el estudiante es algo ventajoso.

Se realizaron las pruebas con la misma estrategia explicada anteriormente, pero esta vez se crearon 40 estudiantes ficticios para cada una de las categorías. Se mantuvieron fijos en cada caso los parámetros de niveles de escolaridad del padre y la madre, y se varía el resto. Los resultados obtenidos de la simulación se muestran en la Tabla 11.

Como se observa, en general, el ajuste de ambas redes neuronales es correcto. Se puede ver el mismo patrón de que los que más aprueban son aquellos que ambos padres son de nivel académico superior, luego los que tienen uno solo y, por último aquellos que no poseen ninguno.

GENERALIZACIÓN Y TRABAJO FUTURO

El trabajo ejecutado y los resultados presentados tienen hasta el momento un carácter de experimento, realizado con el objetivo de valorar la conveniencia o no del uso de las redes neuronales como herramienta para el pronóstico. Por supuesto, para llevarlo a cabo se seleccionaron las materias o asignaturas que más requerían de una herramienta de pronóstico.

Dados los alentadores resultados que se han obtenido en el trabajo, resultaría provechoso continuar trabajando por su mejoramiento y en su generalización. Se debe probar si aún no es posible lograr una mayor eficacia mediante la realización de modificaciones en la arquitectura actual, e inclusive experimentar con otras posibles. También resulta un

Tabla 11. Cantidad de padres que tienen nivel académico superior.

	Total de estudiantes simulados	Pronóstico de aprobados en EDI	Pronóstico de aprobados en EDII
Ambos padres con nivel académico superior	40	34	34
% de aprobados según pronóstico		85	85
% de aprobados según estadística		85,3	82
Con un padre de nivel académico superior	40	29	32
% de aprobados según pronóstico		72,5	80
% de aprobados según estadística		71,8	83
Ningún padre con nivel académico superior	40	27	30
% de aprobados según pronóstico		67,5	75
% de aprobados según estadística		61,6	69,5

campo de experimentación lo referente al algoritmo de entrenamiento a utilizar.

En el sentido de generalización se deberá proceder al diseño de una aplicación que permita el pronóstico para cualquiera de las asignaturas de la carrera, por lo menos en las que pertenecen a las disciplinas de la profesión. Esta aplicación deberá estar conectada con las bases de datos de los sistemas informáticos de control docente existentes en la universidad a los efectos de obtener los conjuntos de datos que permitan entrenar y probar las diferentes redes neuronales.

La aplicación mencionada deberá ser concebida de forma que los parámetros de entrada a usar para la red específica de cada asignatura puedan ser determinados dinámicamente desde su interface, mediante selección. Por supuesto, lo anterior será posible solo para factores en los que se dispongan de los datos en los sistemas informáticos antes mencionados como, por ejemplo, asignaturas precedentes cuyos resultados deberán ser considerados.

Las redes neuronales creadas, entrenadas y probadas deberán ser guardadas en ficheros, lo que es fácil de lograr en MATLAB, a los efectos de su explotación posterior cada vez que se quieran pronosticar los resultados de algún grupo de estudiantes. Por supuesto, deberá existir la facilidad de que se pueda sustituir una red ya existente por otra que demuestre brindar resultados superiores.

CONCLUSIONES

Las redes neuronales constituyen una poderosa herramienta para su aplicación en problemas de pronósticos, en especial para los relacionados con resultados académicos. El cumplimiento de los objetivos del presente trabajo, con buenos índices de eficiencia, es un ejemplo de lo anterior.

Los resultados numéricos obtenidos aquí no necesariamente deben ser tomados como válidos para otras instituciones con una similitud en el currículo, incluso dentro de nuestro propio país y considerando el mismo tipo de variables, puesto que la población considerada no es exactamente la misma en cada caso. No obstante, la estrategia empleada en el trabajo y el modelo para realizar el

pronóstico son aplicables en diversos escenarios, tanto en nuestro contexto como fuera del país.

