ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA MATEMÁTICA

ANÁLISIS Y PROGRAMACIÓN DE LA SIMILARIDAD DE LERMAN ENTRE VARIABLES BINARIAS

Trabajo de Integración Curricular

Tipo: Proyecto de Investigación

Presentado para optar al grado académico de:

MATEMÁTICA

AUTOR: ANABEL DEJANEIRA CÓRDOVA RUIZ

DIRECTOR: Dr. RUBÉN PAZMIÑO MAJI, PhD

Riobamba – Ecuador

2023

Yo, Anabel Dejaneira Córdova Ruiz, declaro que el presente Trabajo de Integración Curricular es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autora asumo la responsabilidad legal y académica de los contenidos de este Trabajo de Integración Curricular; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 28 de Abril de 2023

Anabel Dejaneira Córdova Ruiz

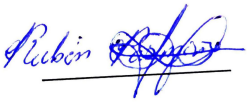
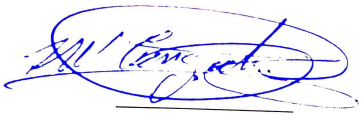
0604610006

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA MATEMÁTICA

El Tribunal del Trabajo de Integración Curricular certifica que: El Trabajo de Integración Curricular; Tipo: Proyecto de Investigación, **ANÁLISIS Y PROGRAMACIÓN DE LA SIMILARIDAD DE LERMAN ENTRE VARIABLES BINARIAS**, realizado por la señorita: **ANABEL DEJANEIRA CÓRDOVA RUIZ**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Integración Curricular, el mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal autoriza su presentación.

	FIRMA	**FECHA**
Ing. María de Lourdes Palacios Robalino **PRESIDENTE DEL TRIBUNAL**		2023-04-28
Dr. Rubén Antonio Pazmiño Maji, PhD **DIRECTOR DEL TRABAJO DE INTEGRACIÓN CURRICULAR**		2023-04-28
Dr. Jorge Washington Congacha Aushay, MSc. **ASESOR DEL TRABAJO DE INTEGRACIÓN CURRICULAR**		2023-04-28

DEDICATORIA

Dedido este trabajo a mis queridos padres, Fausto Córdova y Margarita Ruiz por apoyarme desde el principio y confiar en el potencial de su hija, les agradezco por cuidarme todo el tiempo. A mis hermanos por estar siempre al pendiente de mi y brindarme un consejo. A mis sobrinos por ser la alegría de mi vida y en especial quiero dedicarle a mi pequeña Eliana Salomé por ser mi motor y alegría durante toda la carrera. A mi amiga que ahora es un angelito del cielo, Andrea Nata, estoy segura que desde su cielo bonito me ha guiado y me ha cuidado todo este tiempo, todo esto también lo he logrado pensando en ti. A mi mejor amigo, Bryan Ilbay, has sido mi apoyo desde el día uno en esta carrera, gracias por estar conmigo en cada lucha que hemos tenido en la carrera y ahora vernos cumplir este sueño juntos se siente bien.

Anabel

AGRADECIMIENTO

Agradezco a Dios por todas las bendiciones y las fuerzas que me han dado para culminar esta linda etapa. A mis padres por cada esfuerzo, por todo el apoyo y confianza hacia su pequeña hija para que pueda culminar la carrera universitaria. A todos mis maestros que fueron parte de mi vida universitaria y en especial a mi director de tesis Dr. Rubén Pazmiño Maji, PhD por bridarme su tiempo y sus conocimientos para culminar este trabajo. Agradezco a la ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO por abrirme sus puertas y continuar con mis estudios para convertirme en una profesional de la carrera de mis sueños. Agradezco a todas las personas que formaron parte de mi vida universitaria, pues cada una me dejó una enseñanza, un momento inolvidable, una amistad que llevaré en mi corazón por el resto de mi vida.

Anabel

ÍNDICE DE CONTENIDOS

CAPÍTULO IV

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

RESUMEN

El Análisis estadístico implicativo nació hace más de 40 años, el cual contiene métodos como la llamada similaridad de Lerman que trabaja con variables binarias y fue automatizado en el programa informático CHIC y posteriormente se creó un paquete en R (RCHIC), pero, el Análisis estadístico implicativo (ASI) no cuenta con una programación en el software Matlab que ayude a los ingenieros y matemáticos con el análisis de la gran cantidad de datos que se generan en la actualidad, por lo tanto, el objetivo de esta investigación fue programar en el software Matlab la similaridad de Lerman entre variables binarias y validarla mediante similaridad con el paquete RCHIC. La metodología implementada fue de tipo cuantitativo, se utilizó un diseño pre-experimental de la forma RGXO1 de tipo correlacional puesto que la variable independiente no se la manipula debido a que ya existe una programación con la cual comparar; el colectivo de estudio fue de 100 000 bases aleatorias de datos binarios y de ahí se sacó una muestra de 383 bases aleatorias mediante un muestreo aleatorio simple, las cuales fueron analizadas en el programa hecho en Matlab. Mediante esta metodología se logró la programación de la similaridad de Lerman en el software Matlab y se analizó todas las bases de datos binarios realizando una comparación entre las matrices de similaridad y los nodos significativos. En conclusión, la programación hecha en Matlab tuvo un noventa y cinco por ciento de similaridad con el paquete RCHIC, es por eso que se propone que en estudios futuros se mejore su programación.

Palabras clave: <ANÁLISIS ESTADÍSTICO IMPLICATIVO>, <SIMILARIDAD DE LERMAN>, <PROGRAMACIÓN>, <SOFTWARE>, <DATOS BINARIOS>.

SUMMARY/ABSTRACT

The Implicative Statistical Analysis started more than 40 years ago, which contains methods such as the so-called Lerman similarity that works with binary variables and was automated in the CHIC computer program; then, a package was created in R (RCHIC), but the Implicative Statistical Analysis (ASI) does not have a programming in the Matlab software to help engineers and mathematicians with the analysis of the large amount of data that is generated nowadays. Therefore, the aim of this research was to program the Lerman similarity between binary variables in Matlab and validate it through the similarity existing with the RCHIC package. The methodology implemented was quantitative and a RGXO1 correlational type, pre-experimental design was used, since the independent variable is not manipulated due to the existence of a programming with which it is compared; the study group consisted of 100,000 random bases of binary data, from which a sample of 383 bases was selected by random sampling and analyzed in the Matlab program. By means of this methodology, it was possible to obtain the Lerman similarity programming in the Matlab software; then, all the binary databases were analyzed by making a comparison between the similarity matrices and the significant nodes. In conclusion, the programming carried out in Matlab obtained a ninety-five percent similarity with the RCHIC package, so it is recommended to improve its programming for future studies.

Keywords:<STATISTICAL IMPLICATIVE ANALYSIS>, <LERMAN SIMILARITY>, <PROGRAMMING>, <SOFTWARE>, <BINARY DATA>.

Lic. Paul Rolando Armas Pesántez Mgs.

0603289877

INTRODUCCIÓN

Dentro de la didáctica de la matemática junto con la estadística se desarrolla una herramienta de la Minería de Datos llamada Análisis Estadístico Implicativo que ha sido usado exitosamente para determinar la cuasi-implicación. Contempla la estructuración de datos, interrelacionados a sujetos y variables, y la extracción de reglas inductivas entre las variables. El Análisis Estadístico Implicativo intenta cuantificar cuán probable es que suceda la variable "b" si se ha observado la variable "a". La herramienta se conoce como ASI (*Analyse Statistique Implicative*) fue creado por Régis Gras y surgió hace 43 años, pero ha empezado a sobresalir en los últimos 10 años. Además, permite conocer posibles relaciones de similitud, implicación y cohesión entre los datos (Gras,1951) .

El ASI usa las definiciones de similaridad, cohesion y cuasi.impliación. La herramienta utilizada para automatizar el ASI es el software CHIC donde se crea un paquete RCHIC que contiene funciones como: árbol de similitud, gráfico implicativo, y árbol de cohesión. De particular importancia, aquí hablaremos sobre la similaridad de Lerman que muestra una relación de simetría entre variables o sujetos de una determinada población. Para automatizar la similaridad de Lerman contaremos con ayuda de software Matlab. (Valls, 2014) .

Así pues, el Software Matlab tiene una gran capacidad para trabajar con el lenguaje basado en matrices, con esto buscamos lograr el desarrollo de un pseudocódigo que realice, de manera automática, las relaciones de la similaridad de Lerman entre los datos. Logrando así, que el Software Matlab se convierta en un instrumento más para el ASI. Y por otro lado, nadie ha intentado programar la similaridad de Lerman entre variables binarias en MATLAB.

CAPÍTULO I

1. PROBLEMA DE INVESTIGACIÓN

1.1. Planteamiento del problema

La similaridad de Lerman es uno de los procedimientos utilizados en el ASI (Análisis Estadístico Implicativo) Y fue creada por el matemático Israel Cesar Lerman. El ASI es un método de análisis de datos que trabaja con variables binarias y fue automatizado en el programa informático CHIC y posteriormente se creó un paquete en R (RCHIC). Estos dos software calculan los índices de proximidad de Lerman (Similaridad de Lerman) y además muestra un gráfico (árbol de similaridad) de los mismos. Entonces se pretende realizar una programación similar a como trabajan CHIC Y RCHIC, en el software Matlab.

Actualmente, sabemos que existe una gran cantidad de datos para ser analizados por ende las personas quienes trabajan con estos datos buscan una técnica fácil, eficiente y rápida para analizarlos de forma automática. La herramienta ASI (Análisis Estadístico Implicativo) puede trabajar con datos binarios y a su vez con otro tipo de datos y al ser automatizado en dos software como son Chic y R nos brinda una ventaja al momento de analizar los datos, pero no todos están familiarizados con estos dos software, por otro lado, para muchos se les hace más fácil manejar el Software Matlab ya que se lo ha venido usando por muchos años pero en Matlab no se encuentra automatizado el ASI por lo que muchos prefieren buscar otro método de análisis de datos binarios. Así que en este trabajo pretendemos presentar la programación de la Similaridad de Lerman que viene siendo una parte del ASI y que tenga una similaridad del 70 %.

En este proyecto de investigación sólo vamos a analizar y programar la Similaridad de Lerman que es parte del ASI. Para saber si la programación funciona de manera correcta para cualquier rango de variables se va a trabajar con un colectivo de estudio conformado por 100 000 bases aleatorias de datos binarios. Y solamente se hará el análisis con datos binarios pues se pretende hacer la programación exclusivamente para que ese tipo de datos. Los datos estarán en archivos con extensión CSV debido a que como nos basaremos en la programación de RChic pues este trabaja con archivos de ese tipo de extensión y como el Software Matlab puede leer tranquilamente este tipo de archivos entonces mejoraría la experiencia de los usuarios al momento de usar la programaciń.

1.2. Objetivos

1.2.1. *Objetivo general*

Programar en el Software Matlab la similaridad de Lerman entre variables binarias y validarla mediante similaridad con RCHIC.

1.2.2. *Objetivos específicos*

- Analizar la similaridad de Lerman.
- Determinar la formulación matemática de la similaridad de Lerman.
- Elaborar el seudocódigo de la similaridad de Lerman.
- Programar la similaridad de Lerman.
- Estudiar la similaridad entre los resultados del programa MATLAB (mCHIC) y el paquete RCHIC.

1.3. Justificación

El presente proyecto de investigación está enfocado a presentar la formulación matemática de la similaridad de Lerman y también su programación en el Software Matlab por lo que va acorde con todo lo que se ha aprendido en la carrera de Matemática ya que dentro de la carrera se estudia la materia de Software Matemático, se estudian aplicaciones, también se adquiere conocimiento básico sobre los datos binarios, permitiéndonos así tener unas bases para realizar la investigación.

El objetivo es realizar una programación de la similaridad de Lerman en Matlab por lo que en este trabajo de investigación se pretende realizar una comparación entre los resultados del código hecho en Matlab con los del programa RCHIC tal como: matriz de similaridad, dendograma, copresencias, nodos significativos, nombre del archivo, número de filas que conforman la base de datos, número de columnas que conforman la base de datos, el total de datos.

Por lo que se va a usar un tipo de diseño pre-experimental y un tipo de estudio transversal, es decir, se analizarán los datos de un cierto colectivo de estudio y de este se sacará una muestra que nos ayudará a saber si la programación está bien hecha y sirve para que varios usarios la puedan usar.

CAPÍTULO II

2. MARCO TEÓRICO

2.1. Antecedentes de investigación

2.1.1. *Análisis estadístico implicativo*

El Análisis Estadístico Implicativo mejor conocido por sus siglas ASI (Analyse Statistique Implicative) tiene su origen en Francia, fue creado por el matemático Regis Gras en 1980. Gras trabajó en la minería de datos y utilizó el método ASI para modelizar la cuasi-implicación, dicho de otro modo, deseaba obtener la probabilidad cuantitativa de que suceda una variable "b" si se observa una variable "a" en una población dada E, en donde las variables a y b están definidas (Valls, 2014).

Estamos concientes de que la realidad está formada por hechos y fenómenos que producen relaciones causales, y el pensamiento humano suele captarlos de manera inmediata y los expresa en oraciones formalizadas usando la lógica, y a su vez, estos se convierten en implicaciones porque se relaciona una causa con un efecto. Su representación matemática es: $a \Rightarrow b$, donde a es la causa y b el efecto y se lo enuncia "Si a entonces b".

- Si $a \Rightarrow b$ es verdadera entonces "a es condición suficiente para b"
- Si $b \Rightarrow a$ es verdadera entonces "a es condición necesaria para b"
- Si $a \Leftrightarrow b$ es verdadera entonces "a es condición suficiente y necesaria para b"

En matemática, estas reglas son conocidas como Teoremas puesto que se cumple que toda causa tiene su consecuencia, pero en situaciones reales la regla de causa-efecto tiene sus excepciones conocidas, en matemática, como contradicciones. Las contradicciones motivaron a analizar de mejor manera a los datos y encontrar reglas (Teoremas) en las cuales confiar y así es como surge la Cuasi-Implicación tal que acepta la afirmación "Si ocurre a entonces, generalmente, ocurre b". Para que la regla (Teorema) sea válida dependerá de la fuerza de la Cuasi-Implicación (Zamora, 2009).

