

Artículo 7. Comparación del nivel de desempeño de una competencia usando tres instrumentos, dos basados en rúbrica y otro basado en lógica difusa.

A comparison of the level of competency using three instruments; two rubric based instruments and a fuzzy logic-based instrument.

AUTORES

José Emilio Sánchez García
Alberto Valdez Sandoval
Jesús Eduardo Soto Vega
Brenda Edith Gutiérrez Herrera

Resumen

En la presente investigación, tiene como propósito comparar las aproximaciones en el nivel de desempeño de una competencia usando tres instrumentos de medición, dos basados en rúbrica y otro basado en lógica difusa. Se planteó, para lograr este objetivo, un diseño metodológico de carácter cuantitativo. En los resultados, se encontró suficiente

evidencia para concluir que el instrumento basado en lógica difusa resultó más preciso y exacto. El estudio reafirma la superioridad y poder de cómputo de los modelos matemáticos que la inteligencia artificial pone a disposición contra algoritmos basados en lógica clásica o bivaluadas.

Palabras clave

Evaluación del aprendizaje por competencias, instrumentos de medición, lógica difusa, rúbricas.

Abstract

This investigation is intended to compare approximations in the level of competency using three measurement instruments two of which are rubric based and the third is based on Fuzzy logic. To be able to accomplish this objective, we proposed a qualitative methodological design. Within the results, we found sufficient evidence to conclude that the instrument based on Fuzzy logic was more precise and exact. The study reaffirms the superiority and greater computation power from mathematical models provided by artificial intelligence compared to algorithms based on classical or 2-valued logic.

Keywords

Competence based learning assessment, measurement instruments, Fuzzy logic, rubrics.

Introducción

La lógica difusa (LD) es una rama de la inteligencia artificial (IA), esta última es, a su vez, una rama de las ciencias computacionales que busca resolver problemas mediante modelos matemáticos que emulen procesos mentales humanos por medio de un computador. Los procesos mentales que estudia la IA son los relacionados con el aprendizaje y razonamiento. La LD se enfoca en el razonamiento y no en el aprendizaje, la clase de razonamiento que lleva a cabo una persona cuando necesita tomar una decisión a partir de información donde se tiene incertidumbre; es decir, cuando hay información imprecisa. Gutiérrez y Ferreira (2020) definen la LD de la siguiente forma:

La teoría de los conjuntos borrosos es un acercamiento entre la precisión de las matemáticas clásicas y la imprecisión del mundo real, el cual se le ha intentado ajustar a modelos matemáticos que no permiten lo borroso o lo impreciso, y como consecuencias nos lleva a resultados indeseables (p. 1).

Actualmente, cuando se trata de tomar una decisión al medir el nivel de desempeño de una competencia desarrollada por un estudiante, se utilizan las rúbricas, uno de los instrumentos de medición más aceptados y utilizados por los docentes.

En esencia, las rúbricas son tablas o matrices de doble entrada fáciles de diseñar y aplicar, que integran la evaluación cualitativa, al establecer los criterios de cada rubro a evaluar, con la evaluación cuantitativa al asignar niveles de desempeño en escala para cada rubro antes descrito (Cano, 2015, Gatica & Uribarren, 2012, Martínez, 2008, Picón, 2013).

El objetivo de la presente investigación es comparar las aproximaciones en el nivel de desempeño de una competencia usando tres instrumentos de medición, dos basados en rúbrica y otro basado en lógica difusa. El estudio no se centró en el proceso de evaluación, sino en los resultados que arrojaron los tres instrumentos de medición diseñados y validados, para poder determinar cuál de ellos brindó mayor sensación de justicia en la asignatura de Probabilidad y estadística impartida en el sexto semestre del ciclo escolar febrero-julio 2017 en el Centro de Estudios Tecnológicos industrial y de servicios CETis 108.

Revisión de la literatura

Barlybayev, Sharipbay, Ulyukova, Sabyrov y Kuzenbayev (2016) presentaron un estudio desarrollado en el sector universitario que se centra en evaluar las competencias de los estudiantes en el campo disciplinar de la electrónica. Se tomaron de muestra a 13 estudiantes de ingeniería que cursaban la asignatura de Graficación en 3D, durante 15

semanas. Los aspectos para evaluar, que sirvieron de entrada al modelo difuso fueron: a) conferencias, b) clases prácticas, c) trabajo independiente y d) control final.

Esos datos que alimentaron al sistema difuso fueron tomados a fin de calcular las calificaciones mediante otros tres métodos, a saber: a) cálculo por media aritmética o promedio, b) con el cálculo usando el método seguido en las universidades de Kazajistán, c) con la metodología utilizada en los programas de maestría de la Universidad de Liverpool.

