Empleo de la Neutrosofía para predecir la promoción individual de los estudiantes de la carrera de ingeniería informática

Hugo A. Martínez Noriega[1]

[1]Mastér en Ciencias, Profesor Auxiliar del Departamento de Ciencias Básicas, Facultad # 3. Universidad de las Ciencias Informáticas, Ciudad de la Habana, Cuba. Mail: hugomn@uci.cu

Resumen: En la Universidad de las Ciencias Informáticas de La Habana, Cuba, a partir del 2009, se comenzó a realizar en los estudiantes que arriban al centro una serie de diagnósticos pedagógicos, que junto con los datos generales del expediente del alumno constituyen un estudio inicial de las características personales y de aptitudes cognoscitivas de cada uno de ellos; sin embargo, no hay un uso adecuado de los mismos. En el presente trabajo, se emplean diferentes métodos estadísticos los cuales permiten hacer un estudio de las relaciones de estos datos y el primer corte evaluativo del semestre, con los resultados de promoción limpia al finalizar el primer año de la carrera. Además, se estructura la población de estudiantes de ese año en grupos homogéneos en cuanto a las características encontradas y los grupos obtenidos se utilizan como base de un mecanismo de clasificación que reagrupe a los estudiantes de nuevo ingreso del curso entrante, luego del primer corte evaluativo, permitiendo instaurar un sistema de medidas pedagógicas diferenciadas para cada grupo de estudiantes. Se emplean la Neutrosofía para Neutrosofía para predecir la promoción individual de los estudiantes de la carrera de ingeniería informática, después de un previo análisis con técnicas estadísticas.

Palabras claves: Metodología, clasificación, caracterización individual, estrategias, factores de riesgo, análisis de clúster y mecanismo de clasificación.

1 Introducción

El desarrollo científico-técnico contemporáneo y las transformaciones que lleva consigo exigen un mejoramiento progresivo del sistema educacional. En la actualidad la calidad de la educación superior constituye un factor imprescindible en el avance de nuestra sociedad y las condiciones económicas que enfrenta el país ratifican aún más la necesidad de elevar la eficiencia en este nivel de enseñanza. Entre los objetivos de la enseñanza de la Ingeniería en Ciencias Informáticas, ocupa un lugar importante la creación de habilidades y hábitos para la aplicación sistemática del enfoque de procesos y desarrollo de software. Esto contribuirá al perfeccionamiento y optimización de la enseñanza y el aprendizaje; además de dar respuestas a las necesidades sociales existentes. Para lograr ingenieros de alto nivel en estos tiempos, es necesario prepararlos en la modelación matemática de los procesos que analizan en sí, lo cual significa que estas personas deben ser capaces de redefinir lo que tienen que hacer, volver a aprender, volver a instruirse en cómo hacer las nuevas tareas, a través de algoritmos y procedimientos eficaces.

En consecuencia, a ese objetivo tributa la presente investigación, cuya primera etapa se centrará en la realización de un análisis de los trabajos de descubrimiento de conocimiento en los datos, para el apoyo a la toma de decisiones docentes-pedagógicas; como ejemplos de estos tipos de trabajos se pueden citar los siguientes:

- La evaluación del rendimiento académico de los estudiantes en varias asignaturas en un determinado período, [1] y [2].
- Evaluación del rendimiento académico en una asignatura, a partir del rendimiento de otras, [3].
- Evaluación del rendimiento académico, a partir de su relación con variables demográficas y resultados docentes de la enseñanza previa a la universidad, [4], [5].
- Determinar perfiles de estudiantes como base para el establecimiento de estrategias educativas diferenciadas [6].
- Predicción del promedio de cada año académico en función de las características y los resultados docentes previos de los educandos, [7] y [8].
- Predicción del éxito o fracaso académico de los estudiantes al finalizar el primer año [9] y [10].

En las universidades, el desarrollo de trabajos de descubrimiento de conocimiento en los datos ha sido posible por la presencia de dos factores importantes: la existencia de trabajadores con una amplia experiencia en el análisis de datos y el desarrollo de los sistemas de información universitarios [11]. Los análisis a priori o posteriori mostrados anteriormente han contribuido a desarrollar indicadores de eficiencia universitaria [4]. Estos análisis se basan usualmente en estudios estadísticos transversales o longitudinales, con el objetivo de descubrir conocimiento en la información disponible y, por otra parte, sirven de soporte para la toma de decisiones en el ámbito universitario.

Cuando se decide realizar un trabajo de descubrimiento en los datos se tienen en cuenta tres elementos fundamentales que se interrelacionan: los objetivos, el tipo de datos que se pretende procesar y el grado de conocimiento sobre el tema que tengan los investigadores para obtener e interpretar el modelo [12].

Por ejemplo, Grau pretende, predecir la medida de eficiencia o función objetivo: "Se gradúa en tiempo" (Si/No), la cual como se evidencia es una variable dicotómica. Para ello emplea como variables predictivas datos de los estudiantes individuales previos al comienzo de los estudios universitarios, la facultad y la carrera donde están matriculados. Observe que se utilizan solamente datos predictores "epidemiológicos" y no "clínicos", como podría ser el desempeño de esos estudiantes durante el primer y segundo año de la carrera.

Actualmente, son de gran importancia el empleo de gestores de información, según sea el desarrollo alcanzado en el centro de enseñanza. El formato de los soportes de la información docente es casi siempre tabular. Una tabla que guarde datos docentes de un grupo de estudiantes, casi siempre tiene la estructura siguiente: a cada estudiante (instancia) le corresponde una única fila, las primeras columnas contienen usualmente datos identificatorios y el resto de las columnas (atributos) contienen en general resultados académicos notas y algún que otro parámetro como puede ser la asistencia.

El formato tabular en el que se concentra toda la información oficial es muy positivo; sin embargo, su extensión y contenido no permite que se haga un uso adecuado de esta información [13].

En el primer año de la carrera Ingeniería en Ciencias Informáticas, se confecciona una caracterización individual basada en los datos que se recopilan del expediente estudiantil, los diagnósticos cognoscitivos y el primer corte evaluativo. En la Universidad de las Ciencias Informáticas (UCI) se consideró que esta caracterización individual sería de utilidad para el trazado de estrategias pedagógicas, las cuales tenían como propósito evitar el éxodo de estudiantes al finalizar su primer año de estudios universitarios. Sin embargo, la principal deficiencia que posee la caracterización individual desde un punto de vista práctico, está condicionado al hecho, que la universidad presenta grandes matrículas y diseñar para cada estudiante una estrategia resulta realmente complejo.