Al estudiar la Cuasi-Implicación como objeto matemático dentro del campo de las probabilidades y la estadística, da paso a la construcción de herramientas teóricas, donde se fundamenta el ASI, permitiendo identificar una posible relación causal y formulando hipótesis que describen o predicen.

De forma matemática, Grass enunció al ASI como: Dadas dos variables binarias a y $b \in E$, ¿Cuál es la probabilidad cuantitativa de que suceda una variable b si se observa una variable a? o mejor dicho ¿Es verdad que $a \Rightarrow b$?. El ASI puede trabajar con distintos tipos de variables: binarias, modales, de frecuencia y de intervalo. Este trabajo se va a centrar en el tipo de variables binarias por lo que los hechos solo pueden tomar valores de 1 para afirmar y de 0 para negar. Donde a,$b : E \rightarrow \{0,1\}$, son dos variables binarias arbitrarias. $A = \{x/x \in E, a(x) = 1\}$, $B = \{x/x \in E, b(x) = 1\}$ son conjuntos de valores de verdad. Entonces el ASI trata de medir el grado de verdad de la implicación (Gras, & Kuntz, 2009).

Sean $A_i, A_j \in E$, las variables $a_i \in E$, $a_j \in E$ y la regla $a_i \longrightarrow a_i$. Lo que hace el ASI es comparar el número $n_{a_i \bigcap \overline{a_j}}$ de contraejemplos observados en $A_i \bigcap \overline{A_j}$ con el número $n_{x_i \bigcap \overline{x_j}}$ de contraejemplos obtenidos en una extracción aleatoria de los subconjuntos X_i y $X_j \in E$. También se considera que dado el conjunto E con n individuos y un conjunto $A = \{a_1, a_2, \ldots, a_p\}$ con p características, se supone que $A_i = \{x \in E / a_i(x) = 1\}$. Por lo tanto, se considera la población E con $Card(E) = n$, el conjunto A_i con $Card(A_i) = n_{A_i}$ y el conjunto $A_i \bigcap \overline{A_j}$ con $Card(A_i \bigcap \overline{A_j}) = n_{A_i \bigcap \overline{A_j}}$ (Gras, Suzuki, & Spagnolo, 2009).

Tiempo después, Regis Gras conoce al matemático Israel Cesar Lerman, quien le comenta que ha estado trabajando en su método de análisis de similaridad. Este método era diferente a los varios métodos que se usaban en ese tiempo para poder medir las similaridades. Regis Gras, Israel Cesar Lerman y Rostam, en 1981, deciden trabajar en conjunto y publican un artículo que explica las bases del análisis implicativo para datos binarios. Es aquí donde ya se tenían dos procesos dentro del ASI. Luego se implementó la definición del análisis de cohesiones, es decir, se quería analizar las implicaciones entre implicaciones (Zamora, 2009).

El ASI consta de tres procesos diferentes e independientes, uno de esos procesos se denomina el Análisis de similaridad o clasificación que es una forma de medir distancias.

2.1.2. *La similaridad de Lerman*

La similaridad de Lerman forma parte del Análisis Estadístico Implicativo (ASI), fue creada por el matemático Israel Cesar Lerman. La similaridad se define como una medida de semejanza entre los datos que van a ser agrupados.

El análisis clasificatorio es una técnica de clustering que busca dentro de un conjunto de datos agrupaciones de objetos similares. Se basa en el número de co-presencias (ocurrencia de dos o mas casos juntos en el mismo lugar) (a,b) que existen en el conjunto de datos formando una relación de similaridad. La técnica de clustering es usada para explorar dentro de un conjunto de datos y agruparlos en cluster o grupos más pequeños y que tengan objetos o individuos con características similares. El concepto central del clustering y en el que se basa todo es la llamada similaridad entre los objetos que se están agrupando (Zamora, 2009).

La variable de interés ahora es $Card(X_i \cap X_j)$, la cantidad de individuos que poseen característica a_i y la característica a_j al mismo tiempo.

El análisis clasificatorio fue desarrollado en el sistema infomático CHIC, siendo este creado presisamente para el estudio y aplicación del ASI. En 1990, CHIC resultó ser una herramienta confiable para trabajar la similaridad de Lerman. Actualmente se conoce que Raphael Couturier programó la similidaridad de Lerman, en su totalidad, en C++ (Lenguaje de programación).La construcción de un criterio de agrupamiento depende de la naturaleza de los datos y de la estructura matemática retenida para su representación (Zamora, 2009).

Para formar el árbol de similaridad se calculan los índices de próximidad que definió Isreal Cesar Lerman

2.1.3. *Nodos significativos*

Los nodos significativos representados en el dendograma o mejor conocido como árbol de similaridad, son aquellos valores máximos que se encuentran en cada nivel de similaridad, en sí, son los valores con más similaridad posible (Gras, & Kuntz, 2009).

Definición 2.1. Se llama preorden inicial y global Ω sobre $A \times A$, al preorden inducido por la aplicación S (Similaridad) sobre $A \times A$.

$$G_S(\Omega) = \{(a,b);(c,d) : S(a,b) < S(c,d)\}$$

Sea $S\Pi_k$ el conjunto de pares separados al nivel k y $R\Pi_k$ el conjunto de pares que ya se han reunido hasta este nivel k. $G_S(\Omega) \bigcap [S\Pi_k \times R\Pi_k]$ está formado por los pares de parejas que al nivel k respetan el preorden inicial. Por ejemplo, si se tiene: $S(e,f) < S(a,b)$, entonces $((e,f);(a,b)) \in G_S(\Omega)$ y si

al nivel k, e y f están aún separados, mientras que a y b se reúnen en la clase formada, la pareja $((e,f);(a,b)) \in G_S(\Omega) \bigcap [S\Pi_k \times R\Pi_k]$ (Zamora, 2009).

Ahora, el cardinal de este último conjunto es función de este nivel k, y es un indicador del acuerdo entre el preorden inicial Ω y el preorden Π_k inducido (Zamora, 2009).

Al cardinal de $G_S(\Omega) \bigcap [S\Pi_k \times R\Pi_k]$ se le asocia el índice aleatorio $G_S(\Omega^*) \bigcap [S\Pi_k \times R\Pi_k]$, donde Ω^* es un preorden aleatorio en general, provisto de una probabilidad uniforme, de todos los preordenes del mismo tipo cardinal que Ω (Zamora, 2009).

Este índice tiene:

- **Por esperanza:** $\frac{1}{2} s_k r_k$.
- **Por varianza:** $\frac{s_k r_k (s_k + r_k + 1)}{12}$

Siendo: $Card[S\Pi_k] = s_k$ y $Card[R\Pi_k] = r_k$. El índice centrado se define como:

$$S(\Omega,k) = \frac{Card[G_S(\Omega) \bigcap [S\Pi_k \times R\Pi_k]] - \frac{1}{2} s_k r_k}{\sqrt{\frac{s_k r_k (s_k + r_k + 1)}{12}}}$$

Este índice sirve de **estadística global de los niveles**. Sus variaciones son consideradas para significar la constitución de un nivel significativo, ya que este **indicador de acuerdo** entre el preorden inicial y el definido por la división Π_k se vuelve máximo cuando se alcanza un determinado acuerdo, acuerdo que no puede ser sino provisional a causa de la evolución ascendente de la jerarquía (Zamora, 2009).

Definición 2.2. Se llama **nivel significativo** a todo nivel que corresponde a un máximo local de $S(\Omega,k)$ durante la construcción de la jerarquía. Se puede decir que la división Π_k esta en resonancia parcial con Ω. Si, además, $G(\Omega) \bigcap [S\Pi_k \times R\Pi_k] = [S\Pi_k \times R\Pi_k]$, se dirá que la división Π_k esta en una rasonancia total con Ω (Gras, & Kuntz, 2009).

Se llama **nodo significativo** cualquier nodo formado a un nivel que corresponde a un máximo local de $v(\Omega,k)$, donde:

$v(\Omega,k) = S(\Omega,k) - S(\Omega,k-1)$.

2.1.4. *Análisis comparativo con otros métodos (no ASI)*

Medidas de similaridad para variables binarias

Cuando hablamos de variables binarias nos referimos a aquellas variables que pueden tomar dos valores 0 y 1, que pueden expresar la ausencia o presencia respectivamente o puede mostrar valores como: blanco o negro, si o no, hombre o mujer, verdadero o falso, etc.Para comprender de mejor manera los tipos de medidas de similaridad, se realizarún pequeño ejemplo donde a dos personas se les hace la siguiente pregunta: ¿Posee o no estos aparatos en su hogar?

Las respuestas afirmativas tendrán un valor de 1 y las respuestas negativas un valor de 0.

	x_1	x_2	x_3	x_4	x_5
i	1	1	0	0	1
i_1	0	1	0	1	1

Al relacionar los valores se puede formar una tabla de contingencia de 2x2

i / i_1	1	0
height1	a=2	b=1
0	c=1	d=1

Donde:

1. a representa el número de individuos que toman el valor 1 en cada variable.
2. b indica el número de individuos de la muestra que toman el valor 1 en la variable i y 0 en la i_1 .
3. c es el núumero de individuos de la muestra que toman el valor 0 en la variable i y 1 en la i_1 .
4. d representa el número de individuos que toman el valor 0 en cada variable, al mismo tiempo.

Luego podríamos obtener los totales:

1. $a+c$ muestra el número de veces que la variable i_1 toma el valor 1, independientemente del valor tomado por i.
2. $b+d$ es el número de veces que la variable i_1 toma el valor 0, independientemente del valor tomado por i.

3. $a+b$ es el número de veces que la variable i toma el valor 1, independientemente del valor tomado por i_1 .

4. $c+d$ es el número de veces que la variable i toma el valor 0, independientemente del valor tomado por i_1.

Después de haber explicado cada unas de las variables, procedemos a mostrar las medidas de similitud

- **Medida de Russell y Rao**

$$\frac{a}{a+b+c+d} = \frac{a}{m}$$

donde $m = a+b+c+d$.

Este coeficiente mide la probabilidad de que un individuo elegido al azar tenga el valor 1 en ambas variables. Notemos que este coeficiente excluye la pareja $0-0$, al contar el número de coincidencias pero no lo hace así al contar el número de posibles parejas. Asimismo, esta medida proporciona igual peso a las coincidencias y a las no coincidencias.

- **Medida de parejas simples**

$$\frac{a+d}{a+b+c+d} = \frac{a+d}{m}$$

Este coeficiente mide la probabilidad de que un individuo elegido al azar presente una coincidencia de cualquier tipo, pesando de igual forma las coincidencias y las no coincidencias.

- **Medida de Jaccard**

$$\frac{a}{a+b+c}$$

Esta medida mide la probabilidad condicionada de que un individuo elegido al azar presente un 1 en ambas variables, dado que las coincidencias del tipo $0-0$ han sido descartadas primero y por lo tanto han sido tratadas de forma irrelevante.

- **Medida de Dice**

$$\frac{2a}{2a+b+c}$$

Esta medida excluye el par $0-0$ de forma completa, pesando de forma doble las coincidencias del tipo $1-1$. Se puede ver este coeficiente como una extensión de la medida de Jaccard, aunque su sentido probabilístico se pierde.

- **Medida de Rogers-Tanimoto**

$$\frac{a+d}{a+d+2(b+c)}$$

Este coeficiente puede interpretarse como una extensión de la medida de parejas simples, pesando con el doble valor las no coincidencias.

- **Medida de Kulcynski**

$$\frac{a}{b+c}$$

Esta medida muestra el cociente entre coincidencias y no coincidencias, excluyendo los pares $0-0$.

Existen muchos tipos de medidas:

- $\frac{a+d}{b+c}$
- $\frac{a+d}{a+b+c}$
- $\frac{2a}{2(a+d)+b+c}$
- $\frac{2(a+d)}{2(a+d)+b+c}$
- $\frac{2(a+d)}{2a+b+c}$
- $\frac{a}{a+d+2(b+c)}$
- $\frac{a}{a+2(b+c)}$
- $\frac{a+d}{a+2(b+c)}$

2.1.5. *Formulación matemática*

Considerando un conjunto $E = \{i_1, i_2, \ldots, i_p\}$ formado por n individuos y un conjunto $A = \{a_1, a_2, \ldots, a_p\}$ formado por p características. Y a demás

$$A_i = \{a_i(x) = 1 : x \in I\}$$

donde a_i es una función y $a_i(x) = 1$ si x tiene la característica i, y $a_i(x) = 0$ en caso contrario. Entonces, $Card(I) = n$ y $Card(A_i) = n_{(a_i)}$. Para cada (a_i, a_j), tomando $K = card(A_i \cap A_j)$, se

define el índice de similaridad de Lerman como: $s(a_i, a_j) = Pr\left[Card(A_i \cap A_j) \leq K\right]$ (Zamora, 2009) .

La similaridad entonces se calcula en términos de una probabilidad sobre la variable aleatoria $card(A_i \cap A_j)$, y mide la cantidad de sucesos que ocurren al mismo tiempo entre a_i y a_j. El cálculo de la similaridad que hace CHIC es:

$$s(a_i, a_j) = Pr\left[\frac{Card(X_I \cap X_J) - \frac{n_{a_i} * na_j}{n}}{\sqrt{\frac{n_{a_i} * na_j}{n}}}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Kc} e^{-\frac{1}{2}x^2} dx$$

donde $\quad Kc = \dfrac{K - \frac{n_{a_i} * na_j}{n}}{\sqrt{\frac{n_{a_i} * na_j}{n}}}$

2.1.6. *Ejemplo detallado de aplicación*

Considerando una muestra aleatoria de 20 estudiantes de la carrera de Matemática de la ESPOCH, a quienes se les planteó la la pregunta ¿Le gusta esta materia de la carrera? Las asignaturas escogidas para el análisis son las siguientes encontrandose cada una con su respectiva variable: Lógica Matemática (LM), Geometría Analítica (GA), Topología (TO), Ecuaciones Diferenciales (ED), Análisisi Funcional (AF), Historia de la Matemática (HM) y Álgebra (Al).