En un análisis comparativo entre los cuatro métodos de evaluación, se llevó a cabo una correlación lineal. Se correlacionaron las calificaciones generadas de cada método de evaluación con cada uno de los otros tres métodos restantes y los resultados obtenidos indicaron, mediante el coeficiente de correlación de Pearson, que sólo tuvieron una fuerte correlación lineal positiva el método de promedios con el método de LD y el método utilizado en Kasajistán con el método de LD.

Debido a que el método de LD fue el único que tuvo fuerte correlación con algún otro método de evaluación (con el de promedios y con el de Kasajistán) se sugiere que la LD debe ser aplicada en cualquier contexto, ya que es una herramienta matemática poderosa y que ha sido probada en el ámbito internacional.

Por su parte, Ivanova y Zlatanov (2019) presentaron un estudio orientado a la evaluación de cursos virtuales del idioma inglés, en el cual exaltan algunas características importantes de una justa evaluación que permita a los estudiantes comprender sus logros y desarrollar la confianza de su propia capacidad de aprendizaje y habilidades. La evaluación permite a los maestros recopilar, registrar, calificar e interpretar información sobre lo aprendido en los estudiantes.

En la búsqueda de una evaluación justa, se muestran los resultados de la evaluación a 78 estudiantes de primer grado de informática, la

evaluación constó de 60 preguntas cerradas con valor de un punto cada una, más una pregunta abierta de un valor máximo de 20 puntos.

Se definieron cinco niveles en los conjuntos difusos: reprobado, satisfactorio, bueno, muy bueno y excelente. Al usar conjuntos difusos, se evitan situaciones injustas, como asignar un nivel muy bueno a alguien 61 con puntos y un nivel bueno a alguien con 60. Así pues, se prestó mucha atención a los límites de los niveles de los estudiantes.

Los resultados demostraron que la media y la varianza cambiaron significativamente con el modelo difuso, lo que da evidencia, para concluir, que no se altera la puntuación general del grupo, aunque sí se logró tener una evaluación más justa.

Metodología

En esta sección se describe la metodología empleada en esta investigación con el propósito de comparar las aproximaciones en el nivel de desempeño de una competencia usando tres instrumentos de medición, dos basados en rúbrica y otro basado en lógica difusa.

La investigación se circunscribe desde el enfoque cuantitativo, que se inspira en el positivismo que, a su vez, es una corriente filosófica en la que, según Monje (2011), “lo que importa para el positivismo es la cuantificación, la medición” (p. 12).

En el enfoque cuantitativo, se hace uso del método hipotético deductivo; en la lógica deductiva, que va de lo general a lo particular, se parte de las leyes universales o teorías causales que expliquen la realidad que experimenta el investigador. Los estudios cuantitativos se fundamentan en estudios previos, por lo que la literatura es importante para definir teorías, hipótesis y el diseño mismo de la investigación (Del Canto & Silva, 2013, Hernández, Fernández & Baptista, 2010).

El diseño de investigación aplicado, el plan a seguir para lograr los objetivos, responder las preguntas de investigación y analizar la certeza de las hipótesis (Hernández, Fernández & Baptista, 2010) es de tipo no experimental, ya que no se pretendió tener control sobre las variables de estudio ni se tuvieron grupos de control, no se realizó por parte del investigador ninguna acción de intervención que pudiese alterar las estrategias de evaluación de los docentes; esto es debido a que la investigación no se centra en el proceso de evaluación, se enfoca en los resultados que arrojan los instrumentos de medición en función del modelo matemático que utilizan (variable dependiente e independiente, respectivamente).

La investigación se divide en dos principales fases, la primera de ellas describe cómo se elaboraron y validaron los tres instrumentos de medición a comparar. En esta fase, se utilizó la técnica de recolección de datos, la entrevista abierta o no estructurada a fin de abstraer la base de conocimientos del experto acerca de su concepción de la competencia a evaluar, y plasmar esa base de conocimientos desde la perspectiva de las rúbricas y desde la perspectiva de la lógica difusa.

En el presente estudio, se eligió la competencia argumenta la solución obtenida de un problema, con métodos numéricos, gráficos, analíticos o variacionales, mediante el lenguaje verbal, matemático y el uso de las tecnologías de la información y la comunicación. Ésta se dividió conceptualmente en tres dimensiones: a) teoría, b) práctica y c) actitud, y a su vez en seis atributos, los cuales se muestran en la Tabla 7.1.