Para obtener respuestas de una red neuronal que se asemejen lo más posible a los valores correctos es necesaria una adecuada selección de la arquitectura de la red a utilizar, así como del algoritmo de aprendizaje que se ejecutará. También, para la aplicación correcta de las RNA a una determinada situación, se hace necesario una correcta selección de las variables a utilizar y su codificación.

El ajuste de las predicciones de las redes neuronales a las estadísticas significativas de la muestra de datos seleccionada, constituye una condición para validar que las redes neuronales se encuentran bien entrenadas y que son capaces de realizar predicciones correctas.

REFERENCIAS

- [1] G. Corres, L.I. Passoni, C. Zárate y A. Esteban. "Estudio comparativo de modelos de pronósticos de ventas". Iberoamerican Journal of Industrial Engineering. Vol. 6 N° 11, pp. 113-134. Florianópolis, Santa Catarina, Brasil. 2014.
- [2] R. Fernández Galera, J. Ferrando Cervelló, I. Roldán Torres, M. Monteagudo Viana, C. Núñez Piles, D. Dubois Marques, I. Castro Rodríguez y V. Mora Llabata. "Influencia del ingreso en fin de semana en el pronóstico de pacientes ingresados por síndrome coronario agudo". Revista de la Sociedad Española de Cardiología, publicado por Elsevier España, S.L. Vol. 67, Suplemento 1:504 6011-164. España. 2014.
- [3] O. Meryem, J. Ismail and E.M. Mohammed. "A comparative study of predictive algorithms for time series forecasting". Third IEEE International Colloquium in Information Science and Technology (CIST M014). ISBN 978-1-4799-5979-2, pp. 68-73. DOI: 10.1109/CIST.2014.70165 96, Tetouan, Morocco. October, 2014.
- [4] J. Li, X. Ji, Y. Jia, B. Zhu, G. Wang, Z. Li and X. Liu. "Hard Drive Failure Prediction Using Classification and Regression Trees". 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN).