Teniendo en cuanta lo expresado anteriormente, un aspecto importante a resolver se encuentra relacionado con la forma en que se procesan los datos educacionales para su posterior interpretación y socialización. Es por ello que como resultado clave de esta investigación se presenta una metodología, la cual está dirigida a incrementar la efectividad de las estrategias de intervención pedagógica de los directivos docentes de la UCI. Además, que el conocimiento extraído de la aplicación de la metodología, se pueda socializar e interpretar de manera mejor ha como se hace en la actualidad.

Para lograr validar la metodología se utiliza como caso de estudio el primer año de ingeniería en ciencias informáticas de la Facultad 1 de la UCI, curso 2009-2010.

2 Materiales y métodos

A. *Datos*

Los datos utilizados fueron la ficha de matrícula del estudiante, los diagnósticos pedagógicos, el primer corte evaluativo de los estudiantes de primer año de la Facultad 1 de la (UCI) y las calificaciones finales del semestre, correspondientes al curso 2009-2010. Es importante destacar que en las calificaciones finales del semestre no se incluyeron las correspondientes al examen extraordinario, debido a que la información brindada por las mismas se encontraba fuera del tiempo establecido para aprobar la asignatura y, por consiguiente, se constató que no era favorable para el análisis. Además, las calificaciones finales no se emplean como tal en el estudio, sino que sirven de base para conformar la variable respuesta: "Promueve en tiempo" (Si/No).

La masa de estudiantes se divide en cinco brigadas y el total de estudiantes de primer año de la Facultad 1 de la (UCI) que conforman la muestra es de 140, correspondiente a la cohorte 2009-2010. Los estudiantes que repiten el año y que adicionalmente se le convalidaban asignaturas fueron excluidos del estudio, pues la carga asignada a estos estaba reducida en comparación con el resto del estudiantado.

B. Métodos

Para procesar los datos educacionales se precisa cruzar información que puede provenir de diversas fuentes y generar en consecuencia un modelo que resuma el conocimiento extraído, que se pueda socializar e interpretar de manera mejor a como se hace en la actualidad. Es por ello que la metodología que se propone seguidamente, se basa en combinar métodos estadísticos de tal forma que se pueda indicar los diferentes factores que modifican a la variable respuesta: "Promueve en tiempo" (Si/No).

La metodología se encuentra compuesta de tres fases fundamentales, a partir de los objetivos que se quieren alcanzar en cada una de sus partes. En la primera parte, el objetivo que se pretende alcanzar es el siguiente: identificar los factores que más pueden influir en los resultados finales del semestre. En la segunda parte, el objetivo a lograr es: conformar grupos homogéneos de estudiantes basados en los factores que más influyen en los resultados finales. En la tercera parte el objetivo se centra en: instaurar un mecanismo de clasificación basados en los grupos formados anteriormente para los estudiantes que ingresan a la universidad en el próximo curso.

Para poder identificar los factores que más influyen en los resultados finales del semestre, a partir de la información que se recoge inicialmente, es necesario que exista alguna asociación entre las variables predictoras y la variable de interés "Promueve en tiempo" (Si/No). Esto fue comprobado con los siguientes métodos: análisis univariado de asociación, árbol de decisión CRT y regresión logística binaria. Se utilizaron estos métodos, debido al carácter discreto de la mayoría de las variables, es decir, las variables se encuentran en una escala de medición ordinal u nominal.

El análisis univariado de asociación, está basado en tablas de contingencia [4], pues la mayoría de las variables presentan un carácter discreto. El análisis univariado permitió conocer cuáles variables manifiestan riesgo, es decir, riesgo positivo o riesgo negativo o ni riesgo ni protección en alguna de sus categorías. La exactitud general de este clasificador fue de un 77.2% lo cual es aceptable para aplicaciones en el campo de las ciencias sociales [13]. Las variables asociadas a la muestra general fueron:

- Calificación en el primer corte de matemática I (M.I.1C1), separada en tres categorías ordinales.
- Calificación en el primer corte de matemática discreta I (MD.I.1C1), separada en dos categorías ordinales.
- Centro de Procedencia (C. Procedencia), desglosada en tres categorías ordinales.

El resto de las variables fueron probadas, pero no presentan asociaciones significativas con la variable respuesta: "Promueve en tiempo" (Si/No). Cabe destacar, que las tres variables que poseen asociaciones significativas con la variable respuesta, constituyen factores que influyen en los resultados finales del semestre.

El segundo método que se emplea es el Árbol de Clasificación y Regresión (CRT) [14] y se utiliza este, debido a que el mismo tiene la capacidad de dividir a los datos en segmentos que son los más homogéneos posible respecto a la variable dependiente, lo que permite poder encontrar las variables que manifiestan una incidencia directa en la promoción limpia al finalizar el primer semestre de la carrera Ingeniero en Ciencias Informáticas. Este análisis permitió corroborar los resultados alcanzados en el análisis univariado. Sin embargo, el análisis multivariado (CRT) fue más exhaustivo que el anterior, pues fue capaz de identificar los mismos factores de riesgo que al análisis univariado y por otra parte se detectan otros factores de riesgo que influyen en los resultados de promoción al finalizar el semestre. Los factores fueron los siguientes:

- La estrategia de aprendizaje por autocontrol motivacional (A.C.M), desglosada en sus dos categorías.
- El nivel de escolaridad del padre (N.E.del padre), separado en dos categorías.

La exactitud de la clasificación correcta general alcanzada por el clasificador fue de un 87% aproximadamente, lo cual es muy apropiado para investigaciones en el campo de las Ciencias Sociales. Las variables restantes fueron probadas, pero no manifiestan asociaciones significativas con la variable respuesta:

"Promueve en tiempo" (Si/No). Es importante señalar que las cinco variables que presentan una asociación significativa con la variable respuesta, destacan como factores de riesgo en los resultados de promoción limpia al finalizar el semestre.

El tercer método que se emplea es la regresión logística binaria [15], la cual pretende brindar un modelo predictivo de la condición o estado del estudiante al finalizar el semestre, es decir, saber qué posibilidad tiene el estudiante de promover en tiempo (Si/No). Este instrumento estadístico de análisis multivariado, posibilitó ratificar los resultados alcanzados en los dos análisis anteriores (análisis univariado y análisis multivariado (CRT)), pero a pesar de detectar los mismos factores de riesgo, hubo una variación en el desglose de las categorías. De las seis variables que constituyen factores de riesgo, en cinco se modifican sus categorías. Seguidamente, se puede apreciar cómo se modifican las categorías asociadas a cada variable:

- El nivel de escolaridad del padre solo tuvo en cuenta: es universitario (Si/No).
- La evaluación del corte de matemática I (M.I.1C1) se separa en cuatro categorías:
    - El estudiante se encuentra evaluado de regular (Si/No).
    - El estudiante se encuentra evaluado de bien (Si/No).
- La evaluación en el corte de matemática discreta I (MD.I.1C1) solo tuvo en cuenta: el estudiante está evaluado de bien (Si/No).
- El centro de procedencia solo examino: el estudiante pertenece a un Instituto Preuniversitario Vocacional (Si/No).
- El resultado del test de Autocontrol Motivacional evidencia que el estudiante se encuentra motivado por la carrera (Si/No).