Tabla 7.1*Funciones de membresía de entrada del modelo difuso.*

Atributo	Dimensión
Lleva los conocimientos teóricos a la práctica.	Teoría
Organiza, clasifica, presenta e interpreta información numérica en forma de tablas y gráficas.	Práctica
Calcula e interpreta índices estadísticos como medidas de tendencia central y medidas de dispersión.	Práctica
Presenta los resultados de los índices estadísticos con honestidad, pulcritud de los datos y exactitud.	Actitud
Utiliza las TIC como medio para gestionar su aprendizaje.	Práctica
Cumple con sus actividades con honestidad y puntualidad.	Actitud

Dado que el verbo que define la competencia es argumentar y, a su vez, éste se encuentra en el nivel taxonómico de orden superior, el logro de la competencia por parte de los estudiantes se llevó a cabo de forma gradual con actividades que reforzaran cada uno de los seis atributos descritos. Los atributos por sí mismos tienen un nivel taxonómico de orden inferior que la competencia; sin embargo, al trabajarse de manera conjunta y repetitiva, esto condujo a que los estudiantes fuesen capaces de argumentar sobre los resultados estadísticos al término del semestre, que fue en ese momento en el que se aplicaron los tres instrumentos a comparar.

El modelo matemático empleado con el primer instrumento de rúbrica fue el de promedio o media aritmética: $\text{Calificación} = (\sum_{i=1}^n \text{Atributo}_i) / n$

El segundo instrumento de medición fue elaborado con un modelo matemático basado en LD tipo Mandami. La recolección y el análisis de información en esta etapa ocurrieron simultáneamente, para este fin se utilizó la herramienta Fuzzy Logic Toolbox de Matlab R2013a. El uso de esta herramienta brindó validez al proceso interpretativo, pues se podía ir modelando gráficamente, con exactitud y en tiempo real (durante las

propias entrevistas) la concepción matemática del docente experto. Con el empleo de Fuzzy Logic Toolbox, se modelaron las funciones de membresía de las entradas/atributos y de la salida calificación. Las funciones se detallan en la Tablas 7.2 y 7.3, respectivamente.

Tabla 7.2

Funciones de membresía de entrada del modelo difuso.

Entradas/atributos	Funciones de membresía
1. Lleva los conocimientos teóricos a la práctica. Teoría.	Bajo: $y = \text{trimf}(x, [0 \ 0 \ 4])$ Medio: $y = \text{trapmf}(x, [3 \ 5 \ 6 \ 6.5])$ Alto: $y = \text{gaussmf}(x, [0.927 \ 8.85])$
2. Organiza, clasifica, presenta e interpreta información numérica en forma de tablas y gráficas. Práctica. 4. Presenta los resultados de los índices estadísticos con honestidad, pulcritud de los datos y exactitud. Actitud. 6. Cumple con sus actividades con honestidad y puntualidad. Actitud.	Bajo: $\text{trimf}(x, [0 \ 0 \ 3])$ Regular: $\text{trimf}(x, [1 \ 3 \ 5])$ Bueno: $\text{gbellmf}(x, [1.05 \ 1.656 \ 6.169])$ Alto: $y = \text{gaussmf}(x, [1 \ 9])$
3. Calcula e interpreta índices estadísticos como medidas de tendencia central y medidas de dispersión. Práctica. 5. Utiliza las tic como medio para gestionar su aprendizaje. Práctica.	Bajo: $y = \text{trimf}(x, [0 \ 0 \ 3])$ Medio: $y = \text{trapmf}(x, [2 \ 5 \ 6 \ 8])$ Alto: $y = \text{gaussmf}(x, [1.699 \ 10])$

Tabla 7.3

Funciones de membresía de salida del modelo difuso.

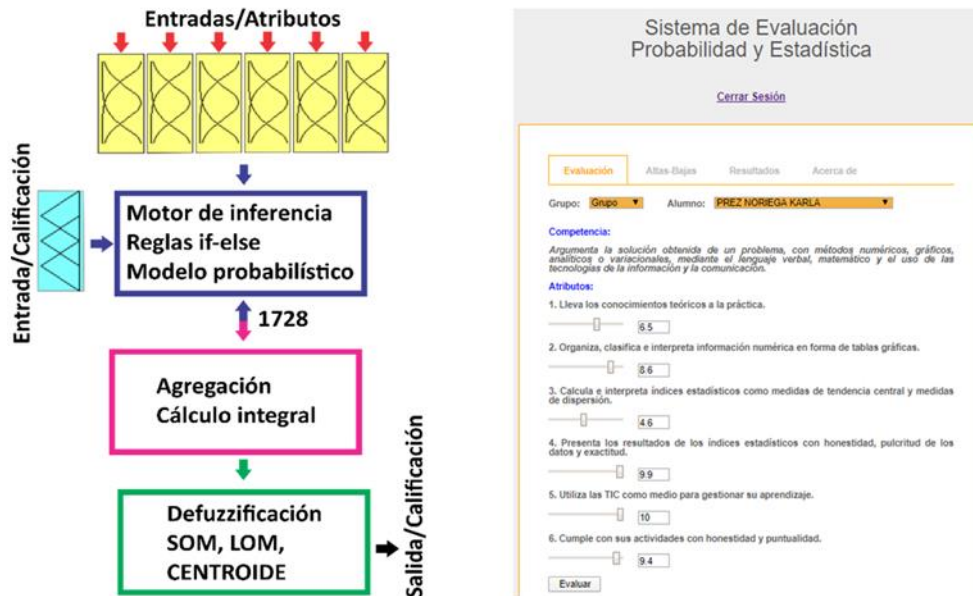
Salida	Funciones de membresía
Calificación	Bajo: $y = \text{gaussmf}(x, [1.293 \ 0])$ Regular: $y = \text{gaussmf}(x, [0.9373 \ 3.72])$ Bueno: $y = \text{gaussmf}(x, [0.7958 \ 7.66])$ Alto: $y = \text{gaussmf}(x, [0.9772 \ 9])$