Tabla 1-2: Matriz de datos binarios

Tabla de datos					
	LM	GA	To	ED	AF
al1	1	1	0	1	1
al2	1	1	0	0	0
al3	1	0	1	1	0
al4	0	0	1	1	1
al5	0	0	1	1	1
al6	0	1	0	1	0
al7	0	0	0	0	0
al8	0	1	0	0	1
al9	0	0	1	1	0
al10	1	1	0	0	1
al11	0	1	0	1	1
al12	1	0	0	1	1
al13	0	0	1	0	0
al14	1	1	1	1	0
al15	0	0	1	0	1
al16	1	1	1	0	0
al17	0	1	0	1	0
al18	1	1	0	1	1
al19	0	0	1	0	1
al20	0	1	1	1	1
n_{a_i}	10	9	11	12	12
n	20	20	20	20	20

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Para cada variable: $n_(LM) = 10$, $n_(GA) = 9$, $n_(TO) = 11$, $n_(ED) = 12$, $n_(AF) = 12$.

Desarrollo:

Para calcular los índices de proximidad $s(a_i, a_j)$ para cada par de variables procedemos a resolver la $Card(A_i \cap A_j)$, tomaremos como muestra el par de variables (LM,GA).

Entonces, la $Card(LM \cap GA) = 6$ es la cantidad de copresencias entre LM y GA, en otras palabras, es el valor del total de 1 que se encuentra tanto en la variable LM como en la variable GA.
Calculemos

$$Kc = \frac{K - \frac{n_{LM} * n_{GA}}{n}}{\sqrt{\frac{n_{LM} * nGA}{n}}}$$

$$Kc = \frac{Card(LM \cap GA) - \frac{n_{LM} * n_{GA}}{n}}{\sqrt{\frac{n_{LM} * nGA}{n}}}$$

$$Kc = \frac{6 - \frac{10*9}{20}}{\sqrt{\frac{10*9}{20}}}$$

$$Kc = 0,707$$

Con estos datos procedemos a calcular la similaridad de Lerman

$$s(LM, GA) = Pr\left[\frac{Card(LM \cap GA) - \frac{n_{LM} * nGA}{n}}{\sqrt{\frac{n_{LM} * nGA}{n}}}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Kc} e^{-\frac{1}{2}x^2} dx$$

Ahora, tomemos como otro ejemplo la similaridad entre las variables LM y TO.

Entonces, la $Card(LM \cap TO) = 5$ es la cantidad de copresencias entre LM y TO, en otras palabras, es el valor del total de 1 que se encuentra tanto en la variable LM como en la variable TO.
Calculemos

$$Kc = \frac{K - \frac{n_{LM} * n_{TO}}{n}}{\sqrt{\frac{n_{LM} * n_{TO}}{n}}}$$

$$Kc = \frac{Card(LM \cap TO) - \frac{n_{LM} * n_{TO}}{n}}{\sqrt{\frac{n_{LM} * n_{TO}}{n}}}$$

$$Kc = \frac{5 - \frac{10*11}{20}}{\sqrt{\frac{10*11}{20}}}$$

$$Kc = 0,213$$

Con estos datos procedemos a reemplazar y calcular la similaridad de Lerman

$$s(LM, GA) = Pr\left[\frac{Card(LM \cap GA) - \frac{n_{LM} * n_{GA}}{n}}{\sqrt{\frac{n_{LM} * n_{GA}}{n}}}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Kc} e^{-\frac{1}{2}x^2} dx$$

Tabla 2-2: Valores de copresencias e índices de similaridad

Variables	Card	kc	s
LM GA	6	0.7071	0.7602
LM TO	5	-0.2132	0.4156
LM ED	6	0	0.5
LM AF	6	0	0.5
GA TO	4	-0.4270	0.3347
GA ED	7	0.6885	0.7544
GA AF	4	-0.6025	0.2734
TO ED	7	0.1557	0.5619
TO AF	6	-0.2335	0.4077
ED AF	5	-0.8199	0.2061

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Tabla 2 podemos observar los índices de similaridad en nivel cero. Obteniendo así una matriz con los índices de similaridad a partir de todas las convinaciones de todas las variables y usando las fórmulas antes vistas.

Tabla 3-2: Matriz de similaridad al nivel 0

	LM	GA	TO	ED	AF
LM	1	0.7602	0.4156	0.5000	0.5000
GA	0.7602	1	0.3347	0.7544	0.2734
TO	0.4156	0.3347	1	0.5619	0.4077
ED	0.5000	0.7544	0.5619	1	0.2061
AF	0.5000	0.2734	0.4077	0.2061	1

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Ahora, se buscan los nuevos índices de similaridad al combinar la clase (a_i, a_j) con el mayor índice de la Tabla 3.

En el nivel uno de la jerarquia se unen las variables LM Y GA ya que tienen como indice de similaridad 0.7602 que viene siendo el numero mayor.

Entonces, calculamos usando la siguiente fórmula:

$$s((a_i,a_j),a_k) = \{Max[s(a_i,a_k);s(a_i;a_k)]\}^2$$

$$s((LM,GA),TO) = \{Max[s(LM,TO);s(GA;TO)]\}^2$$

$$s((LM,GA),TO) = \{Max[0.4156;0,3347]\}^2$$

$$s((LM,GA),TO) = (0.4156)^2$$

$$s((LM,GA),TO) = 0.1727$$

Se puede formar la matriz de primer nivel:

Tabla 4-2: Matriz de similaridad al nivel 1

	LM GA	TO	ED	AF
LM GA	1	0.1727	0.5692	0.2500
TO	0.1727	1	0.5619	0.4077
ED	0.5692	0.5619	1	0.2061
AF	0.2500	0.4077	0.2061	1

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En el nivel dos de la jerarquía se unen las variables ((LM,GA); ED) pues en la Tabla 4 se observa que el índice mayor es 0.5692 que corresponde a dichas variables.

Entonces, calculamos usando la siguiente fórmula:

$$s((a_i,a_j,a_k),a_k1) = \left\{Max[s(a_i,a_k1);s(a_j;a_k1);s(a_k;a_k1)]\right\}^3$$

$$s(((LM,GA),ED),TO) = \{Max[s((LM,GA),TO);s(ED;TO)]\}^3$$

$$s(((LM,GA),ED),TO) = \{Max[0.1727;0.5619]\}^3$$

$$s(((LM,GA),ED),TO) = (0.5619)^3$$

$$s(((LM,GA),ED),TO) = 0.1774$$

Se puede formar la matriz de segundo nivel:

Tabla 5-2: Matriz de similaridad al nivel 2

	LM GA ED	TO	AF
LM GA ED	1	0.1774	0.1250
TO	0.1774	1	0.4077
AF	0.1250	0.4077	1

Fuente: Elaboración propia.
Realizado por: Córdova, Anabel, 2023.

En el nivel tres de la jerarquía se unen las variables (AF; AF) pues en la Tabla 5 se observa que el índice mayor es 0.4077 que corresponde a dichas variables.

Entonces, calculamos usando la siguiente fórmula:

$$s((a_i,a_j,a_k),(a_l,a_m)) = \left\{Max\left[s(a_i,a_l);s(a_i;a_m);s(a_j;a_l);s(a_j;a_m);s(a_k;a_l);s(a_k;a_m)\right]\right\}^6$$

$$s((LM,GA,ED),(TO,AF)) = \{Max\left[s((LM,GA,ED),(TO));s((LM,GA,ED),(AF))\right]\}^6$$

$$s((LM,GA,ED),(TO,AF)) = \{Max\left[0.1774,0.1250\right]\}^6$$

Tabla 6-2: Matriz de similaridad al nivel 3

	LM GA ED	TO AF
LM GA ED	1	0.0177
TO AF	0.0177	1

Fuente: Elaboración propia.
Realizado por: Córdova, Anabel, 2023.

2.1.7. *La programación en MATLAB*

El nombre de MATLAB proviene de "MATrix LABoratory" es un programa donde se pueden realizar operaciones básicas como: + suma, - resta, $*$ multiplicación, división. Y también, estas mismas operaciones ya nombradas se pueden utilizar con matrices a mas de: ' traspuesta, $\wedge$ potenciación, división-izquierda, división-derecha, $.*$ producto elemento a elemento, $.\wedge$ elevar a una potencia elemento a elemento, ./ y $.\backslash$ división elemento a elemento y algunas más. Para crear una matriz en Matlab de cualquier rango, para las columnas se escriben los números separados por

un espacio o por coma y para escribir las filas se separaran por un punto y coma.
Matlab tiene un lenguaje propio de programación para realizar cálculos técnicos. Una de las ventajas que se encuentra en Matlab es que permite crear funciones propias. Luego de ejecutar las funciones se obtiene un resultado o valor numérico al cual se le puede cambiar de formatos de visualización sin que afecte su forma interna (Fernández, 2009).

Según (Fernández, 2009) considera que algunos de los formatos comunmente usados son:

- **format short**, aquel que muestra los 4 dígitos decimales después de la coma (formato que ya viene por defecto en Matlab).
- **format long**, aquel que muestra con 14 o 15 dígitos decimales después de la coma.
- **format short eng**, aquel que muestra en notación científica con 4 dígitos decimales después de la coma y un exponente de 3.
- **format long eng**, aquel que muestra en notación científica con 16 dígitos significativos después de la coma y un exponente de 3.
- **format rat**, aquel que muestra una aproximación racional.
- **format +**, aquel que muestra un signo positivo o negativo, o simplemente un espacio en blanco.

Variables

Las variables en Matlab deben ser escritas con letras minúsculas, sin espacios en blanco y se puede usar hasta máximo 63 caracteres. En primer lugar se tiene la variable **ans**, es aquella variable que almacena el último resultado calculado;**pi** tiene asignado el valor de la razón de una circunferencia con respecto a su diámetro; **inf** es la variable de infinito; **nan** es una magnitud no numérica que muestra NaN cuando no hay ningun valor; **realmin** es la variable que muestra al número real positivo más pequeño que se puede usar; **realmax** es la variable que muestra al número real positivo más grande que se puede usar (Fernández, 2009). Al momento de usar **Clear** se borran las variables del área de trabajo pero no se borra de la ventana de comandos. En caso contrario, si se usa **clc**, borra las variables de la ventana de comando pero no lo hace del área de trabajo (Fernández, 2009).

Funciones

Cuando se realiza una programación en Matlab, las funciones son un conjunton de instrucciones a las cuales se les asigna una tarea en específico. Toda función será escrita en la ventana de de trabajo o de edición de Matlab, a la función se le asigna un nombre único y este será similar al momento

de querer usarlo en cualquier parte del código que se puede estar creando. Las funciones, siempre se las podrá encontrar almacenadas en una carpeta para tener más facilidad de búsqueda (Ojeda, 2014).

En matemática, el concepto de función es bastante importanre y Matlab implementa este concepto al pie de la letra.

Se puede entender mejor el concepto y cómo trabaja Matlab con una función. (Ojeda, 2014) considera el siguiente ejemplo:
functión n=menor(a, b)
si b<a
n=b
else
n=a
end

En donde:

- **n** es la variable que va a entregar el resultado.
- **menor** es el nombre asignado a la función.
- **a,b** son los llamados parámetros donde se ingresan los datos que va a calcular la función.

Comandos de programación

- El comando **if** hace que se ejecute todas las órdenes o instrucciones que estan dentro de ese bloque siempre cuando la expresión sea verdadera, es decir, que me de como resultado un número.
- El comando **for** hace que se ejecute una orden en forma de bucle, es decir, se repite esta orden pero solo por un un determinado número de veces.
- El comando **while** hace que se ejecute todas las órdenes o instrucciones dentro de ese bloque pero en forma de bucle, es decir, se repite esta orden hasta cuando la expresión es verdadera.
- Al comando **end** se lo usa para finalizar un bloque donde se encuentran las órdenes o instrucciones.

CAPÍTULO III

3. MARCO METODOLÓGICO

Por el paradigma de investigación es de tipo cuantitativo ya que trabajaremos con bases aleatorias de datos binarios. También se trabajaran con fórmulas que nos darán como resultado un número. El tabajo de investigación se encuentra en el nivel correlacional puesto que es un tipo de método de investigación no experimental en el cual se va medir dos variables.

El tipo de diseño utilizado es pre-experimental de la forma RGXO1, donde RG es el grupo de base de datos estadísticas aleatorias obtenidas mediante un muestreo aleatorio simple. X es el tratamiento (la ejecución) y O1 es es la correlación del programa diseñado con el respetivo de RCHIC.

Como es de conocimiento pues el presente trabajo tiene un tipo de diseño pre-experimental por lo que a la variable independiente no se la va a manipular debido a que ya existe un experimento (una programación) con el cual se va a comparar solamente.

El trabajo es teórico pues se ha realizado una investigación a fondo que van a aportar a la matemática en el sentido de que al presentar los resultados, estos sean utilizados por diferentes usuarios que tienen conocimiento básico de matemáticas y en sí sobre progrmación y manejo de variables binarias.

El colectivo de estudio lo conforman las 100000 bases de datos aleatorias formadas por un máximo de 1000 observaciones y 100 variables, por la amplitud es un estudio de muestreo conformado por 383 bases de datos aleatorias binarias.

El cálculo aproximado del tamaño de la población se muestra en forma detallada en el Apéndice. Por el gran tamaño de la población, se escogió una muestra utilizando el método de muestreo aleatorio simple con parámetro de interés la media, se consideró la fórmula para el cálculo de la muestra $n = \dfrac{S^2}{\frac{E^2}{Z^2_{\frac{\alpha}{2}}} + \frac{S^2}{N}}$

Para aplicar la fórmula se utilizaron los parámetros desviación estándar $= 1$; $\alpha = 5\,\%$; $Z = 1,96$; $E = 10,01\,\%$; $N = 100000$ y se generó un tamaño de la muestra de $383,2$ que redondeado es 383.

CAPÍTULO IV

4. MARCO DE ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

Al momento de programar en Matlab se tuvo en cuenta la estructura general que tiene el programa, todas las herramientas que posee y las operaciones que es capaz de realizar. Por lo tanto, el programa nos ha permitido programar la Similaridad de Lerman para variables binarias. Así que se consiguió el objetivo principal que era obtener la programación, por otro lado también se logró cumplir la meta de presentar a las personas que trabajan con con el método ASI una nueva herramienta con la cual puedan trabajar y verificar datos.