Después de haber analizado los resultados que se obtienen tras la aplicación de los métodos anteriores, se evidencia que las variables obtenidas constituyen factores de riesgo de los resultados de promoción limpia al finalizar el semestre. Dichas variables son modeladas a través de un modelo de recomendación neutrosófico para

para recomendar las variables atender en la predicción de la promoción individual de los estudiantes de la carrera de ingeniería informática.

La Neutrosofía es una nueva rama de la filosofía que estudia el origen, naturaleza y alcance de las neutralidades, así como sus interacciones con diferentes espectros ideacionales, creada por el Profesor Florentin Smarandache [16]. Su teoría fundamental afirma que toda idea tiende a ser neutralizada, disminuida, balaceada por las ideas como un estado de equilibrio.

El término "neutrosófico" se propuso porque "neutrosófico" proviene etimológicamente de la "neutrosofía", que significa conocimiento del pensamiento neutro, y este tercer neutral representa la distinción principal, es decir, la parte neutra indeterminada desconocida (además de la "verdad" "pertenencia" y "falsedad" Componentes de "no pertenencia" que aparecen en la lógica borrosa conjunto). Lógica Neutrosófica es una generalización de la lógica difusa de Zadeh [17], y especialmente de la lógica difusa intuitiva de Atanassov [18], y de otras lógicas.

3 Resultados

A. *Identificación de los factores que más influyen en los resultados finales.*

Los resultados obtenidos por cada una de las técnicas utilizadas para identificar a los factores que más influyen en los resultados finales se detallan en los próximos sub acápites.

A.1 Análisis univariado

El procedimiento univariado se empleó con un total de 36 variables, arrojó que solo tres de estas variables presentaban asociación con la variable respuesta “Promueve en tiempo” (Si/No). Las variables que fueron detectadas como posibles factores de riesgo, a través de la técnica de árboles de decisión “Chaid Exhaustivo” con un estadístico Chi-cuadrado de alta significación fueron las siguientes:

- La “Calificación del primer corte de Matemática I”, con sus correspondientes tres categorías.
- La “Calificación del primer corte de Matemática Discreta I”, con solo dos de sus categorías originales.
- El “Centro de Procedencia”, con solo tres de sus categorías originales.

En la tabla 1, se muestran las variables seleccionadas, con sus categorías, la significación del test exacto de Fisher, la V-Cramer (Phi), el riesgo relativo y su intervalo de confianza para las categorías asociadas a los posibles factores de riesgo.

Categorías	Sig. del test exacto de Fisher	V de Cramer (Phi)	Riesgo Relativo (Si/No)	Intervalo de Confianza 95% para el riesgo
E.Mal.M.I.1.C.1	0.000	0.500	2.393	1.776-3.224
E. Regular.M.I.1.C.1	0.001	0.008	1.729	1.017-1.419
E. Bien.M.I.1.C.1	0.000	- 0.518	0.313	0.212-0.464
E. Regular.MD.I.1.C.1	0.000	0.467	2.240	1.686-2.976
E.Bien.MD.I.1.C.1	0.000	- 0.467	0.219	0.108-0.446
T.Medio.C.Proced.	0.01	- 0.031	0.946	0.699-1.280
IP. C.Proced.	0.02	0.252	1.555	1.181-2.046
IPV. C.Proced.	0.04	0.241	0.572	0.402-0.814

Tabla 1: Medidas de asociación en la tabulación cruzada entre cada categoría de "Posibles factores de riesgo" y "No promueve en tiempo". **Fuente**: Elaboración propia.

El resto de las variables fueron probadas, pero no aparecen en la tabla anterior, a pesar que el árbol de decisión fue forzado a romper por estas variables y el mismo fue incapaz de discretizarlas para obtener categorías asociadas a la variable respuesta "Promueve en tiempo" (Si/No). Sumando los valores positivos de Phi máximos de cada variable y los valores negativos de Phi mínimos resultan *PhiMax*=1.219 y *PhiMin*=-1.295. Se puede hacer entonces un pronóstico del riesgo de no terminar en tiempo para cada estudiante individual. Para ello bastará sumar algebraicamente los valores Phi correspondientes a la categoría de cada variable presente en ese estudiante y estandarizarlo según la fórmula 1:

$$PhiScore = \frac{Phi - PhiMin}{PhiMax - PhiMin} \quad (1)$$

Por ejemplo, un estudiante con calificación de regular en el primer corte de la asignatura de Matemática I (M.I) y calificación de mal o regular en el primer corte de la asignatura de Matemática Discreta I (MD.I), reporta un valor total de Phi igual a 0.475, que estandarizado con PhiMin y PhiMax se convierte en 0.70. Cuando se compara con el umbral (0.51 que se evidencia en la Figura 1) el pronóstico fue entonces "No promueve en tiempo" (0.70>0.51). Análogamente, otro estudiante con calificación de bien en el primer corte de la asignatura Matemática I (M.I), calificación de bien en el primer corte de la asignatura de Matemática Discreta I (MD.I) y que provenga de un Instituto Preuniversitario Vocacional (IPV), alcanza un valor total de Phi igual a - 0.744, que estandarizando con PhiMin y PhiMax se transforma en 0.22. Cuando se compara con el umbral (0.51 que se evidencia en la Figura 4) el pronóstico fue entonces "Promueve en tiempo" (0.22<0.51). Realmente ambos ejemplos existen en la base de datos y al finalizar el semestre se evidenció, que las causas principales que incidieron en el caso del estudiante que no promueve en tiempo, fueron las asignaturas de matemática.

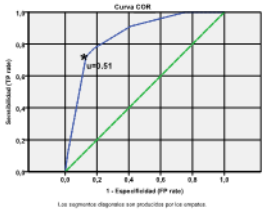

Figura 1: Curva continua (ROC) muestra la relación entre la razón de Verdaderos Positivos y los Falsos Positivos en el clasificador cuando el umbral se mueve en el intervalo [0; 1]. El punto marcado con asterisco está suficientemente cerca del vértice superior izquierdo del cuadrado y corresponde a **u=0.51**. **Fuente**: Elaboración propia.