Para el cálculo de las reglas de inferencia, se utilizó un modelo probabilístico de combinaciones, por regla de multiplicación, se obtuvieron un total de 1 728 reglas al multiplicar $3 \times 4 \times 3 \times 4 \times 3 \times 4$. Esto debido a que los atributos 1, 3 y 5 tienen tres niveles (bajo, medio y alto); los atributos 2, 4 y 6 tienen cuatro niveles (bajo, regular, bueno y alto). En el modelo matemático, se utilizaron los métodos de defuzzificación SOM (Smallest Of Maximum), LOM (Largest Of Maximum) y centroide.

El instrumento basado en LD se desarrolló con HTML5, CSS, JavaScript, PHP y MySQL, el diagrama a bloques del modelo matemático empleado y la interfaz gráfica se pueden observar en la Figura 7.1.

Figura 7.1

Estructura del sistema difuso e interfaz gráfica (Web).



Se empleó un tercer instrumento de medición basado en rúbrica con un modelo matemático por regla de tres, dándole a cada atributo distintas ponderaciones o pesos de 10%, 20%, 15%, 20%, 15% y 20% en ese orden para cada uno de los seis atributos, el modelo matemático es el siguiente:

$$\text{calificación} = \frac{\sum_{i=1}^n \text{Puntos obtenidos en el atributo}_i * \text{Porcentaje de ponderación en el atributo}_i}{\text{Puntaje máximo en el atributo}}$$

Las ponderaciones se eligieron de esa manera en un intento de acercar el modelo de regla de tres al modelo difuso en el siguiente sentido: a los atributos que resultaron con más niveles en las funciones de membresía se les asignó mayor peso, 20 por ciento.

La segunda fase describe los pasos seguidos en el diseño metodológico no experimental, en el cual se empleó el instrumento modelado con LD para la recolección de datos, a manera de cuestionario autoadministrado con preguntas cerradas. La recolección se realizó en un

solo momento, lo que lo hace un diseño de tipo transversal, al concluir el semestre febrero-julio 2017, ya que para ese momento los estudiantes habrían realizado múltiples desempeños y generado evidencias suficientes con lo que los docentes podrían establecer un juicio de valor para cada uno de ellos.

Las hipótesis de investigación planteadas son las siguientes.

Hipótesis de diferencia entre grupos.

H_i: 80% o más de las mediciones calculadas con ambas rúbricas presentan diferencias significativas (más de una unidad), en escala de cero a diez.

H_i: 20% o más de las mediciones calculadas con el modelo basado en LD presentan diferencias significativas (más de una unidad) con referencia a las dos rúbricas, en escala de cero a diez.

Hipótesis estadística: ANOVA.

H₀: No hay diferencia en las medias de las mediciones calculadas con los tres instrumentos. H₀: $\mu_1 = \mu_2 = \mu_3$

H₁: Existe diferencia en las medias de las mediciones calculadas con los tres instrumentos. H₁: $(\mu_1 \neq \mu_2 \neq \mu_3)$

Hipótesis de diferencia entre grupos.

H_i: El modelo basado en LD por ser más complejo e imitar el razonamiento humano es menos predecible que los modelos por promedio y regla de tres.

A fin de agregar validez al estudio, se muestran los tres grupos que tenía asignado el docente experto de quien se extrajo la base de conocimientos, más otros tres grupos de otro docente que también impartió la asignatura de Probabilidad y estadística; de esta forma, se

tendrían dos opiniones referentes a la buena o mala aproximación de la calificación de los instrumentos.

En total se mostraron seis grupos de nueve posibles, sumando un total de 143 estudiantes de un universo de 217, esto equivale a una muestra de 65.9% de todos los estudiantes matriculados en esa asignatura. La Tabla 7.4 presenta un concentrado de los grupos, así como la cantidad de alumnos que los conformaban.

Tabla 7.4

Población de grupos de Probabilidad y Estadística del semestre febrero-julio 2017.