Aquípresentaremos de forma detallada cada proceso que tuvo la programación de la Similaridad de Lerman.

Diagrama de flujo

Antes de programar se debe tener en cuenta cuáles son las operaciones y cálculos que hace la Similaridad de Lerman. Por lo que se realizó un diagrama de flujo que detalla cada una de ellas.

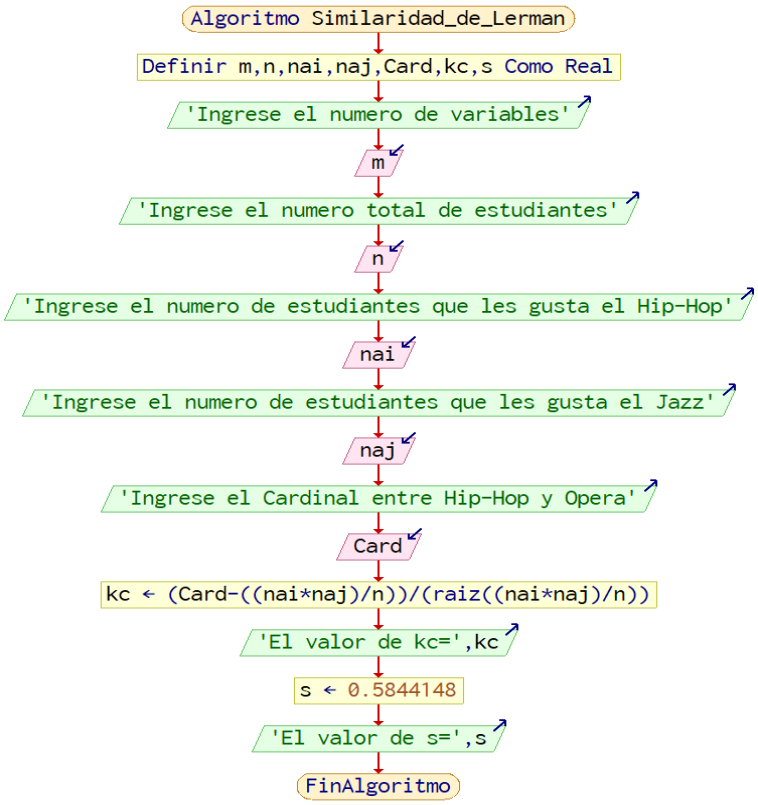


Ilustración 1-4: Diagrama de Flujo.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

- m corresponde al número de variables.
- n corresponde al número de estudiantes a quienes se les realiza la encuesta.
- n_{a_i} número de estudiantes o personas que dijeron que si les gusta el Hip-Hop , es decir tienen el valor de 1.
- n_{a_i} número de estudiantes o personas que dijeron que les gusta el Jazz, es decir tienen el valor de 1 (diferente variable).
- *Card* es el número de copresencias que se encuentran tanto es la variables de Hip-Hop como en Jazz.
- *kc* corresponde a la fómula: $Kc = \dfrac{K - \frac{n_{a_i} * na_j}{n}}{\sqrt{\frac{n_{a_i} * na_j}{n}}}$
- Finalmente, se calcula la similaridad que corresponde a la fórmula: $s(a_i, a_j) = Pr\left[\dfrac{Card(X_I \cap X_J) - \frac{n_{a_i} * na_j}{n}}{\sqrt{\frac{n_{a_i} * na_j}{n}}}\right] = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{Kc} e^{-\frac{1}{2}x^2} dx$

4.1. Procesamiento, análisis e interpretación de resultados

Toda programación tiene su procedimiento, en este capítulo presentaremos paso a paso de cómo se fue desarrollando el código de la Similaridad de Lerman en Matlab.

En primer lugar, se conoció la programación hecha en Rchic.

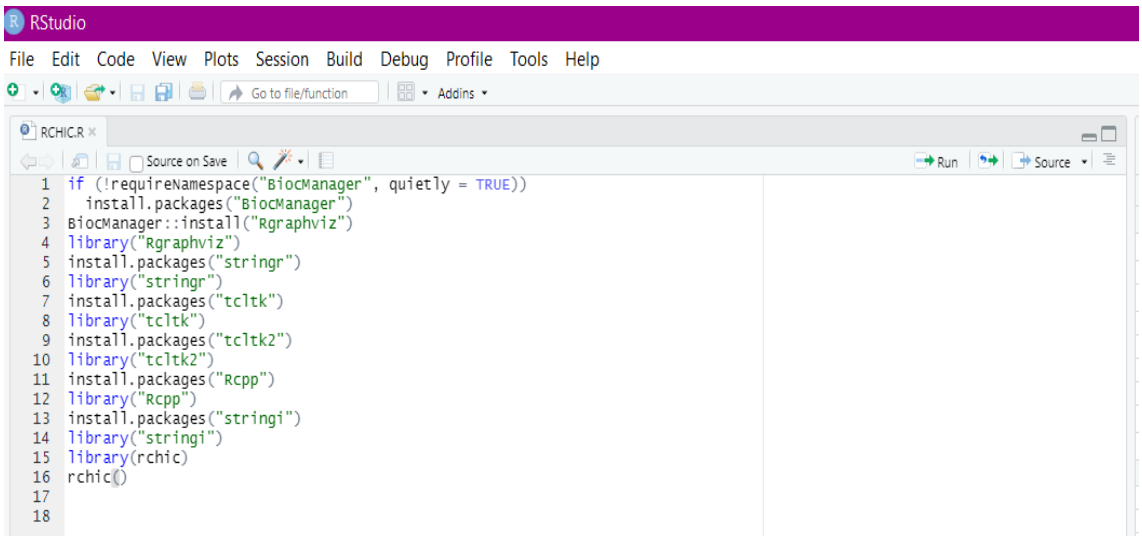


Ilustración 2-4: Paquetes en R del Análisis Estadístico Implicativo.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Se activan cada uno de los paquetes:

- BiocManager
- Rgraphviz
- stringr
- tcltk
- tcltk2
- Rcpp
- stringi
- rchic

Para obtener una nueva ventana:

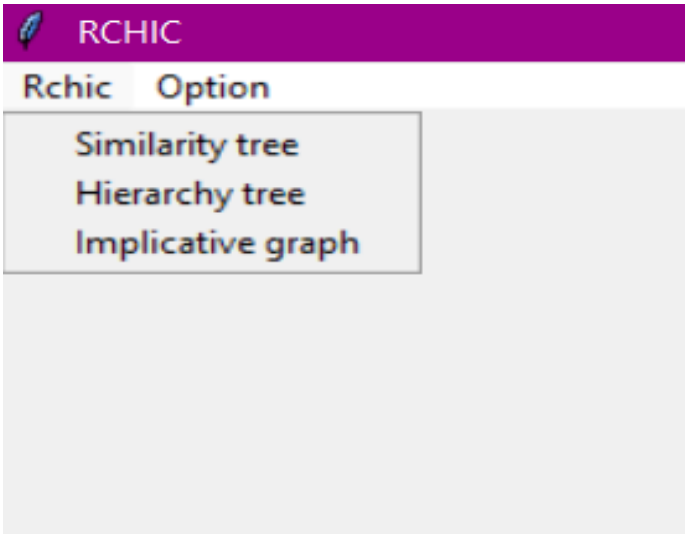


Ilustración 3-4: Ventana de RCHIC.

Fuente: Elaboración propia.

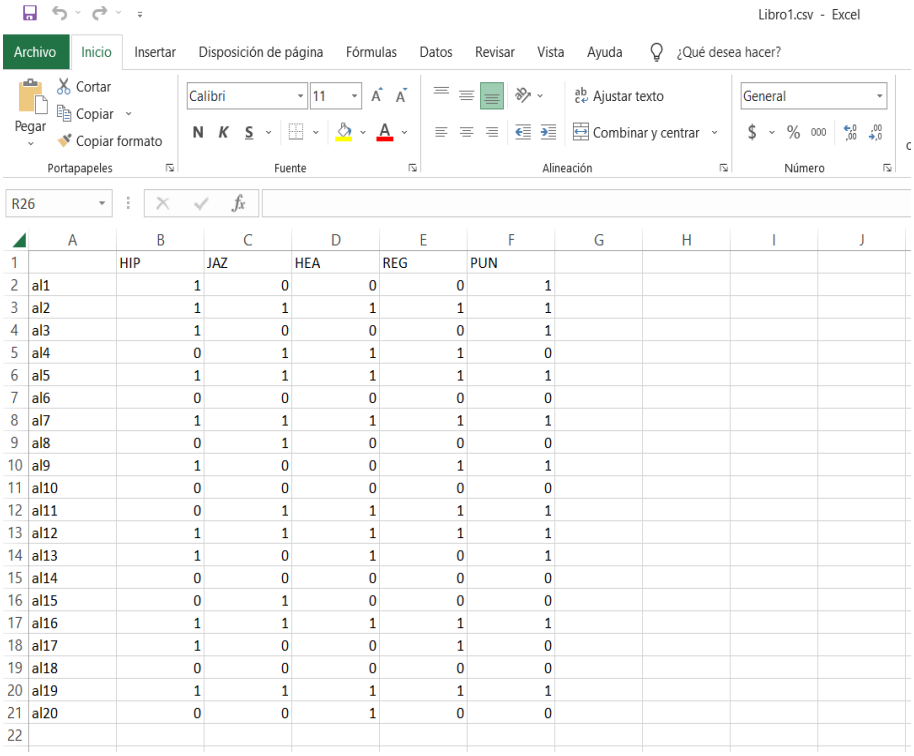
Realizado por: Córdova, Anabel, 2023.

La cual nos permite trabajar con tres opciones:

- El árbol de similaridad
- Análisis implicativo
- El árbol jerárquico

En esta investigación particularmente se va a trabajr con el árbol de similaridad. Pa comprobar su funcionamiento, escogeremos un archivo Excel con extensión ".csv"que contiene una tabla con los

datos binarios que pertenecen a la encuesta ¿Le gusta este tipo de música ? . El documento contiene 5 variables (HIP, JAZ, HEA, REG, PUN) que va a comparar.



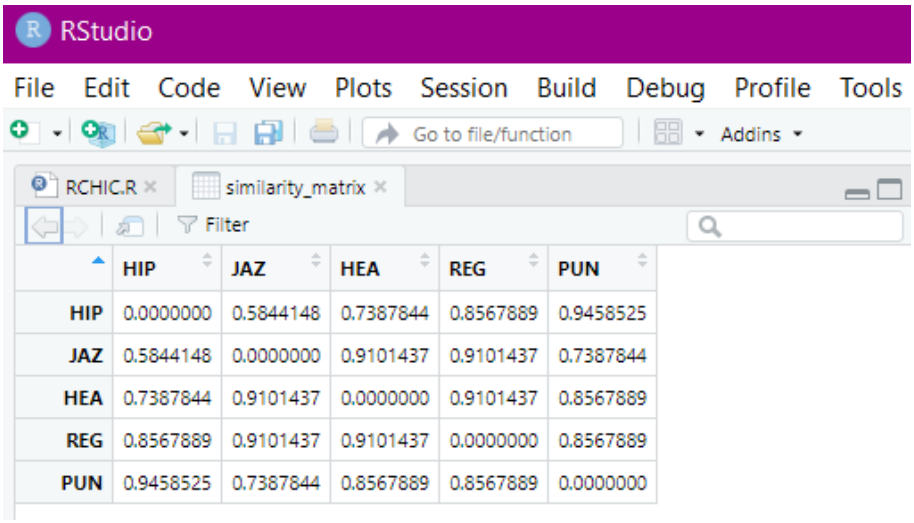
	A	B	C	D	E	F
1		HIP	JAZ	HEA	REG	PUN
2	al1	1	0	0	0	1
3	al2	1	1	1	1	1
4	al3	1	0	0	0	1
5	al4	0	1	1	1	0
6	al5	1	1	1	1	1
7	al6	0	0	0	0	0
8	al7	1	1	1	1	1
9	al8	0	1	0	0	0
10	al9	1	0	0	1	1
11	al10	0	0	0	0	0
12	al11	0	1	1	1	1
13	al12	1	1	1	1	1
14	al13	1	0	1	0	1
15	al14	0	0	0	0	0
16	al15	0	1	0	0	0
17	al16	1	1	1	1	1
18	al17	1	0	0	1	0
19	al18	0	0	0	0	0
20	al19	1	1	1	1	1
21	al20	0	0	1	0	0
22						

Ilustración 4-4: Datos en Excel.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Al ejecutar se observa que obtenemos la matriz de similaridad en RCHIC.



	HIP	JAZ	HEA	REG	PUN
HIP	0.0000000	0.5844148	0.7387844	0.8567889	0.9458525
JAZ	0.5844148	0.0000000	0.9101437	0.9101437	0.7387844
HEA	0.7387844	0.9101437	0.0000000	0.9101437	0.8567889
REG	0.8567889	0.9101437	0.9101437	0.0000000	0.8567889
PUN	0.9458525	0.7387844	0.8567889	0.8567889	0.0000000

Ilustración 5-4: Matriz de similaridad en R.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Y de igual manera el programa mostrará el árbol similaridad.

Ilustración 6-4: Árbol de similaridad.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Todo lo presentado anteriormente corresponde al paquete RCHIC creado por Raphael Couturier para R. Ahora bien, vamos a presentar el paso a paso de lo que se hizo para obtener una programación similar en el programa Matlab.

En primero lugar, usamos como base un ejemplo del archivo [**?**] el cual contenía ya todos los resultados con los cuales nos ayudaremos para ir haciendo la programación.