Con este umbral se puede tener una razón de Falsos Positivos de 0.180 y una razón de Verdaderos Positivos de 0.72, así como una exactitud general de 77.2%. Se explícita, que aquí "positivo" significa "riesgo de no promover en tiempo", es decir, si se establece una analogía con la epidemiología donde "positivo" significa "posiblemente enfermo". Los resultados obtenidos anteriormente no son sorprendentes, no obstante, evidencian que se puede realizar un pronóstico con este análisis univariado.

A.2. Árbol de Decisión CRT

En la figura 2 se muestra el árbol de decisión obtenido para la muestra en general. Las interacciones que se detectan son interesantes, pues en primer lugar se conjugan todas las variables utilizadas hasta el momento en los anteriores análisis, es decir: el Centro de Procedencia del estudiante (C.Procedencia), la calificación en el primer corte en las asignaturas de matemáticas (MD.I y M.I); también, se detectan la Estrategia de Aprendizaje por Autocontrol Motivacional (A.C.M) y el Nivel de Escolaridad del Padre (N.E. del padre). El total de nodos en este árbol es igual a 11, el número total de nodos terminales del árbol es igual a 6, que a su vez tienen asociados 6 caminos o trayectorias para llegar a cada uno de ellos. Para predecir la medida de eficiencia desde el nodo raíz hasta el nodo terminal, se define en cada caso una regla.

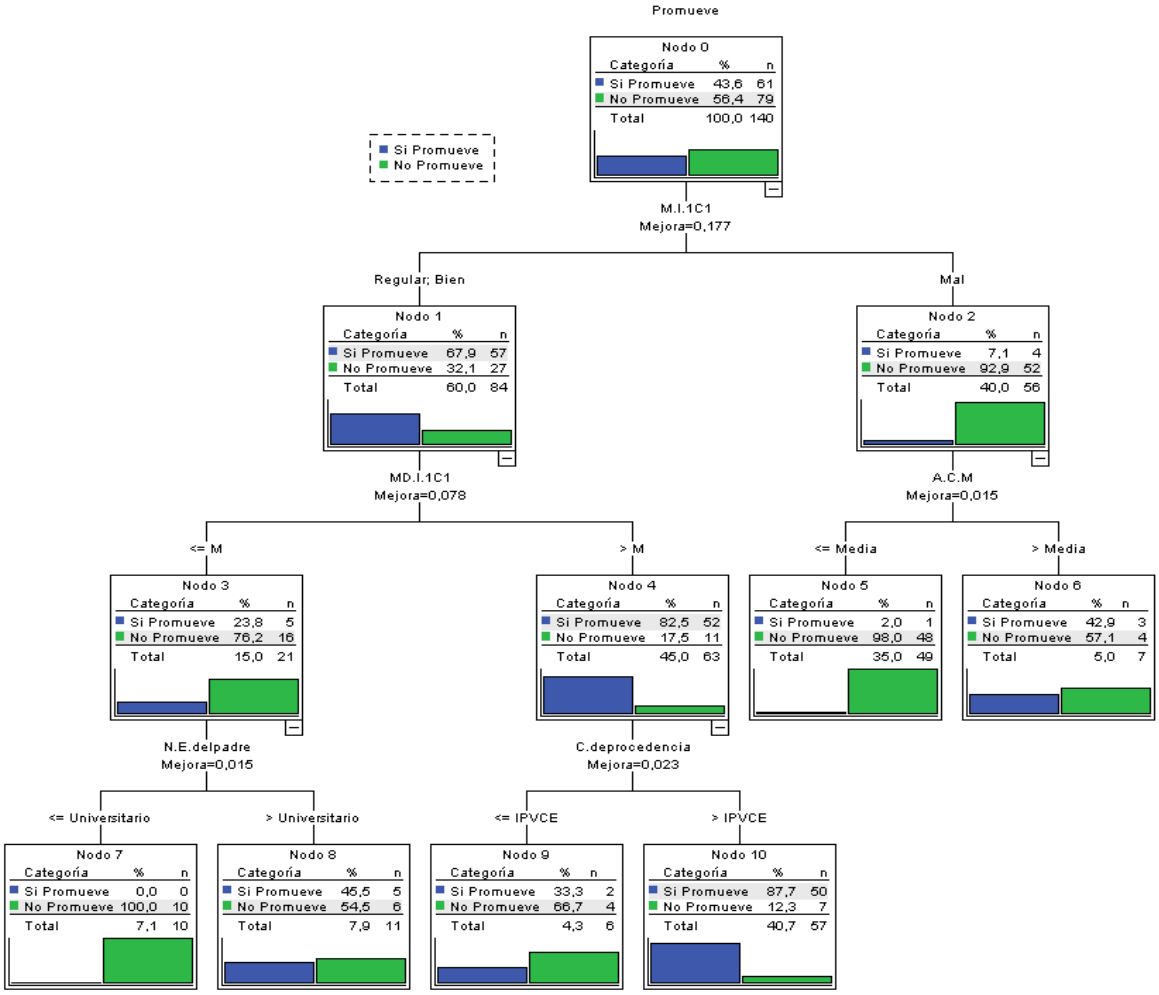

Figura 2: Árbol de decisión obtenido por la técnica CRT. **Fuente**: Elaboración propia.

La matriz de confusión del clasificador empleado anteriormente, evidencia los excelentes resultados que se obtienen al utilizar este clasificador multivariado. El porcentaje de estudiantes con tendencia a no promover en tiempo es de un 91% aproximadamente y el porcentaje de estudiantes con tendencia a promover en tiempo es de un 82%, como se puede apreciar el porcentaje en ambos casos es bueno, demostrando la potencia del clasificador. Igualmente, la exactitud general alcanza un buen nivel del 87% aproximadamente, lo cual es muy apropiado para investigaciones en el campo de las Ciencias Sociales, Tabla 2.

Clasificación

Observado	Pronosticado		
	Si Promueve	No Promueve	Porcentaje correcto
Si Promueve	50	11	82,0%
No Promueve	7	72	91,1%
Porcentaje global	40,7%	59,3%	87,1%

Tabla 2: Desempeño del clasificador multivariado. **Fuente**: Elaboración propia

A.III Regresión Logística Binaria

La regresión logística binaria pretende a partir de los datos educacionales obtener un modelo predictivo de la condición o estado del estudiante al finalizar el semestre, es decir, saber qué posibilidad tiene el estudiante de Promover en tiempo (Si/No).