Docente	Grupo	Hombres	Mujeres	Grupal (total)
No muestreado	6AM-CT	10	17	27
Apoyo	6BM-CT	5	16	21
Apoyo	6AV-CT	10	10	20
Experto	6AM-EL	12	2	14
No muestreado	6AV-EL	14	0	14
No muestreado	6AM-MT	32	1	33
Experto	6AM-PR	16	7	23
Experto	6AV-PR	16	9	25
Apoyo	6AM-SM	19	21	40
	Suma	134	83	217

En la Tabla 7.4, se observan un total de 143 (78 hombres y 65 mujeres) alumnos muestreados de 217 posibles, lo que equivale a 65.9% del alumnado, por la cantidad puede ser considerada una muestra bastante representativa.

Se utilizó, como instrumento de recolección de datos, el software diseñado con lógica difusa, dado que éste contaba con una interfaz gráfica que permitía dar de alta los grupos y sus estudiantes.

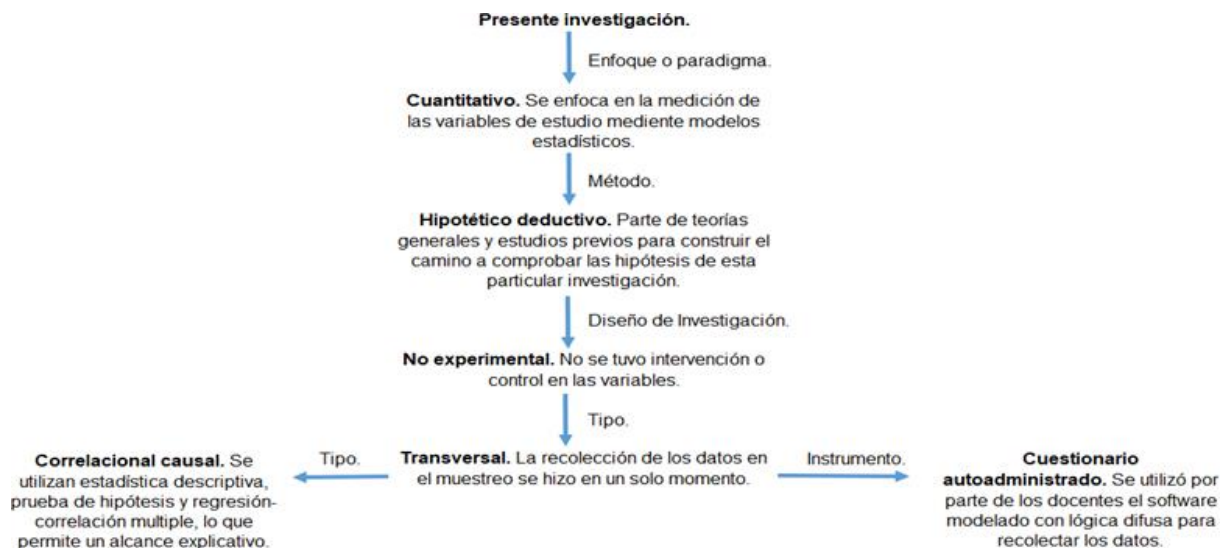
Una vez que la base de datos contenía los datos de alta de todos los estudiantes, los dos docentes procedieron a ingresar para cada uno de sus

alumnos, un valor de cero a diez como nivel de desempeño en cada uno de los seis atributos. El software calculó automáticamente la calificación resultante por estudiante. Posteriormente, se pudieron realizar consultas grupales.

Se aplicaron, en el diseño metodológico, estadística descriptiva, prueba de hipótesis ANOVA con test de Tukey, regresión-correlación múltiple, con la intención de comprender más a fondo el comportamiento de los modelos matemáticos empleados en las rúbricas y el modelo difuso, y lograr que la investigación tenga un alcance de carácter explicativo. En la Figura 7.2 se muestra el diseño metodológico, el camino elegido para lograr los objetivos generales y específico de la investigación.

Figura 7.2

Diseño de investigación empleado en la investigación.



Resultados

Una primera rúbrica se modeló matemáticamente con un promedio; una segunda rúbrica con la regla de tres, con las siguientes ponderaciones: 10%, 20%, 15%, 20%, 15% y 20% en ese orden para cada uno de los seis atributos; el instrumento restante se modeló con LD con intención de imitar el razonamiento humano.

En la Tabla 7.5 se encuentran las medidas de tendencia central y de dispersión de las calificaciones arrojadas por los distintos instrumentos de medición a fin de comprender la distribución de los datos en cada uno de ellos.

Tabla 7.5

Estadísticos de las calificaciones obtenidas de los instrumentos de medición.