- DOI: 10.1109/DSN.2014.44, pp. 383-394. Atlanta, Georgia, USA. 23-26 June 2014.
- [5] A. Arango, J.D. Velásquez. y C.J. Franco. “Técnicas de lógica difusa en la predicción de índices de mercados de valores: Una revisión de literatura”. Revista Ingenierías, Universidad de Medellín. Vol. 11 N° 22, pp. 115-124. Medellín, Colombia. Enero-junio de 2013. ISSN: 1692-3324.
- [6] U. Johansson, H. Bostrom and T. Lofstrom. “Conformal Prediction Using Decision Trees”. 13th IEEE International Conference on Data Mining (ICDM). ISSN: 1550-4786, pp. 330-339 DOI: 10.1109/ICDM.2013.85. Dallas, Texas. 7-10 December 2013.
- [7] J.J. Heredia, A.G. Rodríguez y J.A. Vilalta. “Predicción del rendimiento en una asignatura empleando la regresión logística ordinal”. Estudios Pedagógicos. Vol. 40 N° 1, pp. 145-162. 2014.
- [8] Organización Internacional de Normalización. “IWA-2: Sistemas de gestión de la calidad para organizaciones educativas”. México. 2002.
- [9] J. Freeman y D. Skapura. “Redes Neuronales. Algoritmos, aplicaciones y técnicas de programación”. Addison Wesley. ISBN: 9780201601152. 1993.
- [10] H. Zeng, L. Zhou, L. Li Li and Y. Wu. “An improved prediction of protein secondary structures based on a multi-mold integrated neural network”. Eighth International Conference on Natural Computation (ICNC), ISBN: 978-1-4577-2130-4. pp. 376-379. DOI: 10.1109/ICNC.2012.6234679. Chongqing, China. 29-31 May 2012.
- [11] I. Lorenzo Fonseca, F. Maciá Pérez, F.J. Mora Gimeno, R. Lau Fernández, J.A. Gil Martínez-Abarca and D.M. Jorquera. “Intrusion Detection method using Neural Networks based on the Reduction of Characteristics”. International Workshop/Work - Conference on Artificial Neural Networks (IWANN). Salamanca, España. 2009. Lecture Notes on Computer Science. Part I, Vol. 5517, pp. 1296-1303. Springer Verlag, Heidelberg. 2009.
- [12] C. Zambrano Matamala, D. Rojas Díaz, K. Carvajal Cuello y G. Acuña Leiva. “Análisis de rendimiento académico estudiantil usando data warehouse y redes neuronales”. Ingeniare. Revista chilena de ingeniería, Vol. 19 N° 3, pp. 369-381. 2011.
- [13] S. Karamouzis and A. Vrettos. “An Artificial Neural Network for Predicting Student Graduation Outcomes”. Proceedings of the World Congress on Engineering and Computer Science. ISBN: 978-988-98671-0-2, pp. 991-994. San Francisco, USA. October, 2008.
- [14] V.O. Oladokun, A.T. Adebajo and O. E. Charles-Owaba. “Predicting Students Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course”. Pacific Journal of Science and Technology. Vol. 9 N° 1, pp.72-79. Nigeria. 2008.
- [15] A.M. Esteban, C.N. Zárate, V.A. Mortara y M.B. Berardi. “Obtención de variables exógenas para mejorar la performance del pronóstico a partir de análisis multivariado”. Iberoamerican Journal of Industrial Engineering. Vol. 6 N° 12, pp. 200-213. Florianópolis, Santa Catarina, Brasil. 2014.
- [16] A. Pitarque, J.C. Ruiz y J.F. Roy. “Las redes neuronales como herramientas estadísticas no paramétricas de clasificación”. Psicotherma 2000. ISSN 0214-9915. Vol. 12 N° 2, pp. 450-463. Valencia, España.
- [17] J.V. Collantes. “Predicción con redes neuronales, comparación con las tecnologías Box y Jenkins”. Tesis para optar al título Magister Scientae en Estadística Aplicada. Mérida, Venezuela. 2001.
- [18] H.O. Sarmiento y W.M. Villa. “Inteligencia Artificial en pronóstico de demanda de energía eléctrica: una aplicación en optimización de recursos energéticos”. Revista Colombiana de Tecnologías de Avanzada. ISSN: 1692-7257. Vol. 2 N° 12. 2008.
- [19] M.G. Longoni, E. Porcel, M.V. López y G.N. Dapozo. “Modelos de Redes Neuronales Perceptrón Multicapa y de Base Radial para la predicción del rendimiento académico de alumnos universitarios”. XVI Congreso Argentino de Ciencias de la Computación (CACIC 2010). ISBN: 978-950-9474-49-9, pp. 692-701. Facultad de Ciencias Exactas y Naturales y Agrimensura. Universidad Nacional del Nordeste. Argentina. 2010.
- [20] Get into PC. “Matlab R2014a”. Fecha de actualización: 11 de diciembre de 2014.

- Fecha de consulta: 8 de abril de 2015. URL: <http://getintopc.com/software/development/matlab-r2014a-free-download/>.
- [21] IBM. “Clementine 11.1 - data mining”. Fecha de actualización: 7 de agosto de 2007. Fecha de consulta: 8 de abril de 2015. URL: <http://www-304.ibm.com/partnerworld/gsd/solutiondetails.do?solution=10387&expand=true>
- [22] Open for innovation “KNIME. KNIME Beginner’s Luck. A Guide to KNIME Data Mining Software for Beginners”. Fecha de actualización: 12 de marzo de 2015. Fecha de consulta: 8 de agosto de 2015. URL: <http://www.knime.org/knimepress/beginners-luck>
- [23] H. Demuth and M. Beale. “Neural Networks Toolbox for use with MATLAB. Manual de Usuario, Versión 4”. The MathWorks Inc. USA. 2002.
- [24] E. Porcel. “Predicción del rendimiento académico de alumnos de primer año de la FACENA (UNNE) en función de su caracterización socioeducativa”. Revista Electrónica de Investigación Educativa (REDIE). ISSN: 1607-4041. Vol. 12 N° 2. Argentina. 2010.
- [25] K. Barker, T. Trafalis and T.R. Rhoads. “Learning from Student Data”. Proceedings of the 2004 IEEE Symposium on Systems and Information Engineering Design. ISBN: 0-9744559-2-X, pp. 79-86. April 2004.