La ecuación de regresión logística que se construye en el presente estudio sería:

$$P(\text{Estado} = \text{No Promueve}) = \frac{1}{1+\exp(3.879-1.560xN.Edelpadre\_U+2.920xAC.M-1.425xM.I.1C1\_R-2.978xM.I.1C1\_B-2.361xMD.I.1C1\_\text{B-}1.548\text{xC.deprocedencia\_IPV})} \quad (2)$$

La ecuación 2, puede ser utilizada para predecir la probabilidad de tener el resultado (Estado) de "No Promueve" de un estudiante que presenta los factores de riesgo anteriormente expuestos. Así, un estudiante que su padre no tenga un nivel de escolaridad universitario, que se evaluación en matemática I sea de mal, que su evaluación en matemática discreta I sea de mal o de regular y que el centro de procedencia no sea un instituto preuniversitario vocacional, presenta una probabilidad de *No Promover en tiempo* igual a:

$$P(\text{Estado} = \text{No Promueve}) = \frac{1}{1+\exp(3.879-1.560x1+2.920x0.5-1.425x1-2.978x1-2.361x1\text{-}1.548x1)} = 0.98 \quad (3)$$

Por lo tanto, con esta probabilidad predicha, como es mayor que 0.65 se clasificaría como "Estado=No Promueve en tiempo".

La tabla 3 es un extracto de los pronósticos hechos por el modelo de regresión logística, se conoce que las seis variables que conforman al modelo presentan dos categorías (1 ó 0). Por lo tanto, se tendrían 2^6 combinaciones posibles a partir de los dos niveles en los que se miden cada variable predictora del modelo, lo que equivale a 64 evaluaciones del modelo de regresión logística. Se puede apreciar que el modelo es capaz de diferenciar de forma correcta los estudiantes que promueven de los que no promueven, lo cual es muy útil para la toma de decisiones de los directivos docentes.

N.E.P	A.C.M	M.I.1C1_R	M.I.1C1_B	MD.I.1C1_B	C.P_IPV	Pronóstico	Promueve
0	0	1	1	1	0	0,95	No Promueve
0	0	1	1	1	1	0,99	No Promueve
1	0	1	1	1	0	0,99	No Promueve
1	0	1	1	1	1	1	No Promueve
0	1	1	1	1	0	0,22	Promueve
0	1	0	1	1	0	0,14	Promueve
1	1	1	1	1	1	0,96	No Promueve
1	1	0	1	1	1	0,9	No Promueve
0	0	0	0	0	0	0,02	Promueve
0	0	0	0	0	1	0,09	Promueve
0	1	0	0	0	0	0,01	Promueve
0	0	1	0	0	0	0,29	Promueve

Tabla 3: Resultados de la Regresión Logística. **Fuente:** Elaboración propia.

B. *Análisis de Clusters*

Para formar los grupos de estudiantes se utilizaron los siguientes factores de riesgo:

- El estudiante proviene de un Instituto Preuniversitario Vocacional de Ciencias Exactas (IPVCE), (Si/No).
- El padre es universitario, (Si/No).
- El resultado del test de Autocontrol Motivacional evidencia que el estudiante se apega a este, (Si/No).
- El estudiante está evaluado de Bien en el primer corte de la asignatura Matemática I, (Si/No).
- El estudiante está evaluado de Bien en el primer corte de la asignatura de Matemática Discreta I, (Si/No).

En la selección de las anteriores variables como factores de riesgo intervinieron los pedagogos y directivos docentes, los cuales consideran que los factores detectados son de utilidad para formarse una idea del posible desempeño de los estudiantes al finalizar el semestre.

Para construir los conglomerados se empleó el método de K-Modas, con la distancia de coincidencia simple de Kant. En la tabla 4 se muestra una caracterización de los cinco conglomerados que se conformaron. La tabla 4 se compone de tres columnas con las siguientes particularidades: en la primera columna se etiqueta a cada conglomerado, en la segunda columna se reflejan las características que distinguen a cada clúster y en la tercera columna se exhibe la cantidad de estudiantes que conforman al conglomerado y el porciento que estos representan en la muestra utilizada. El color rojo representa a las características que constituyen factores de riesgo, mientras que el color azul representa a los factores de confianza o protección.

Número	Conglomerado	Composición (%)
1	**Padre no universitario**, **Evaluación de B en M.I**, **Provienen de IPVCE**, **Evaluación de B en MD.I**.	**26 (18.57%)**
2	**No evaluados de B en MD.I**, **No evaluados de B en M.I**, **Se apegan a A.C.M**, **Provienen de IPVCE**, **Padre universitario**.	**36 (25.71 %)**
3	**No se apegan a A.C.M**, **No evaluados de B en MD.I**, **Evaluación de B en M.I**, **Proceden de IPVCE**.	**34 (24.29 %)**

4	**No se apegan a A.C.M**, **No evaluados de B en M.I**, **No evaluados de B en MD.I**.	**24 (17.14 %)**
5	**Padre universitario,** **Se apegan a A.C.M**, **Evaluados de B en M.I**, **Evaluados de B en MD.I**.	**20 (14.29%)**

Tabla 4. Caracterización de Conglomerados**. Fuente:** Elaboración propia.

A partir de la caracterización de cada uno de los conglomerados se propone la escala de riesgo para la promoción de los estudiantes, la cual servirá de apoyo para definir las estrategias a seguir con cada uno de los grupos formados (Ver Figura 3).

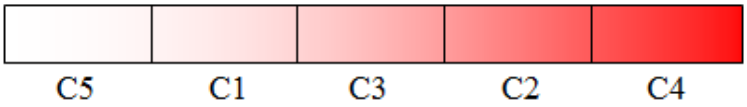

Figura 3. Escala de riesgo en los conglomerados. **Fuente**: Elaboración propia.

C. *Mecanismo de Clasificación*

Para establecer un mecanismo de clasificación de la nueva matrícula, se utilizará la técnica multivariante de análisis discriminante. Se procesaron los 140 casos válidos, lo cual representa el 100% de la muestra escogida; por lo tanto, no se excluye del análisis ningún caso por tener al menos un valor perdido en alguna de las variables discriminantes.

Los resultados de la clasificación arrojan que hay sólo 4 casos mal clasificados y el porcentaje de clasificación general fue de un 97.1%. Sin embargo, utilizando validación cruzada se obtuvo un 92.1% de casos bien clasificados. Los resultados alcanzados, evidencian que el empleo de las funciones discriminantes sería de gran utilidad para elaborar procedimientos de clasificación sistemática de individuos nuevos.