Pasamos la base de datos binarios de forma manual.

```
clc
X=[1 0 0 0 1; 1 1 1 1 1;1 0 0 0 1;0 1 1 1 0;1 1 1 1 1;0 0 0 0 0;1 1 1 1 1;0 0 0 0 0;1
0 0 1 1;0 0 0 0 0;0 1 1 1 1;1 1 1 1 1;1 1 1 0 1;0 0 0 0 0;0 0 0 0 0;1 1 1 1 1;1 0
0 1 0;0 0 0 0 0;1 1 1 1 1;0 1 1 0 0 ]
S= sum(X)
n=20

CardHIPPUN= 10;
CardHIPJAZ=6;
CardHIPHEA=7;
CardHIPREG=8;
CardPUNJAZ=7;
CardPUNHEA=8;
CardPUNREG=8;
CardJAZHEA=8;
CardJAZREG=8;
CardHEAREG=8;
Card=[CardHIPPUN;CardHIPJAZ;CardHIPHEA;CardHIPREG;CardPUNJAZ;CardPUNHEA;CardPUNREG;CardJAZHEA;CardJAZREG;CardHEAREG;]

Kc1=((Card(1,1))-((S(1,1)*S(1,5))/n))/(sqrt((S(1,1)*S(1,5))/n));
Kc2=((Card(2,1))-((S(1,1)*S(1,2))/n))/(sqrt((S(1,1)*S(1,2))/n));
Kc3=((Card(3,1))-((S(1,1)*S(1,3))/n))/(sqrt((S(1,1)*S(1,3))/n));
Kc4=((Card(4,1))-((S(1,1)*S(1,4))/n))/(sqrt((S(1,1)*S(1,4))/n));
Kc5=((Card(5,1))-((S(1,5)*S(1,2))/n))/(sqrt((S(1,5)*S(1,2))/n));
Kc6=((Card(6,1))-((S(1,5)*S(1,3))/n))/(sqrt((S(1,5)*S(1,3))/n));
Kc7=((Card(7,1))-((S(1,5)*S(1,4))/n))/(sqrt((S(1,5)*S(1,4))/n));
Kc8=((Card(8,1))-((S(1,2)*S(1,3))/n))/(sqrt((S(1,2)*S(1,3))/n));
Kc9=((Card(9,1))-((S(1,2)*S(1,4))/n))/(sqrt((S(1,2)*S(1,4))/n));
Kc10=((Card(10,1))-((S(1,3)*S(1,4))/n))/(sqrt((S(1,3)*S(1,4))/n));
kc=[Kc1;Kc2;Kc3;Kc4;Kc5;Kc6;Kc7;Kc8;Kc9;Kc10]
```

```
Si1 = normcdf(kc(1,1),0,1);
Si2 = normcdf(kc(2,1),0,1);
Si3 = normcdf(kc(3,1),0,1);
Si4 = normcdf(kc(4,1),0,1);
Si5 = normcdf(kc(5,1),0,1);
Si6 = normcdf(kc(6,1),0,1);
Si7 = normcdf(kc(7,1),0,1);
Si8 = normcdf(kc(8,1),0,1);
Si9 = normcdf(kc(9,1),0,1);
Si10 = normcdf(kc(10,1),0,1);
Si=[Si1;Si2;Si3;Si4;Si5;Si6;Si7;Si8;Si9;Si10;] V= 'HIPPUN';'HIPJAZ';'HIPHEA';'HIPREG';'PUNJAZ';'PUNHEA';'PUNREG';'JAZHEA';'JAZREG';'HEAREG';
T = table(V,Card,kc,Si)

simi0='HIP';'JAZ';'HEA';'REG';'PUN';
HIP=[1;Si2;Si3;Si4;Si1;];
JAZ=[Si2;1;Si8;Si9;Si5;];
HEA=[Si3;Si8;1;Si10;Si6;];
REG=[Si4;Si9;Si10;1;Si7;];
PUN=[Si1;Si5;Si6;Si7;1;];
T1 = table(simi0,HIP,JAZ,HEA,REG,PUN)
IM= max(Si)

simi1='HIP,PUN';'JAZ';'HEA';'REG';;
HIPPUNJAZ=(max(Si2,Si5))^2;
HIPPUNHEA = (max(Si3, Si6))^2;
HIPPUNREG = (max(Si4, Si7))^2;
HIPPUN = [1;HIPPUNJAZ;HIPPUNHEA;HIPPUNREG;];
JAZ1 = [HIPPUNJAZ;1;Si8;Si9;];
HEA1 = [HIPPUNHEA;Si8;1;Si10;];
REG1 = [HIPPUNREG;Si9;Si10;1;];
T1 = table(simi1,HIPPUN,JAZ1,HEA1,REG1)
A = [HIPPUNJAZ,HIPPUNHEA,HIPPUNREG,Si8,Si9,Si10];
IM1 = max(A)

simi2='HIP,PUN';'JAZ,HEA';'REG';;
a=[Si2,Si5,Si3,Si6]
HIPPUNJAZHEA=(max(a))^4;
JAZHEAREG = (max(Si9, Si10))^2;
HIPPUN1 = [1;HIPPUNJAZHEA;HIPPUNREG];
JAZHEA = [HIPPUNJAZHEA;1;JAZHEAREG];
REG2 = [HIPPUNREG;JAZHEAREG;1;];
T2 = table(simi2,HIPPUN1,JAZHEA,REG2)
B = [HIPPUNJAZHEA,HIPPUNREG,JAZHEAREG,];
IM2 = max(B)

simi3='HIP,PUN';'JAZ,HEA,REG';;
f=[Si2, Si3,Si4, Si5,Si6, Si7];
HIPPUNJAZHEAREG=(max(f))^6;
HIPPUN2 = [1;HIPPUNJAZHEAREG];
JAZHEAREG = [HIPPUNJAZHEAREG;1];
T3 = table(simi3,HIPPUN2,JAZHEAREG)


s1=19;

s2=18;

s3=16;

s4=10;

R1=1;

R2=2;

R3=4;
```

```
R4=10;

Card1=18;

Card2=29;

Card3=51;

Card4=38;

S1=(Card1-(1/2*(s1*R1)))/(sqrt((s1*R1*(s1+R2))/12));

S2=(Card2-(1/2*(s2*R2)))/(sqrt((s2*R2*(s2+R3))/12));

S3=(Card3-(1/2*(s3*R3)))/(sqrt((s3*R3*(s3+R4))/12));

S4=(Card4-(1/2*(s4*R4)))/(sqrt((s4*R4*(s4))/12));

V1=S1;

V2=S2-S1;

V3=S3-S2;

V4=S4-S3;

Nivel = '1';'2';'3';'4';

sk = [s1;s2;s3;s4];

rk = [R1;R2;R3;R4];

Card=[Card1;Card2;Card3;Card4];

S=[S1;S2;S3;S4];

V=[V1;V2;V3;V4];

T = table(Nivel,sk,rk,Card,S,V)
```

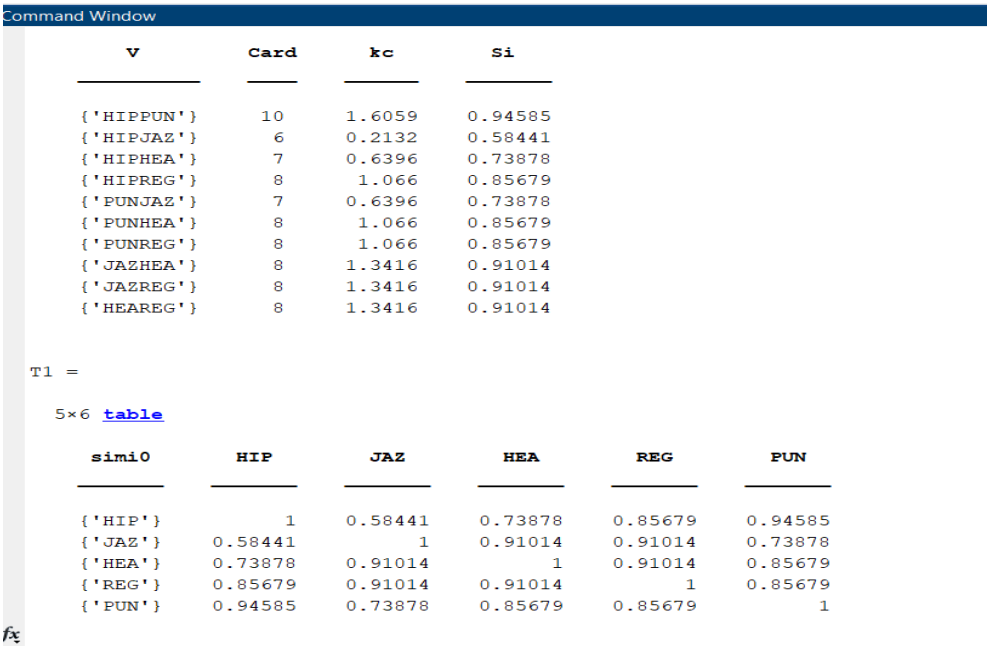


Ilustración 7-4: Primera programación en Matlab de la similaridad de Lerman.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Podemos darnos cuenta que la matriz de similaridad que nos presenta Matlab es la misma que nos presentó RCHIC anteriormente.

Esta primera programación hecha funciona de forma manual, es decir, se va pasando dato a dato, pero no es tan útil debido a que lo que el objetivo es que funcione para cualquier número de variables.

Programación del dendograma (árbol de similaridad)

```
F=[Si;IM;IM1;IM2;HIPPUNJAZHEAREG]

tree = linkage(F,'average');

dendrogram(tree)
```

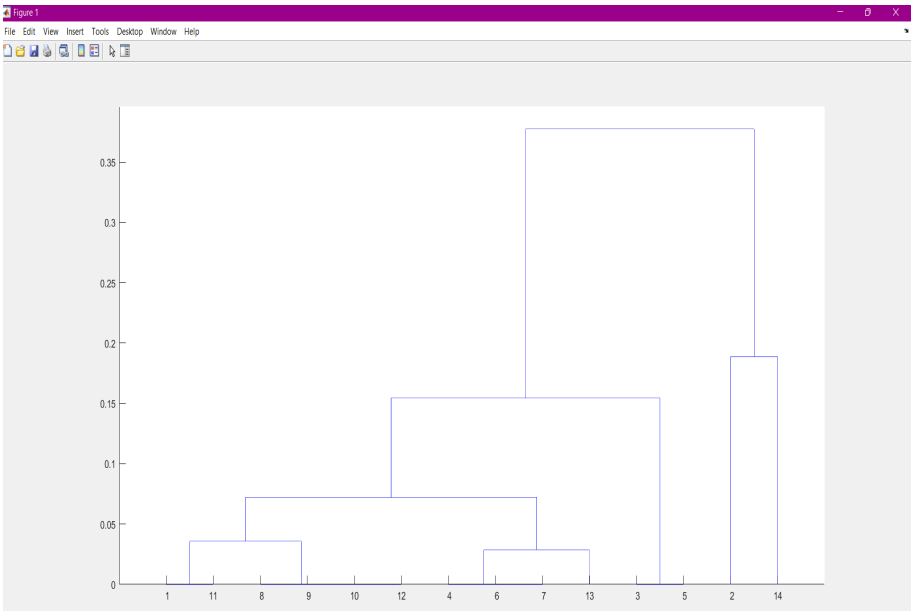


Ilustración 8-4: Dendograma (árbol de similaridad).

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Se puede observar que el dendograma de la Figura 9 tiene un parecido con el árbol de similaridad presentado por RCHIC de la Figura 6, pero no son iguales debido a que Matlab va midiendo las distancias más no la similaridad entre cada variable.

Programación para generar una base aleatoria de n variables y m filas

La segunda programación realizada logra generar una base aleatoria de datos binarios con cualquier número de variables y cualquier número de filas. Y cada una de estas bases eran calculadas y va presentado las matrices de la cardinalidad, las copresencias estandarizadas y los índices de similaridad.

clc

n=input('Ingrese el número de filas=');

m=input('Ingrese el número de columnas=');

X=randi([0,1],n,m);

X

Total de 1 en cada columna

S=sum(X)

Número total de filas y columnas
m,n

= size(X)

Cardinalidad

```
for i=[1:n]

for j=[1:n]

if i =j

CAR(i,j)= (sum(X(:,i).* X(:,j)));

end

end

end

CAR
```

copresencias estandarizadas

```
for i=[1:n]

for j=[1:n]

if i =j

Kc(i,j)=(CAR(i,j)-((S(1,i).*S(1,j))/m))/(sqrt((S(1,i).*S(1,j))/m));

end

end

end

Kc
```

índices de similaridad

```
for i=[1:n]

for j=[1:n]

if i =j

Si(i,j)=normcdf(Kc(i,j),0,1);

end

end

end

Si
```

Índice máximo de una matriz

```
M = max(Si);

M1=max(M)
```

DENDOGRAMA

```
Z = linkage(Si,'complete');

dendrogram(Z)
```

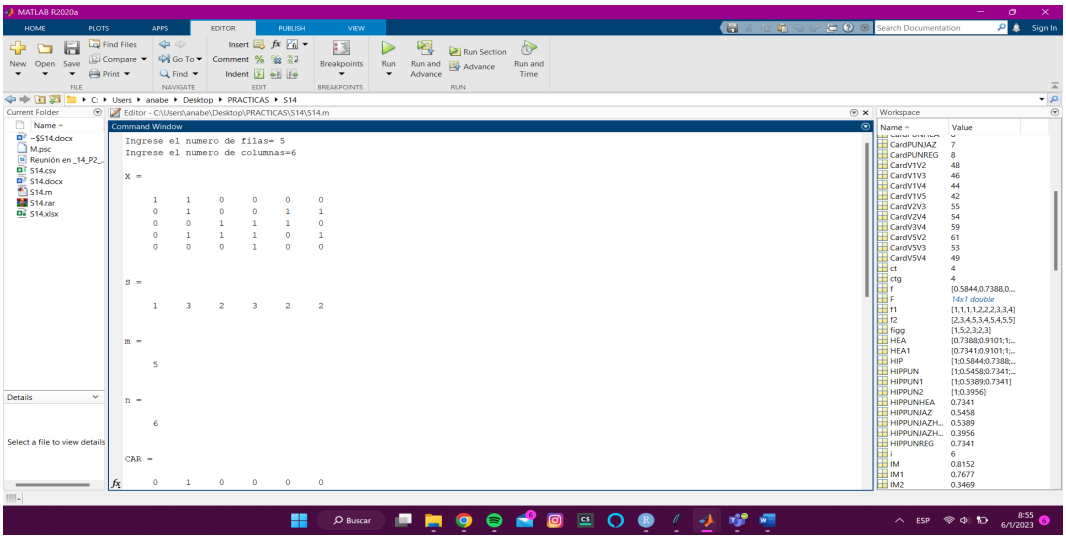


Ilustración 9-4: Segunda programación en Matlab.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

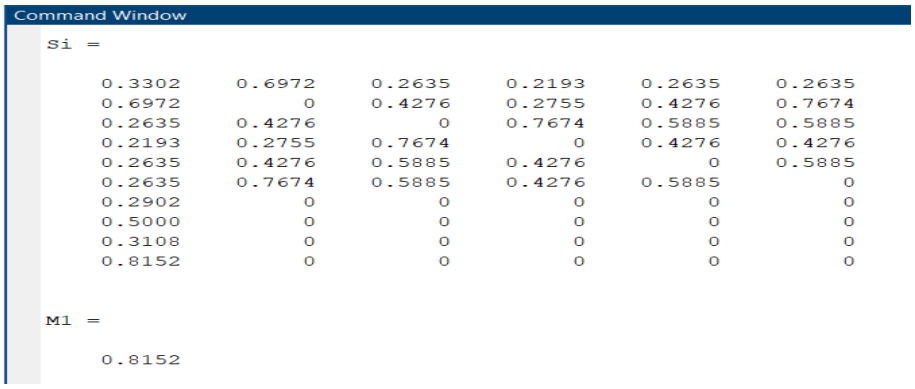


Ilustración 10-4: Segunda parte de la Segunda programación en Matlab.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Se habia logrado una parte del objetivo pero esta programación no muestra los diferentes niveles de similaridad por lo que aún no está completo.