Instrumento	Media	Desviación estándar	Mediana	Moda	N para moda
Modelo difuso	7.457	2.253	7.8	7.8	33
Rúbrica con promedio	6.907	2.227	7.2	6.7	7
Rúbrica con regla de tres	6.921	2.239	7.2	6.8	6

Con los datos que arrojó el software Minitab en la tabla 5 se puede observar que la media o promedio no cambia significativamente (una unidad o más) si se mide con un instrumento u otro; es decir, el promedio grupal de calificaciones no se ve afectado con demasía al utilizar uno u otro instrumento. La desviación estándar es muy cercana entre los tres instrumentos, esto sugiere que la dispersión de los datos es muy similar, lo que hace pensar que los instrumentos mantienen aproximadamente la misma precisión; sin embargo, al poderse clasificar en varias muestras, tres en total (una por cada instrumento) la desviación estándar puede ser engañosa, ya que no permite conocer la precisión de un instrumento (muestra) respecto a los otros dos.

Es por esta razón que se calculó el coeficiente de variación $CV = \text{desviación estándar}/\text{media}$, que es una medida de dispersión relativa, porque mide la precisión de los datos con respecto a la media, y al ser un coeficiente, el resultado se interpreta a manera porcentual, los resultados se aprecian en la Tabla 7.6.

Tabla 7.6

Coeficiente de variación entre los tres instrumentos de medición.

Instrumento	Media	Desviación estándar		Coeficiente de variación CV	
Modelo difuso	7.457	2.253	0.302132	=	30.21 %
Rúbrica con promedio	6.907	2.227	0.322427	=	32.24 %
Rúbrica con regla de tres	6.921	2.239	0.323508	=	32.35 %

Aunque la desviación estándar sugiere que los tres instrumentos tienen casi la misma precisión, siendo el modelo difuso el que muestra menos, pues la desviación es ligeramente más alta, el coeficiente de variación revela que, al compararse de manera porcentual, la precisión realmente de los tres instrumentos es el modelo difuso el que muestra mayor precisión de los tres. Esto porque conforme más se acerca el coeficiente a 100% más imprecisión hay entre los datos. El modelo difuso aumenta la precisión en los resultados en 2% respecto a las dos rúbricas, a su vez, las dos rúbricas tienen prácticamente la misma precisión.

Después, al ordenar las calificaciones de los instrumentos de mayor a menor o viceversa, la calificación que queda en el centro de esos datos y que, a su vez, divide las observaciones en dos grupos de igual tamaño, es la mediana. Se aprecia en la tabla 5 que la mediana en ambas rúbricas es 7.2 y para el modelo difuso 7.8; por tanto, tampoco varía significativamente (una unidad o más) de un instrumento a otro. Significa que, en los resultados de las rúbricas, la mitad de las calificaciones están por debajo de 7.2 y la otra mitad de las calificaciones está por encima de ese mismo valor.

En el modelo difuso, la mitad está por debajo de 7.8 y la otra mitad por encima de él.

El estadístico descriptivo que sufre un abrupto contraste es la moda, siendo este parámetro la calificación que más se repite. Para las rúbricas, la moda es muy similar en valor y frecuencia: 6.7 y 6.8 con frecuencia de 7 y 6 repeticiones, respectivamente; sin embargo, en el modelo difuso, la moda se aleja de estos valores significativamente (una unidad o más) resultando 7.8 y con una frecuencia de repetición de 33, significa que de 143 estudiantes muestreados 33 de ellos obtuvieron 7.8 de calificación con este instrumento, lo que equivale a 23 por ciento.

Los datos arrojados con los tres instrumentos de evaluación arrojaron una distribución normal, los resultados de cada modelo matemático se observan en la tabla 7 (Anexo A).

La primera hipótesis de investigación H_1 : 80% o más de las mediciones calculadas con ambas rúbricas no presentan diferencias significativas (más de una unidad), en escala del cero al diez, puede aceptarse como válida, ya que 100% de las calificaciones calculadas con una rúbrica u otra no sufrieron cambios significativos, esto quiere decir que utilizar la rúbrica por promedios o bien asignar las distintas ponderaciones arrojarán prácticamente el mismo resultado sin diferencia significativa en una unidad o más.

Respecto a la segunda hipótesis de investigación (H_2): 20% o más de las mediciones calculadas con el modelo basado en LD presentan diferencias significativas (más de una unidad) con referencia a las dos rúbricas, en escala del cero al diez, ésta también se toma como válida.

Realizado un conteo se obtuvo que 42 calificaciones cambiaron significativamente (una unidad o más) del modelo difuso a la rúbrica por promedio, equivale a 29.4% y 43 calificaciones cambiaron significativamente del modelo difuso a la rúbrica por regla de tres, lo que corresponde a 30 por ciento.

Buscando tener más certeza sobre estos porcentajes de cambio, se aplicó en Minitab intervalos de confianza para la proporción, resultando los valores de la tabla 8 en los que se visualiza claramente que los intervalos sugieren que más de 22% de los resultados tendrán cambios significativos.

Tabla 7.8

Intervalos de confianza para la proporción de cambios significativos entre instrumentos.