Para obtener las funciones discriminantes se utilizan los coeficientes de clasificación propuestos por Fisher en 1936 (ver, (Peña; 2002)) los cuales se utilizan únicamente para la clasificación Luego, a partir de los coeficientes se obtiene una función de clasificación para cada grupo; cada una de estas funciones se evalúa para un sujeto dado y se clasifica al sujeto en el grupo en el cual la función obtenga una mayor puntuación.

Función discriminante para el grupo 1:

$$D_1 = -6.724 - 0.565C.\Pr\_IPVCE + 8.589N.E.Padre\_U + 5.438A.C.M - 1.288M.I.\_B + 0.713MD.I\_B$$

Función discriminante para el grupo 2:

$$D_2 = -8.270 - 0.850C.\Pr\_IPVCE - 1.198N.E.Padre\_U + 2.844A.C.M + 10.675M.I.\_B + 7.583MD.I\_B$$

Función discriminante para el grupo 3:

$$D_3 = -14.634 - 0.940C.\Pr\_IPVCE - 0.270N.E.Padre\_U + 21.758A.C.M + 2.026M.I.\_B + 7.058MD.I\_B$$

Función discriminante para el grupo 4:

$$D_4 = -25.385 - 1.871C.\Pr\_IPVCE - 1.033N.E.Padre\_U + 23.335A.C.M + 16.547M.I.\_B + 9.568MD.I\_B$$

Función discriminante para el grupo 5:

$$D_5 = -3.820 + 4.700C.\Pr\_IPVCE - 1.340N.E.Padre\_U + 1.985A.C.M - 0.534M.I.\_B + 1.429MD.I\_B$$

Basado en los resultados obtenidos se utiliza un modelo de recomendación, el cual es útil para recomendar las variables atender para predecir la promoción individual de los estudiantes de la carrera de ingeniería informática. El modelo de recomendación a desarrollar parte de la información que recojan estos modelos y de los algoritmos utilizados para generar las recomendaciones, en este sentido se distinguen las técnicas referidas por [19, 20].

Los modelos de recomendación basados en conocimiento realizan sugerencias haciendo inferencias sobre las necesidades de los expertos según [20, 21]. El enfoque basado en conocimiento se distingue en el sentido que usan conocimiento sobre cómo el objeto de estudio, en particular, puede satisfacer las necesidades requeridas, y por lo tanto tiene la capacidad de razonar sobre puede satisfacer las necesidades del usuario, y por lo tanto tiene la capacidad de razonar sobre la relación entre una necesidad y la posible recomendación que se mostrará.

Este tipo de modelo se basa en la construcción de perfiles de usuarios como una estructura de conocimiento que apoye la inferencia la cual puede ser enriquecida con la utilización de expresiones que emplea lenguaje natural [20, 22]. El flujo de trabajo en el presente estudio se basa en la propuesta de Cordón [20, 23] para sistemas de recomendación basados en conocimiento permitiendo representar términos lingüísticos y la indeterminación mediante conjuntos neutrosóficos de valor único (SVN), [24]. En la figura 4 se muestra el flujo de trabajo.

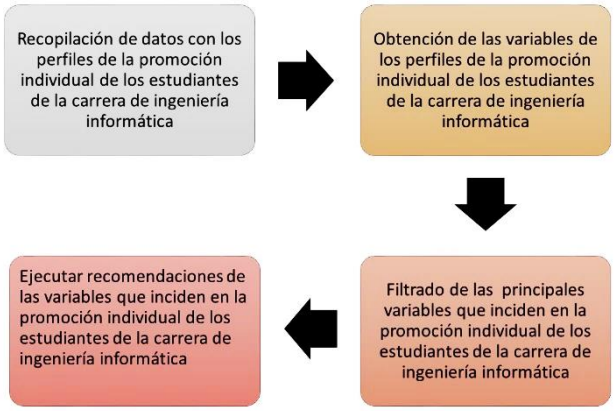

Figura 4. Modelo de recomendación propuesto. **Fuente**: Elaboración propia.

La descripción detallada de cada una de sus actividades y del modelo matemático que soporta la propuesta es presentada a continuación:

1 Recopilación de datos con los perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática

Cada una de las concepciones se describen por un conjunto de características que conformarán los perfiles de las concepciones del derecho a la vida.

$$C = \{c1, \dots, ck, \dots, cl\} \tag{4}$$

Para la obtención de la base de datos de las variables de los perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática se obtiene mediante números neutrosóficos de valor único (SVN) [25, 26].

Sea $A* = (A1*, A2*, .., An*)$ sea un vector de números SVN tal que $Aj* = (aj*, bj*, cj*)$ $j = (1,2, \dots, n)$ y $Bi = (Bi1, Bi2, \dots, Bim)$ $(i = 1,2, \dots, m)$ sean m vectores de n (SVN) números tal que y $Bij = (aij, bij, cij)$ $(i = 1,2, \dots, m), (j = 1,2, \dots, n)$ entonces la distancia euclidiana es definida como. Las Bi y $A*$ resulta [25]:

$$d_i=\left(\frac{1}{3}\sum_{j=1}^{n}\left\{\left(|a_{ij}-a_j^*|\right)^2+\left(|b_{ij}-b_j^*|\right)^2+\left(|c_{ij}-c_j^*|\right)^2\right\}\right)^{\frac{1}{2}}$$
$$(i = 1,2, \dots , m) \quad (5)$$

A partir de la obtención de la distancia euclidiana se define una medida de similitud según refiere [27].

En la medida en que la alternativa Ai se encuentre más cercana al perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática (si) mayor será la similitud, lo que permite establecer un orden entre alternativas según [28]. El perfil de la promoción individual de los estudiantes de la carrera de ingeniería informática puede ser obtenido de forma directa a partir de los expertos, para ello se utiliza la ecuación 6.

$$F_{a_j}=\{v_1^j,\dots,v_k^j,\dots v_l^j\},j=1,\dots n \quad (6)$$

Las valoraciones de las variables de los perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática, aj, serán expresadas utilizando la escala lingüística S, $v_k^j \in$S donde $S = \{s1, \dots, sg\}$ correspondiente al segundo conjunto de término lingüísticos definidos para evaluar las variables c_k utilizando los números SVN. Para esto los términos lingüísticos a emplear son definidos una vez descrito el conjunto de variables de los perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática y se representan según la expresión 7.

$$A = \{a1, \dots, aj, \dots, an\} \quad (7)$$

El conjunto de las variables de los perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática se guarda en una Base de Datos previamente creada.