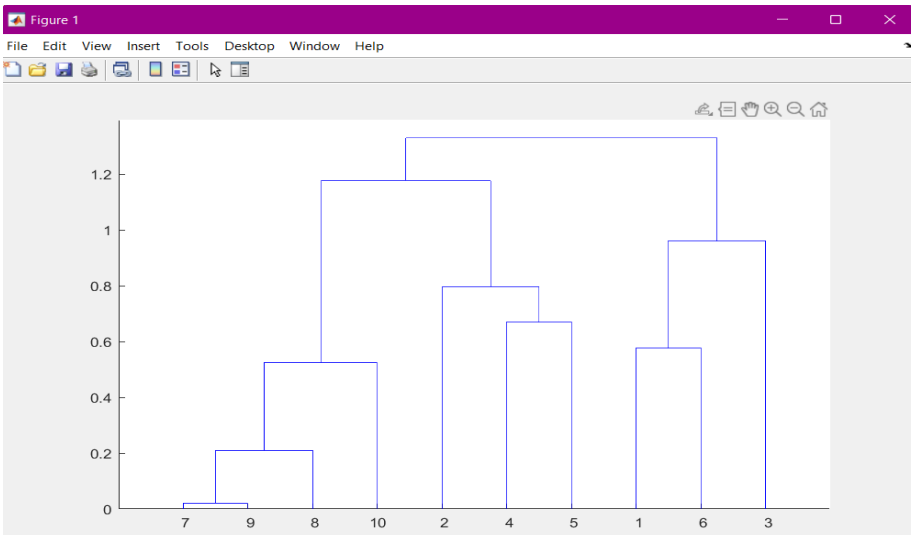


Ilustración 11-4: Dendograma segunda programación.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Observamos que el dendograma de la Figura 11 sigue midiendo las distancias más no la similaridad entre cada variable. Por lo tanto no logra aún coincidir con el gráfico que muestra RCHIC.

Programación donde se ingresen el número de interacciones

En la tercera programación hecha se puede ingresar el número de interacciones, es decir, cuantas bases con datos binarios deseamos que el programa genere. Luego de generarlas automáticamente, las bases aleatorias estarán compuesta hasta de 100 variables y 1000 filas. Cada base va a imprimirse en un archivo excel, el cual nos ayudó para comparar los resultados de este programa con los de RCHIC.

Código

```
clc Int=input('Ingresar en numero de interacciones=')
v=[1:Int];
for g=1:length(v)
X=randi([0,1],randi([1,1000]),randi([1,100]));
no=genvarname(['X',num2str(g)]);
eval([no, '=X']);
m n
=size(X)
S=sum(X)
for i=[1:n]
for j=[1:n]
if i =j
CAR(i,j)= (sum(X(:,i).* X(:,j)));
end
end
end
CAR
for i=[1:n]
```

```
for j=[1:n]
if i =j
Kc(i,j)=(CAR(i,j)-((S(1,i).*S(1,j))/m))/(sqrt((S(1,i).*S(1,j))/m));
end
end
end
Kc
for i=[1:n]
for j=[1:n]
if i =j
Si(i,j)=normcdf(Kc(i,j),0,1);
end
end
end
Si

si=genvarname(['si',num2str(g)]);
eval([si, '=Si']);

Z = linkage(Si,'complete');
ZE=genvarname(['Z',num2str(g)]);
eval([ZE, '=Z']);

end



figure F1 = dendrogram(Z1,'Orientation','top','ColorThreshold','default'); set(F1,'LineWidth',2)
figure F2 = dendrogram(Z2,'Orientation','top','ColorThreshold','default'); set(F2,'LineWidth',2)
figure F3 = dendrogram(Z3,'Orientation','top','ColorThreshold','default'); set(F3,'LineWidth',2)
figure F4 = dendrogram(Z4,'Orientation','top','ColorThreshold','default'); set(F4,'LineWidth',2)
figure F5 = dendrogram(Z5,'Orientation','top','ColorThreshold','default'); set(F5,'LineWidth',2)
figure F6 = dendrogram(Z6,'Orientation','top','ColorThreshold','default'); set(F6,'LineWidth',2)
figure F7 = dendrogram(Z7,'Orientation','top','ColorThreshold','default'); set(F7,'LineWidth',2)
figure F8 = dendrogram(Z8,'Orientation','top','ColorThreshold','default'); set(F8,'LineWidth',2)
figure F9= dendrogram(Z9,'Orientation','top','ColorThreshold','default'); set(F9,'LineWidth',2)
figure F10 = dendrogram(Z10,'Orientation','top','ColorThreshold','default'); set(F10,'LineWidth',2)


xlswrite('Datos1.xlsx',X1)

xlswrite('Datos2.xlsx',X2)

xlswrite('Datos3.xlsx',X3)

xlswrite('Datos4.xlsx',X4)

xlswrite('Datos5.xlsx',X5)

xlswrite('Datos6.xlsx',X6)

xlswrite('Datos7.xlsx',X7)

xlswrite('Datos8.xlsx',X8)

xlswrite('Dataos9.xlsx',X9)

xlswrite('Datos10.xlsx',X10)
```

Por ejemplo, al hacer correr el programa con 10 interacciones se obtienen 10 bases con datos binarios distintas con diferente número de filas y columnas, estas a su vez se guardan en la carpeta donde está guardado el archivo de la programación en archivos de Excel y de igual manera

presenta las 10 matrices con los índices de similaridad

.

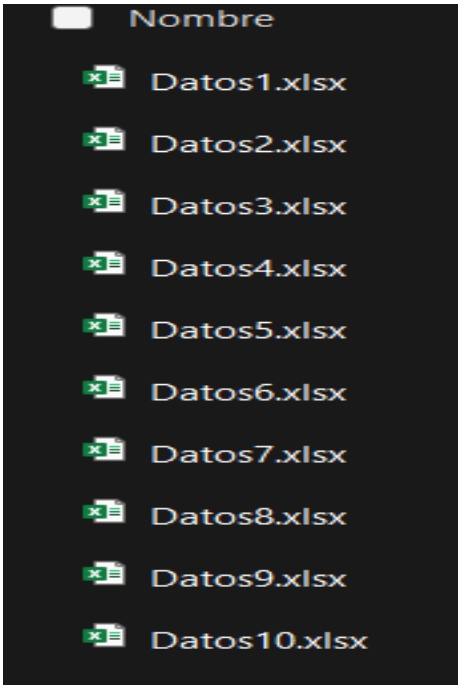


Ilustración 12-4: Bases de datos binarios creadas por Matlab.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Ahora, observemos que de la primera interacción, el programa calculo la primera matriz con los índices de similaridad, y lo mismo sucederá con las otras 9 interacciones más. Luego comparamos con RCHIC y observamos que coinciden con sus respuestas.

si1 | si2 | si3 | si4 | si5 | si6 | si7 | si8 | si9 | si10

75x75 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	0.3468	0.6236	0.3606	0.5133	0.3295	0.6847	0.5562	0.8048	0.4548	0.5590	0.5891	0.5936	0.6
2	0.3468	0	0.6778	0.3114	0.3805	0.4129	0.4836	0.5496	0.4841	0.5158	0.5485	0.3201	0.1706	0.3
3	0.6236	0.6778	0	0.6633	0.1629	0.3524	0.5597	0.7306	0.7032	0.3824	0.4782	0.2638	0.3215	0.4
4	0.3606	0.3114	0.6633	0	0.3309	0.3517	0.3517	0.2593	0.1127	0.5644	0.2134	0.4873	0.8087	0.3
5	0.5133	0.3805	0.1629	0.3309	0	0.5108	0.7761	0.3207	0.5805	0.3543	0.6541	0.4160	0.6226	0.5
6	0.3295	0.4129	0.3524	0.3517	0.5108	0	0.2575	0.6280	0.3756	0.5274	0.4954	0.5220	0.6654	0.4
7	0.6847	0.4836	0.5597	0.3517	0.7761	0.2575	0	0.3524	0.5280	0.5963	0.3596	0.3164	0.6654	0.6
8	0.5562	0.5496	0.7306	0.2593	0.3207	0.6280	0.3524	0	0.4946	0.5122	0.6075	0.5148	0.2656	0.3
9	0.8048	0.4841	0.7032	0.1127	0.5805	0.3756	0.5280	0.4946	0	0.4637	0.8981	0.6674	0.3365	0.7
10	0.4548	0.5158	0.3824	0.5644	0.3543	0.5274	0.5963	0.5122	0.4637	0	0.5731	0.6118	0.6631	0.7

Ilustración 13-4: Matriz de similaridad de la primera interacción en Matlab.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

RchicInstall.R × similarity_matrix ×

Filter Cols: « < 1 - 50 > »

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14
A1	0.0000000	0.3467908	0.6236328	0.3606353	0.51330632	0.32949209	0.68472004	0.55624074	0.8047968	0.45476437	0.5589634	0.5891485	0.5935639	0.8436089
A2	0.3467908	0.0000000	0.6778027	0.3114412	0.38047966	0.41292214	0.48361170	0.54963809	0.4841489	0.51580876	0.5484602	0.3201036	0.1706498	0.3517954
A3	0.6236328	0.6778027	0.0000000	0.6632949	0.16293482	0.35239354	0.55971247	0.73064321	0.7031702	0.38243073	0.4782466	0.2638300	0.3214901	0.4803876
A4	0.3606353	0.3114412	0.6632949	0.0000000	0.33091697	0.35172653	0.35172653	0.25934052	0.1126518	0.56441259	0.2133673	0.4873003	0.8087130	0.3859991
A5	0.5133063	0.3804797	0.1629348	0.3309170	0.00000000	0.51081610	0.77614009	0.32070974	0.5805302	0.35429904	0.6540570	0.4160345	0.6226164	0.5883443
A6	0.3294921	0.4129221	0.3523935	0.3517265	0.51081610	0.00000000	0.25750709	0.62801814	0.3755509	0.52740645	0.4953749	0.5220370	0.6654158	0.4895773
A7	0.6847200	0.4836117	0.5597125	0.3517265	0.77614009	0.25750709	0.00000000	0.35239354	0.5280332	0.59633583	0.3596489	0.3164459	0.6654158	0.6280181
A8	0.5562407	0.5496381	0.7306432	0.2593405	0.32070974	0.62801814	0.35239354	0.00000000	0.4946115	0.51217419	0.6074858	0.5148181	0.2656330	0.3515458
A9	0.8047968	0.4841489	0.7031702	0.1126518	0.58053017	0.37555087	0.52803320	0.49461147	0.0000000	0.46366644	0.8980605	0.6673600	0.3365353	0.7629703
A10	0.4547644	0.5158088	0.3824307	0.5644126	0.35429904	0.52740645	0.59633583	0.51217419	0.4636664	0.00000000	0.5730616	0.6118131	0.6630527	0.1689429

Ilustración 14-4: Matriz de similaridad de la primera interacción en RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Dendograma

Observemos que el dendograma de Matlab y Rchic respectivamente no se parecen y esto sucede porque cuando las bases superan las 30 variables, Matlab ya no presenta todas las variables en el dendograma.