Diferencias	N	Evento	Muestra p	IC de 95% para p
Modelo difuso y rúbrica con promedio	143	42	0.293706	(0.220572, 0.375605)
Modelo difuso y rúbrica con regla de tres	143	43	0.300699	(0.226916, 0.382947)

Si bien entre los estadísticos descriptivos que se muestran en la tabla 9 se observa que la media de la muestra no cambió significativamente al usar cualquiera de los tres instrumentos, aun así para darle mayor validez a este supuesto, se analizó mediante una prueba de hipótesis ANOVA.

Tabla 7.9

Resultados del análisis de varianza ANOVA.

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Instrumentos	2	29.49	14.743	2.94	0.054
Error	426	2136.68	5.016	-----	--
Total	428	2166.16	-----	-----	--

De todos esos valores de la tabla 9 el que interesa es el valor p, utilizado en la decisión de aceptar o rechazar las hipótesis estadísticas.

H0: No hay diferencia en las medias de las mediciones calculadas con los tres instrumentos. $H_0: \mu_1 = \mu_2 = \mu_3$

H1: Existe diferencia en las medias de las mediciones calculadas con los tres instrumentos.

$$H1: \mu_1 \neq \mu_2 \neq \mu_3$$

Debido a que el valor p es mayor al nivel de significancia de 5% (0.050), se encuentra evidencia suficiente para aceptar H0 como válida, se afirma, entonces, que al medir con cualquiera de los tres instrumentos no cambia significativamente el promedio general en los grupos en donde se utilicen. El test de Tukey agregado al ANOVA arrojó más evidencia como se puede observar en la Tabla 7.10.

Tabla 7.10

Resultados del test de Tukey.

Instrumento	N	Media	Agrupación
Difuso	143	7.457	A
Regla	143	6.903	A
Promedio	143	6.898	A

Como se aprecia, los tres instrumentos pertenecen a un mismo grupo A. Esto se interpreta, en este caso, de igual manera que ANOVA, entre los tres grupos no hay evidencia de que exista un cambio significativo en la media.

Con las pruebas ANOVA y Tukey, se tiene mayor evidencia para afirmar que, aunque se use el modelo difuso o las rúbricas, el promedio general de todas las calificaciones que estos instrumentos generen no cambiará significativamente.

A fin de comprender más el comportamiento de los modelos matemáticos que implementa cada uno de los tres instrumentos de medición utilizados, se calculó la ecuación de regresión múltiple y el coeficiente de correlación múltiple en Minitab, los resultados se presentan en la tabla 11. En este análisis, se busca conocer un modelo predictivo de la calificación, además de qué tanta correlación guardan los seis atributos evaluados con la calificación que emiten los instrumentos.

Tabla 7.11

Ecuación de regresión y coeficiente de correlación múltiple para los instrumentos de medición.

Instrumento	Ecuación de regresión (modelo predictivo)	R-cuad. (grado de correlación)
Modelo difuso	Calificación = 0.903 – 0.2278 Atrib. 1 + 0.4865 Atrib. 2 + 0.1863 Atrib. 3 + 0.0779 Atrib. 4 + 0.1513 Atrib. 5 + 0.2688 Atrib. 6	91.86%
Rúbrica con promedio	Calificación = 0.00038 + 0.16825 Atrib. 1 + 0.17103 Atrib. 2 + 0.16338 Atrib. 3 + 0.16361 Atrib. 4 + 0.16549 Atrib. 5 + 0.16959 Atrib. 6	99.98%
Rúbrica con regla de tres	Calificación = 0.02885 + 0.09723 Atrib. 1 + 0.19744 Atrib. 2 + 0.14820 Atrib. 3 + 0.20436 Atrib. 4 + 0.15094 Atrib. 5 + 0.20021 Atrib. 6	99.99%

Al calcular la calificación de cada instrumento con su respectiva ecuación de regresión y compararse este resultado contra la calificación real que arrojó el instrumento, se tiene que en el modelo difuso sólo 11 de las 143 calificaciones calculadas con la ecuación de regresión fueron iguales a las que arrojó el modelo difuso.

Comparando las calificaciones que arrojó la rúbrica por promedios contra las calificaciones que arrojó su ecuación de regresión, 141 de las 143 calificaciones resultaron iguales. En el caso de la rúbrica por regla de tres, fueron 138 de las 143 calificaciones resultaron iguales.

Esto no es de extrañarse, ya que como se observó en la tabla 11, el coeficiente R-cuadrado es para las rúbricas muy cercano a 100 por ciento.

Estos resultados dan evidencia de que el instrumento con LD al ser más complejo es menos predecible, y que en la ecuación de regresión los atributos sólo pueden explicar 91.86% de la correlación.

Las ecuaciones y los coeficientes de correlación brindan soporte a la decisión de aceptar como válida la hipótesis de investigación.

Hi: El modelo basado en LD por ser más complejo e imitar el razonamiento humano es menos predecible que los modelos por promedio y regla de tres.