2 Obtención de las variables de los perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática

En esta fase se obtiene las principales variables de los perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática, almacenándose ellas en un perfil tal y como se muestra en la expresión 8.

$$P_e=\{P_1^e,\dots,P_q^e,\dots,P_l^e\} \quad (8)$$

Este perfil estará integrado por un conjunto de atributos que para su interpretación se representan a través de la expresión 9.

$$C_e=\{c_1^e,\dots,c_k^e,\dots,c_l^e\} \quad (9)$$

Donde: $c_k^e \in S$

El perfil relacionado con las variables de los perfiles de la promoción individual de los estudiantes de la carrera de ingeniería informática se obtiene mediante el llamado enfoque conversacional o mediante ejemplos los cuales pueden ser adaptados según refiere [29].

3 Filtrado de las principales variables que inciden en la promoción individual de los estudiantes de la carrera de ingeniería informática

En esta fase se filtran las principales variables que inciden en la promoción individual de los estudiantes de la carrera de ingeniería informática para encontrar cuáles son las más adecuadas. Para ello se calcula la similitud entre los perfiles que incide en la promoción individual de los estudiantes de la carrera de informáticas, Pe, y cada variable relativa a cada perfil, aj, registrada en la Base de Datos. Para el cálculo de la similitud total se emplea la expresión 10.

$$S_i = 1 - \left(\left(\frac{1}{3} \sum_{j=1}^{n} \left\{ (|a_{ij} - a_j^*|)^2 + (|b_{ij} - b_j^*|)^2 + (|c_{ij} - c_j^*|)^2 \right\} \right)^{\frac{1}{2}} \right) \quad (10)$$

La función S calcula la similitud entre los valores de los atributos del perfil de cada concepción relacionada con el derecho a la vida y la de cada concepción analizada en el presente estudio, aj [27].

4 Ejecutar recomendaciones

Calculada la similitud entre los perfiles que incide en la promoción individual de los estudiantes de la carrera de informáticas y cada variable relativa a cada perfil, los resultados se ordenan de acuerdo a la similitud obtenida, ellas se representan según el vector de similitud que se representa en la expresión 11.

$$D=(d_1,\ldots,d_n) \quad (11)$$

Las variables atender serán aquellas que mejor satisfagan las necesidades de los perfiles que incide en la promoción individual de los estudiantes de la carrera de informática, es decir las que poseen mayor similitud con las variables relacionadas con la promoción individual de los estudiantes.

A partir del modelo de recomendación propuesto se obtienen las valoraciones de las variables relacionadas con la promoción individual de los estudiantes de la carrera de ingeniería informática a través de la expresión definida en 7, $A = \{a1, a2, a3, a4, a5\}$. Estas variables se describen por el conjunto de atributos $C = \{c1, c2, c3, c4, , c5\}$.

El conjunto de atributos se valorará a través de la escala lingüística que se muestra en la tabla 5. Estas valoraciones se almacenaron en una Base de Datos, previamente creada para recomendar cuales son las variables a tener en cuenta para promoción individual de los estudiantes de la carrera de ingeniería informática.

Término lingüístico	**Números SVN**
Extremadamente buena (EB)	(1,0,0)
Muy muy buena (MMB)	(0.9, 0.1, 0.1)
Muy buena (MB)	(0.8,0,15,0.20)
Buena(B)	(0.70,0.25,0.30)
Medianamente buena (MDB)	(0.60,0.35,0.40)
Media(M)	(0.50,0.50,0.50)
Medianamente mala (MDM)	(0.40,0.65,0.60)
Mala (MA)	(0.30,0.75,0.70)
Muy mala (MM)	(0.20,0.85,0.80)
Muy muy mala (MMM)	(0.10,0.90,0.90)
Extremadamente mala (EM)	(0,1,1)

Tabla 5: Términos lingüísticos empleados [25].

Las recomendaciones dada la información relacionada con las variables estudiadas en el presente estudio y de acuerdo a los términos lingüísticos que se muestran en la tabla 1, se muestran en la expresión 12.

$$Pe = \{MB, MMB, MB\} \quad (12)$$

Basado en la expresión 12, la variable 1 relacionada con el nivel de escolaridad del padre solo tuvo en cuenta: es universitario, obtiene valores muy bueno (MB), la variables 2 relacionada con el centro de procedencia solo examino: el estudiante pertenece a un Instituto Preuniversitario Vocacional, obtiene valores muy buenos (MMB), para variable 3, relacionada con los resultado del test de Autocontrol Motivacional evidencia que el estudiante se encuentra motivado por la carrera obtiene valores medianamente buenos (MB).

Los resultados obtenidos en las recomendaciones sostienen que, de las tres variables de mayor incidencia en la promoción individual de los estudiantes de la carrera de informática, han tenido recepción en la literatura, destacándose la variable 2, relativa al centro educacional de procedencia del estudiante.

Obtenidas las recomendaciones se calcula la similitud entre las variables relacionadas con la promoción individual de los estudiantes de la carrera de informática y las caracteristicas de los perfiles de los estudiantes de la carrera de informática, específicamente de las tres variables estudiadas se obtienen los resultados que se muestran en la tabla 6.

$\boldsymbol{a}1$	$\boldsymbol{a}2$	$\boldsymbol{a}3$
0.52	0.90	0.80

Tabla 6: Similitud entre las variables relacionadas con la promoción individual de los estudiantes de la carrera de informática y las caracteristicas de los perfiles de los estudiantes. **Fuente:** Elaboración propia.

Basado en los resultados obtenidos se recomienda aquellas variables que más se acerquen al perfil relacionado con con la promoción individual de los estudiantes de la carrera de informática. Un ordenamiento de las caracteristica de acuerdo con la comparación sería $\{a2, a3, a1\}$.

En caso de una recomendación de los perfiles relacionado con la promoción individual de los estudiantes de la carrera de informática y las caracteristicas de los perfiles de los estudiantes, se recomienda en el presente estudio atender los dos perfiles más cercanos, ello serían las recomendaciones, $a2$, $a3$, correspondiente con la procedencia de los estudiante y sobre las que poseen resultado del test de Autocontrol Motivacional, donde se evidencia que los estudiantes se encuentra motivado por la carrera.

Conclusiones

Como resultado de este trabajo, se obtuvo:

- Una metodología para la clasificación de los estudiantes en cuanto aquellas características que influirán más en su futura promoción en tiempo al final del curso, a partir de los datos que se proporcionan en la etapa inicial del primer año de la carrera de Ingeniería en Ciencias Informáticas.
- Se logró determinar a partir de las variables disponibles en el estudio, los factores que más influyen en los resultados finales del semestre.
- Se conformaron grupos homogéneos de estudiantes basados en los factores que más influyen en los resultados finales del semestre.
- Se diseñó un mecanismo de clasificación para los nuevos estudiantes que arriban a la universidad, apoyado en los grupos anteriores.
- La metodología planteada puede ser empleada para diagnosticar a los estudiantes casi al comienzo de la carrera de Ingeniería en Ciencias Informáticas, pues permite determinar a qué tipo preestablecido de estudiantes pertenece en cuanto a sus características iníciales en función de su posibilidad de promoción

limpia, cómo habían solicitado los pedagogos que tienen la responsabilidad de su correspondiente atención diferenciada, pero masiva en función del tamaño de las matrículas.