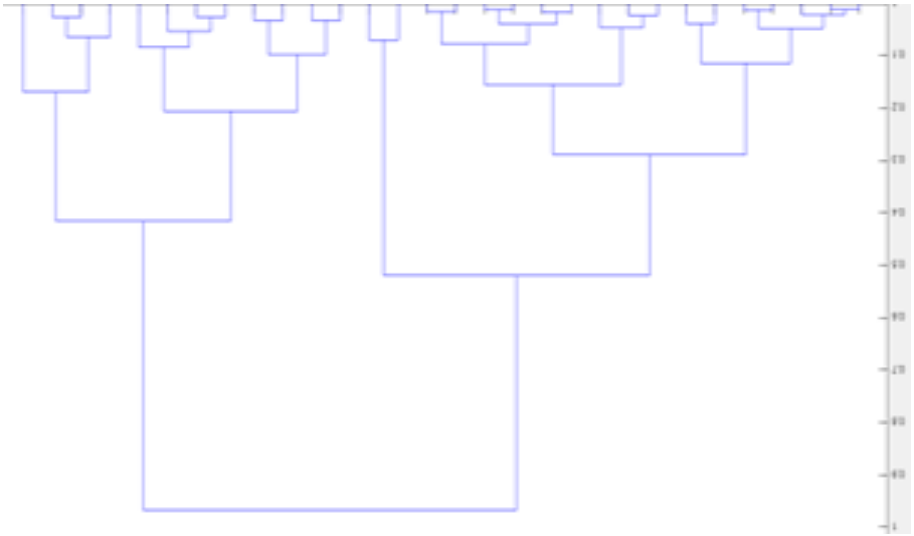


Ilustración 15-4: Dendograma en Matlab.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

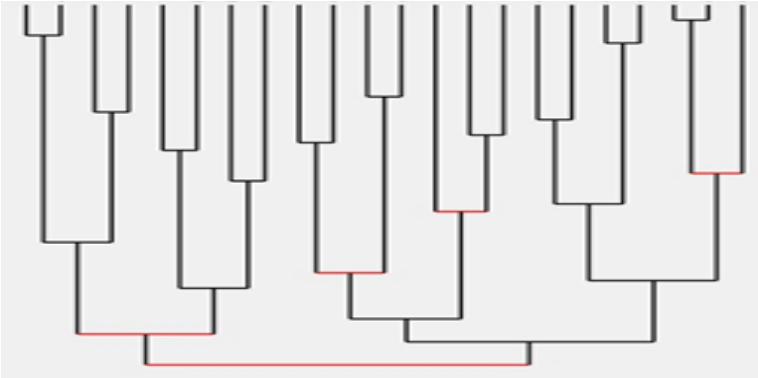


Ilustración 16-4: Árbol de similaridad en RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Entonces con la tercera programación hecha comprobamos que los índices de similaridad coinciden en su totalidad pero no coincide el dendograma. Pero no se ha logrado sacar los siguientes niveles de similaridad por lo que la programación no está completa. Es decir, aún no se cumple el objetivo. Esta programación será muy útil para quienes deseen generar bases aleatorias con datos binarios.

Ahora veamos la programación final de la Similaridad de Lerman.

Tengamos claro que para comparar la similaridad de los dos programas se estableció una muestra de 3– de basea a ser analizadas. Por lo que esta última programación no trabajará con interacciones, si no que leerá el archivo Excel con extensión ".xlsx"que contiene los variables y filas que serán analizadas.

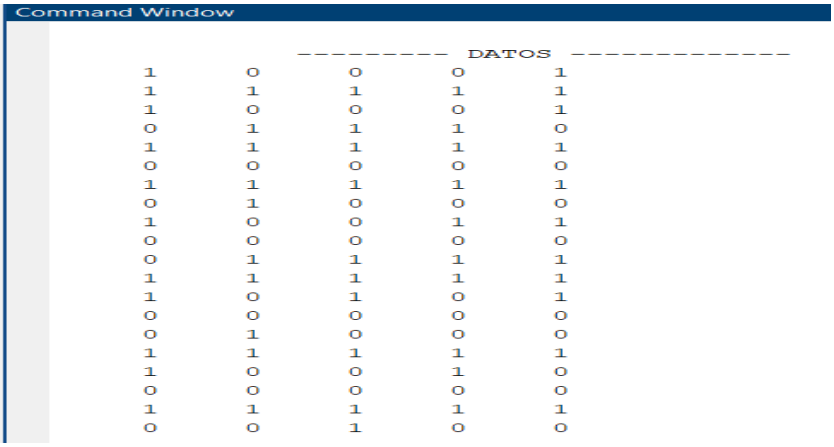


Ilustración 17-4: Base de datos binarios.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Command Window

```
            --------- TABLA -------------
      Variables        Card        kc         s
    1.0000    2.0000    6.0000    0.2132    0.5844
    1.0000    3.0000    7.0000    0.6396    0.7388
    1.0000    4.0000    8.0000    1.0660    0.8568
    1.0000    5.0000   10.0000    1.6059    0.9459
    2.0000    3.0000    8.0000    1.3416    0.9101
    2.0000    4.0000    8.0000    1.3416    0.9101
    2.0000    5.0000    7.0000    0.6396    0.7388
    3.0000    4.0000    8.0000    1.3416    0.9101
    3.0000    5.0000    8.0000    1.0660    0.8568
    4.0000    5.0000    8.0000    1.0660    0.8568
```

Ilustración 18-4: Tabla de valores de copresencias estandarizadas e índices de similaridad.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Command Window

```
            --------- RESULTADOS -------------

----- Matriz de similaridad al nivel 0 -----------
    1.0000    0.5844    0.7388    0.8568    0.9459
    0.5844    1.0000    0.9101    0.9101    0.7388
    0.7388    0.9101    1.0000    0.9101    0.8568
    0.8568    0.9101    0.9101    1.0000    0.8568
    0.9459    0.7388    0.8568    0.8568    1.0000


----- Matriz de similaridad al nivel 1 -----------
    1.0000    0.5458    0.7341    0.7341
    0.5458    1.0000    0.9101    0.9101
    0.7341    0.9101    1.0000    0.9101
    0.7341    0.9101    0.9101    1.0000


----- Matriz de similaridad al nivel 2 -----------
    1.0000    0.4617    0.7341
    0.4617    1.0000    0.8284
    0.7341    0.8284    1.0000


----- Matriz de similaridad al nivel 3 -----------
    1.0000    0.3956
    0.3956    1.0000
```

Ilustración 19-4: Niveles de jerarquía.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

Fijemonos que ahora si muestra los niveles de jerarquía, la cantidad de niveles es diferente para cada base aleatoria, podrán tener más o menos que en este ejemplo. Lo importante es que ahora

la programación calcula de forma automática todo. Lo único que aún no se ha podido hacer es el dendograma (árbol de similaridad) debido a que cuando una base posee más de 30 variables solo logra graficar las 30 variables, no más, por lo que no se puede hacer una comparación entre el árbol de similaridad de RCHIC y el dendograma de Matlab.

4.2. Discusión

4.2.1. *Estadísticos descriptivos: matriz similaridad (mCHIC y RCHIC), matriz similaridad (mCHIC y CHIC), gráfica (RCHIC y CHIC), nodos (mCHIC y RCHIC), nodos (mCHIC y CHIC)*

Tabla 1-4: Comparaciones estadísticas

Estadísticas							
Variable	**Conteo total N**	**N***	**Porcentaje**	**PrcAcum**	**Media**	**Desv.Est.**	
Matriz Similaridad mCHIC y RCHIC	383	383	0	100	100	0.91942	0.03229
Matriz Similaridad mCHIC y CHIC	383	383	0	100	100	0.94308	0.04958
Gráfica RCHIC y CHIC	383	383	0	100	100	1.0000	0.000000
Nodos mCHIC y RCHIC	383	383	0	100	100	0.99000	0.000000
Nodos mCHIC y CHIC	383	383	0	100	100	0.99000	0.000000
Variable	**Varianza**	**CoefVar**	**Mínimo**	**Q1**	**% Mediana**	**Q3**	**Máximo**
Matriz Similaridad mCHIC y RCHIC	0.00104	3.51	0.80000	0.92000	0.93000	0.94000	0.98000
Matriz Similaridad mCHIC y CHIC	0.00246	5.26	0.90000	0.90000	0.90000	1.00000	1.00000
Gráfica RCHIC y CHIC	0.000000	0.00	1.0000	1.0000	1.0000	1.0000	1.0000
Nodos mCHIC y RCHIC	0.000000	0.00	0.99000	0.99000	0.99000	0.99000	0.99000
Nodos mCHIC y CHIC	0.000000	0.00	0.990000	0.99000	0.99000	0.99000	0.99000
Variable	**IQR**	**Modo**	**N para moda**	**Asimetría**	**Curtosis**		
Matriz Similaridad (mCHIC y RCHIC)	0.02000	0.94	116	-1.75	2.79		
Matriz Similaridad (mCHIC y CHIC)	0.10000	0.9	218	0.28	-1.93		
Gráfica (RCHIC y CHIC)	0.000000	1	383	*	*		
Nodos (mCHIC y RCHIC)	0.000000	0.99	383	*	*		
Nodos (mCHIC y CHIC)	0.000000	0.99	383	*	*		

Fuente: Elaboración propia.
Realizado por: Córdova, Anabel, 2023.

En la tabla 7 de las comparaciones estadísticas, se observa que al realizar el análisis, la comparación entre la matriz de similaridad Mchic y la matriz de similaridad Chic tiene un mayor promedio de 0.94308, valor que se da del conteo total de las 383 bases que son el 100 % de las bases de la muestra, obteniendo una desviación estándar de 0.04958, con una varianza de 0.00246, un

coeficiente de variación de 5.26; el valor mínimo analizado de los 383 datos es 0.90000, en el primer cuartil que corresponde al 25 % del análisis de los 383 datos se obtiene un valor de 0.90000, en el segundo cuartil que corresponde a la mediana de los 383 datos analizados obteniendo un valor de 0.90000, en el tercer cuartil que corresponde al 75 % del análisis de los 383 datos se obtiene un valor de 1.00000 y el valor máximo analizado de los 383 datos es 1.00000. El rango intercuartílico (IQR), que viene siendo la diferencia entre el tercer cuartil y el primer cuartil, tiene un valor de 0.10000, el valor de la moda de los 383 datos analizados es 0.9 con un número de 218 veces repetidas, se tiene una simetría de los datos de 0.28 y una curtosis de -1.93.

Por otro lado, la comparación entre la matriz de similaridad Mchic y la matriz de similaridad Rchic tiene un menor promedio de 0.91942 con una desviación estandar de 0.03229, con una varianza de 0.00104, un coeficiente de variación de 3.51; el valor mínimo analizado de los 383 datos es 0.80000, en el primer cuartil que corresponde al 25 % del análisis de los 383 datos se obtiene un valor de 0.92000, en el segundo cuartil que corresponde a la mediana de los 383 datos analizados obteniendo un valor de 0.93000, en el tercer cuartil que corresponde al 75 % del análisis de los 383 datos se obtiene un valor de 0.94000 y el valor máximo analizado de los 383 datos es 0.98000. El rango intercuartílico (IQR), que viene siendo la diferencia entre el tercer cuartil y el primer cuartil, tiene un valor de 0.02000, el valor de la moda de los 383 datos analizados es 0.94 con un número de 116 veces repetidas, se tiene una simetría de los datos de -1.75 y una curtosis de 2.79. Con lo que concluimos que la programación de la similaridad de Lerman hecha en Matlab tiene mayor similutd con el programa Chic.

A su vez, la compración entre las gráficas (dendograma o árbol de similaridad) de M chic y Rchic nos da un promedio de 1,0000, valor que se da del conteo total de las 383 bases que son el 100 % de las bases de la muestra, obteniendo una desviación estándar de 0.000000, con una varianza de 0.000000, un coeficiente de variación de 0.00; el valor mínimo analizado de los 383 datos es 1.0000, en el primer cuartil que corresponde al 25 % del análisis de los 383 datos se obtiene un valor de 1.0000, en el segundo cuartil que corresponde a la mediana de los 383 datos analizados obteniendo un valor de 1.0000, en el tercer cuartil que corresponde al 75 % del análisis de los 383 datos se obtiene un valor de 1.0000 y el valor máximo analizado de los 383 datos es 1.0000. El rango intercuartílico (IQR), que viene siendo la diferencia entre el tercer cuartil y el primer cuartil, tiene un valor de 0.000000, el valor de la moda de los 383 datos analizados es 1 con un número de 383 veces repetidas, no se tiene una simetría de datos ni de curtosis; lo cual indica que las gráficas son idénticas.

Luego, la comparación entre los nodos que muestra Mchic con los nodos de Rchic y Chic tienen un promedio del 0.99000, valor que se da del conteo total de las 383 bases que son el 100 % de las bases de la muestra, obteniendo una desviación estándar de 0.000000, con una varianza de 0.000000, un coeficiente de variación de 0.00; el valor mínimo analizado de los 383 datos es 0.990000, en el primer cuartil que corresponde al 25 % del análisis de los 383 datos se obtiene un valor de 0.99000, en el segundo cuartil que corresponde a la mediana de los 383 datos analizados obteniendo un valor de 0.99000, en el tercer cuartil que corresponde al 75 % del análisis de los 383 datos se obtiene un valor de 0.99000 y el valor máximo analizado de los 383 datos es 0.99000. El rango intercuartílico (IQR), que viene siendo la diferencia entre el tercer cuartil y el primer cuartil, tiene un valor de 0.000000, el valor de la moda de los 383 datos analizados es 0.99 con un número de 383 veces repetidas tanto para la comparación de los nodos entre Mchic con los nodos de Rchic y Chic, no se tiene una simetría de datos ni de curtosis; es decir que se muestra en su mayoría el mismo número de nodos en los tres programas.

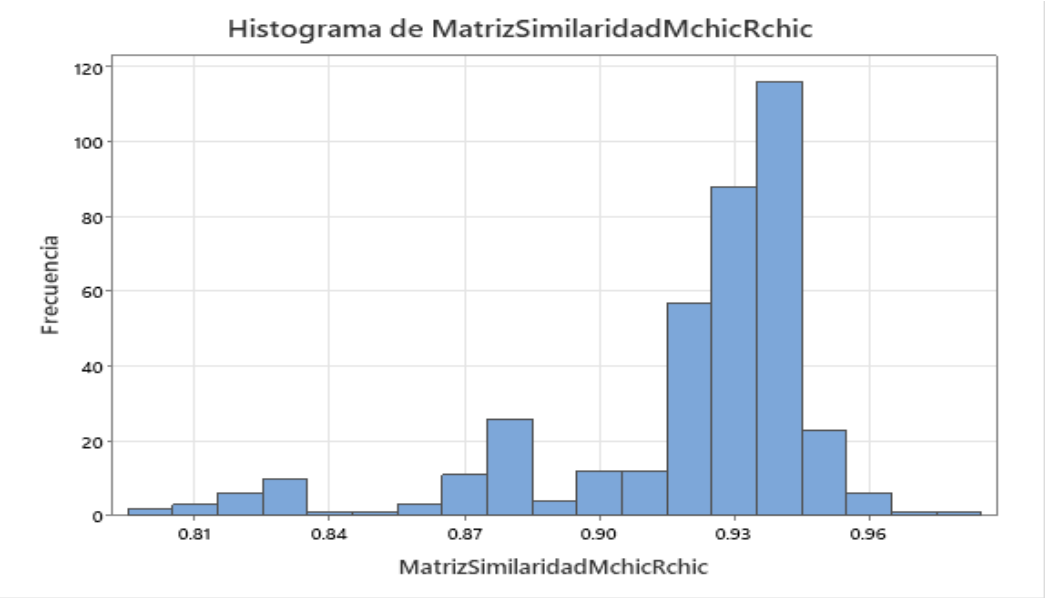


Ilustración 20-4: Histograma de Matriz Similaridad entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 20 se observa el histograma de la comparación entre la matriz de similaridad de Mchic y la matriz de similaridad de Rchic donde observamos que al menos una base de datos binarios tiene los siguientes valores de similaridad: 0.8, 0.84, 0.85, 0.97 y 0.98. De 2 a 10 bases de datos binarios tienen los siguientes valores de similaridad: 0.81, 0.82, 0.83, 0.86, 0.87, 0.90, 0.91, 0.96. De 20 a 60 bases de datos binarios tienen los siguientes valores de similaridad: 0.88, 0.92, 0.95. De

80 hasta cerca de 120 bases de datos binarios tienen los siguientes valores de similaridad: 0.93, 0.94. Por lo que podemos concluir que existe la moda en el valor 0.94 con 116 bases de datos binarios que contienen ete valor de similaridad. Hay que tener en cuenta que mientras más grande sea el número de variables de una base de datos binarios se va a tener mejor similaridad.