Las ecuaciones de regresión múltiple y los coeficientes de correlación múltiple aplicados a los instrumentos ponen de manifiesto que el modelo matemático empleado en el instrumento con lógica difusa, al ser más complejo, es menos predecible. Esto cobra relevancia cuando se recuerda que el modelo imita el razonamiento humano en la toma de decisiones al existir incertidumbre.

Desde la perspectiva de ambos docentes, al analizar uno a uno los resultados para cada estudiante, el modelo difuso fue 95.8% más exacto, brindando, en esos casos, mayor sensación de justicia al comparar el juicio de valor con la calificación resultante. Además, recordando el coeficiente de variación expuesto en la tabla 5, el modelo difuso resultó ser también 2% más preciso que ambas rúbricas. Por otra parte, en ANOVA y test de Tukey, mostraron que, al no haber variación en las medias grupales, hay consistencia entre los tres instrumentos; por tal motivo, la validación de éstos se considera correcta.

Discusión

Existe consistencia en los datos empíricos obtenidos en la presente investigación comparados con los resultados encontrados en Ivanova y Zlatanov (2019), debido a que los resultados experimentales de la evaluación tradicional basada en promedios comparada con los resultados obtenidos mediante lógica difusa demuestran que las evaluaciones son más justas. Los investigadores consideran que la evaluación es importante, porque permite monitorear el progreso del estudiante y también la

evaluación brinda a los docentes la oportunidad de obtener información para mejorar su propia práctica docente.

Conclusiones

El modelo difuso, por su naturaleza compleja, es un modelo matemático menos predecible que el promedio y regla de tres. Existen algunos aspectos no conocidos de incertidumbre que influyen en la calificación que éste arroja; por tal motivo, las entradas al modelo (calificaciones de los seis atributos) sólo pueden explicar 91% de la calificación final. En las rúbricas, las mismas entradas explican 99% de la calificación final.

El modelo difuso fue más efectivo en más de 95% de los casos según la percepción de ambos docentes al brindar mejor sensación de justicia respecto de las rúbricas. Entonces, el modelo difuso resultó ser más exacto; de tal forma, si, por ejemplo, el modelo arrojó una calificación de 8.3, ésta representa 83% de grado de desarrollo de la competencia.

El modelo difuso fue 95% más efectivo que las rúbricas dado que este modelo matemático permitió emular el razonamiento de los docentes al momento de emitir un juicio de valor, se considera, entonces, que el modelo difuso es tanto un instrumento de evaluación (razona y evalúa mediante un cerebro artificial) y es también un instrumento de medición, a diferencia de las rúbricas que son instrumentos de medición.

Se puede concluir que las rúbricas tienen un modelo matemático en donde la medición es puntual (discreta), a base de escalas y, en el modelo difuso, la medición es continua; esto en la manera en que se procesan las variables. Además, en el modelo difuso, los cálculos se procesan más integralmente.

Referencias

- Barlybayev, A., Sharipbay, A., Ulyukova, G., Sabyrov, T., & Kuzenbayev, B. (2016). Student's performance evaluation by fuzzy logic. *Procedia Computer Science*, 102, 98-105.
- Biggs, J. (2006). *Calidad del aprendizaje universitario*. Madrid, España: NARCEA.
- Cano, E. (2015). Las rúbricas como instrumento de evaluación de competencias en educación superior: ¿uso o abuso? *Profesorado. Revista de Currículum y Formación de Profesorado*, 19(2), 265-280.
- Del Canto, Ero, & Silva Silva, Alicia (2013). Metodología cuantitativa: abordaje desde la complementariedad en ciencias sociales. *Revista de Ciencias Sociales (Cr)*, III(141), 25-34.
- Gatica, F., & Uribarren, T. (2012). ¿Cómo elaborar una rúbrica? *Investigación en Educ. Médica*, 2(1), 61-62.
- Gutiérrez, A. & Ferreira, W. (2020). Un modelo de regresión lineal aplicando lógica difusa. *Revista Sextante*, 23, 58-64.
- Hernández, R., Fernández, C., & Baptista, M. (2010). *Metodología de la investigación*. 5ª. ed. México: McGraw Hill.
- Ivanova, Vanya, & Zlatanov, Boyan (2019). Application of fuzzy logic in online test evaluation in english as a foreign language at university level. *AIP Conference Proceedings* 2172, 040009 (2019). Recuperado de <https://doi.org/10.1063/1.5133519>
- Martínez, J. (2008). Las rúbricas en la evaluación escolar: su construcción y uso. *Avances en Medición*, 6, 129-138.
- Monje, C. (2011). *Metodología de la investigación cuantitativa y cualitativa*. Guía didáctica. Neiva: Universidad Sur Colombiana.

Picón Jácome, É. (2013). La rúbrica y la justicia en la evaluación. Íkala, Revista de Lenguaje y Cultura, 18(3), 79-94.