- A través del modelo de recomendación neutrosófico se obtuvo las recomendaciones correspondientes con la procedencia de los estudiantes y las que poseen resultado del test de Autocontrol Motivacional, donde se evidencia que los estudiantes se encuentran motivado por la carrera, el modelo de recomendación neutrosófico siguió un enfoque basado en conocimiento, específicamente el modelo se basa en el empleo de los números SVN para expresar términos lingüísticos.

Referencias

[1] Garnica, E. El rendimiento estudiantil: una metodología para su medición. Revista de Economía, (1997), 13: 5-26.
[2] Moral de la Rubia, J. Predicción del rendimiento académico universitario. Perfiles Educativos, (2006), 28(113): 43-61.
[3] Rodríguez G., y. C. Análisis Multivariado. La Habana, (2009). Instituto Superior Politécnico "José Antonio Echeverría".
[4] Grau, R. La eficiencia en la graduación universitaria analizada con descubrimiento de conocimientos en la base de estudiantes de la Universidad Central de las Villas, (2012), 9.
[5] Gallander, M., B. Dilouya. Academic achievement in first-year university: who maintains their high school average, (2011). The transition to university project. Canada.
[6] Delavari, N. a. M. R. B. Data Mining Application in Higher Learning Institutions. Informatics in Education, (2008), 7(1): 31-54.
[7] Espinosa, I. a. S. P. Obtención de Reglas y Patrones en el Proceso Académico de la Universidad de Ciencias Informáticas. CEIS. La Habana, Instituto Superior Politécnico José Antonio Echeverría. (2007). Tesis de Diploma.
[8] Brito, R. Minería de Datos aplicada a la Gestión Docente del Instituto Superior Politécnico José Antonio Echeverría. CEIS. La Habana, (2008). Instituto Superior Politécnico José Antonio Echeverría. Master.
[9] Kant., S. S. K. a. S. Computation of initial modes for K-modes clustering algorithm using evidence accumulation in Proceedings of the 20th international joint conference on Artificial intelligence (IJCAI): (2007). 2784-2789.
[10] Remón, N. (2009). Análisis para la predicción del éxito o fracaso académico de estudiantes de la Universidad de las Ciencias Informáticas mediante la teoría de conjuntos aproximados. Facultad 5. La Habana, Universidad de las Ciencias Informáticas. Tesis de Diploma.
[11] Rico, J. J. H. Análisis de datos en apoyo a la productividad en el proceso de formación de ingenieros Facultad de Ingenieria Industrial. La Habana, Instituto Superior Plitécnico José Antonio Echevarría: (2011), 1-127.
[12] Rodríguez, A. a. J. H. Rediseño de procesos de gestión de la enseñanza basado en tecnologías informativas. Novena Semana Tecnológica. Las TIC presente y futuro. (2009). La Habana, Cuba, MIC.
[13] Álvarez, J. H. Q. Procesamiento del Diagnóstico Pedagógico mediante Algoritmos de Minería de Datos. Departamento de Ciencias Básicas, Facultad 1. La Habana, Universidad de las Ciencias Informáticas (UCI): (2012), 1-83.
[14] Trevor Hastie, R. T. a. J. F. The Elements of Statistical Learning.Data Mining, Inference, and Prediction, Second Edition. (2008), Springer Series in Statistics.
[15] Peña, D. Análisis de Datos Multivariantes. España, (2002). McGraw-Hill.
[16]. M. Leyva, F. Samarandache. Neutrosofía: Nuevos avances en el tratamiento de la incertidumbre, 2018. Pons Publishing House / Pons asbl Quai du Batelage, 5 1000 – Bruxelles, Belgium. DTP: George Lukacs ISBN 978-1-59973-572-6, Bruselas.
[17]. Zadeh, L.A., Fuzzy sets. Information and Control, 1965.8(3): p. 338-353.
[18]. K.T. Atanassov. Intuitionistic fuzzy sets. Fuzzy sets and Systems, 1986. 20(1): p. 87-96.
[19]. J. Dietmar. Tutorial: Recommender Systems, inInternational Joint Conference on Artificial Intelligence Beijing, August 4, 2013. 2013.

[20]. Cordón, L.G.P., Modelos de recomendación con falta de información. Aplicaciones al sector turístico, 2008. Universidad de Jaén.
[21] [J. B. Freire, et al. Modelo de recomendación de productos basado en computación con palabras y operadores OWA [Aproduct recommendation model based on computing withword and OWA operators]. International Journal ofInnovation and Applied Studies, 2016. 16(1): p. 78.
[22] Herrera, F. and L. Martínez, A 2-tuple fuzzy linguistic representation model for computing with words. Fuzzy Systems, IEEE Transactions on, 2000. 8(6): p. 746-752.
[23] M.R.M. Arroyave, A. F. Estrada, and R.C. González. Modelo de recomendación para la orientación vocacional basado en la computación con palabras [Recommendation models for vocational orientation based on computing with words]. International Journal of Innovation and Applied Studies, 2016. 15(1): p. 80.
[24] H. Wang, et al. Single valued neutrosophics sets. Review of the Air Force Academy, 2010(1): p. 10.
[25] Şahin, R. and M. Yiğider, A Multi-criteria neutrosophic group decision making metod based TOPSIS for supplier selection. arXiv preprint arXiv:1412.5077, 2014.
[26] J. Ye. Single-valued neutrosophic minimum spanning treeand its clustering method. Journal of intelligent Systems, 2014. 23(3): p. 311-324.
[27] K. Pérez-Teruel, M. Leyva-Vázquez, and V. Estrada-Sentí. Mental model's consensus process using fuzzy cognitivemaps and computing with words. Ingeniería y Universidad, 2015. 19(1): p. 173-188.
[28] M.Y. L. Vázquez, et al. Modelo para el análisis de escenarios basados en mapas cognitivos difusos: estudio de caso en software biomédico. Ingeniería y Universidad: Engineering for Development, 2013. 17(2): p. 375-390
[29] L. Pérez. Modelo de recomendación con falta de información. Aplicaciones al sector turístico. 2008, Tesis-doctoral. Universidad de Jaén.