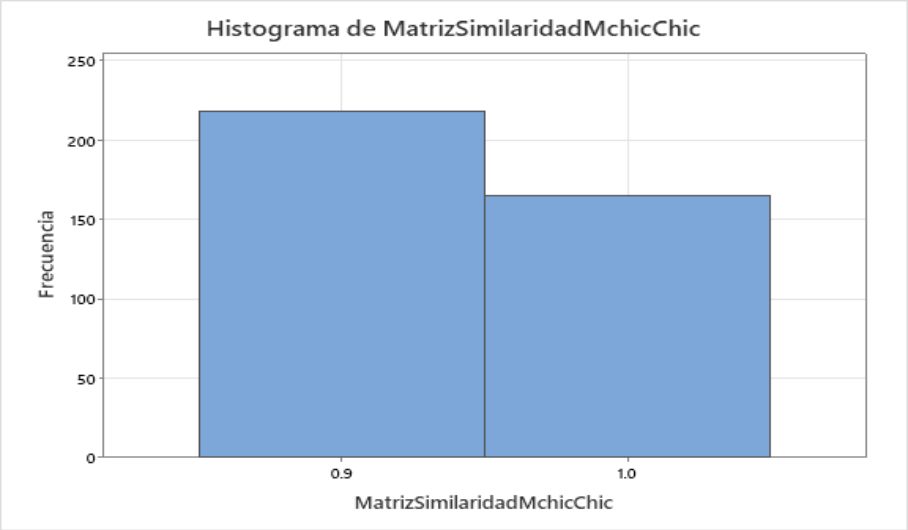


Ilustración 21-4: Histograma de Matriz Similaridad entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 21 se observa el histograma de la comparación entre la matriz de similaridad de Mchic y la matriz de similaridad de Chic donde observamos que 218 bases de datos binarios tienen una similaridad de 0.9 y 165 bases de datos binarios tienen una similaridad de 1.0.

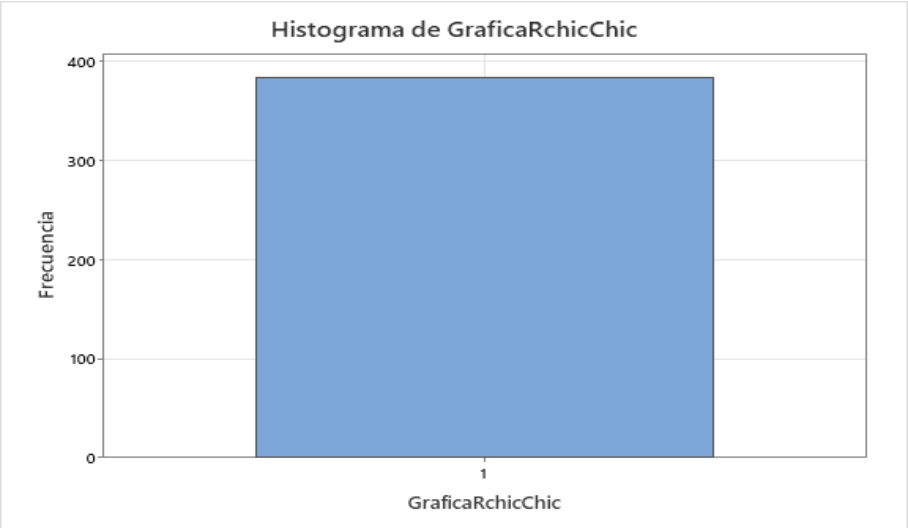


Ilustración 22-4: Histograma de Gráfica entre RCHIC y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 22 se observa el histograma de la comparación entre la gráfica (dendograma o árbol de similaridad) de Rchic y la gráfica (dendograma o árbol de similaridad) de CHIC donde observamos que todos los dendogramas de las 383 bases de datos binarios son iguales con un valor de 1.

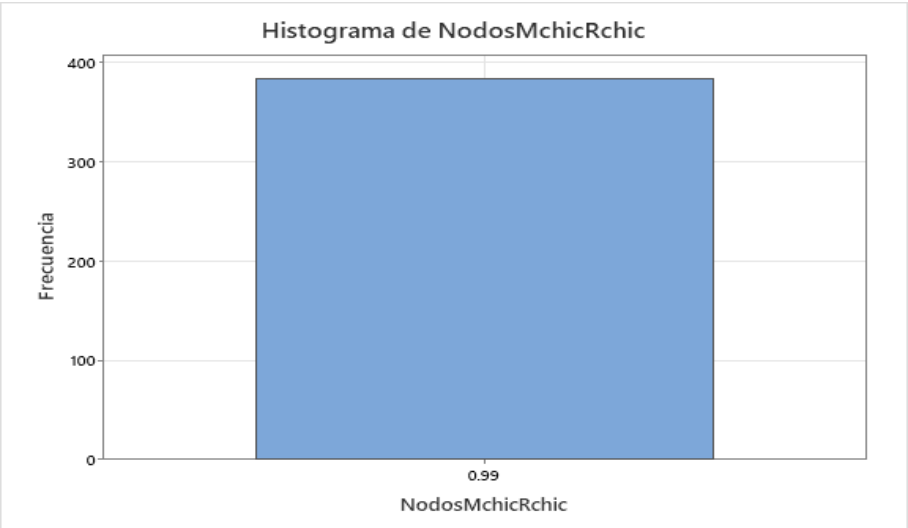


Ilustración 23-4: Histograma de Nodos entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 23 se observa el histograma de la comparación entre los nodos que muestra Mchic y los nodos que muestra Rchic donde observamos que todas las 383 bases de datos binarios tienen una similaridad de 0.99.

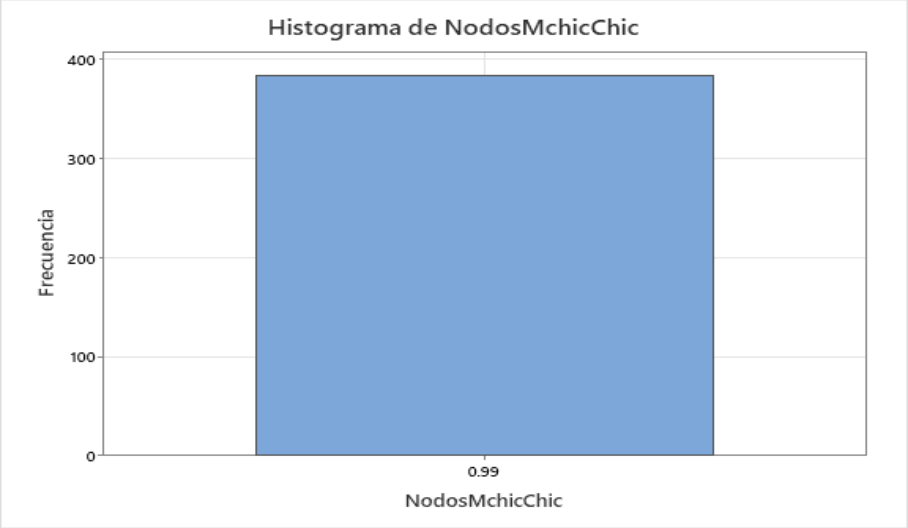


Ilustración 24-4: Histograma de Nodos entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 24 se observa el histograma de la comparación entre los nodos que muestra Mchic y los nodos que muestra Chic donde observamos que todas las 383 bases de datos binarios tienen una similaridad de 0.99.

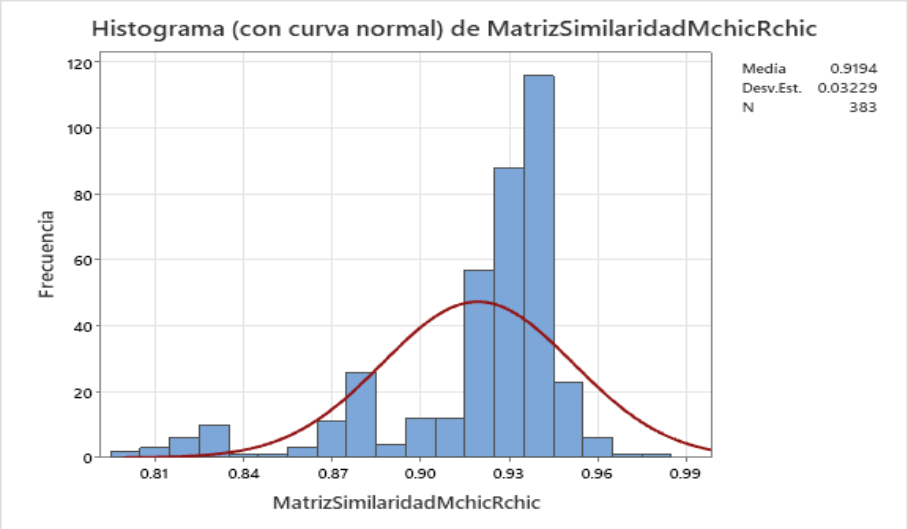


Ilustración 25-4: Histograma (con curva normal) de Matriz Similaridad entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 25 del histograma (con curva normal) de Matriz de Similaridad entre Mchic y Rchic

se observa que presenta un leve sesgo a la izquiera que indica que es asimétricamente negativo con un valor de -1.75, se tiene una media de 0.9194 y una desviación estandar de 0.03229.

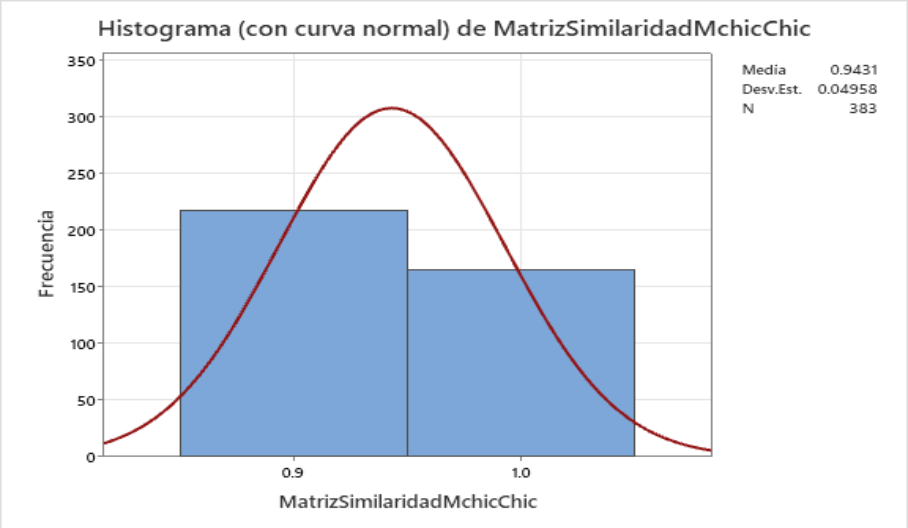


Ilustración 26-4: Histograma (con curva normal) de Matriz Similaridad entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 26 del histograma (con curva normal) de Matriz de Similaridad entre Mchic y Chic se observa que presenta un leve sesgo a la derecha que indica que es asimétricamente positivo con un valor de 0.28, se tiene una media de 0.9431 y una desviación estandar de 0.04958.

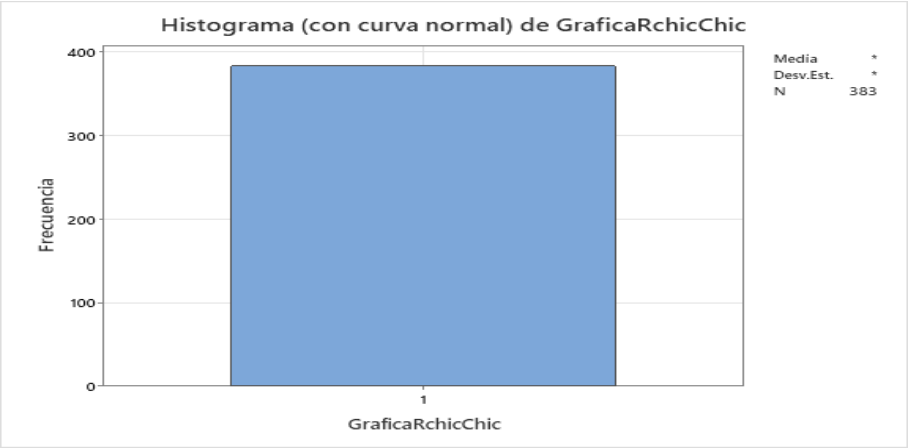


Ilustración 27-4: Histograma (con curva normal) de Gráfica entre RCHIC y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

* NOTA * La distribución no se pudo ajustar. El número de filas de datos distintas en GraficaRchicChic debe ser mayor que o igual al número de parámetros de distribución estimados.

En la Figura 27 del histograma de comparación entre la gráfica de Rchic y Chic no se presenta una curva normal ya que las gráficas son identicas.

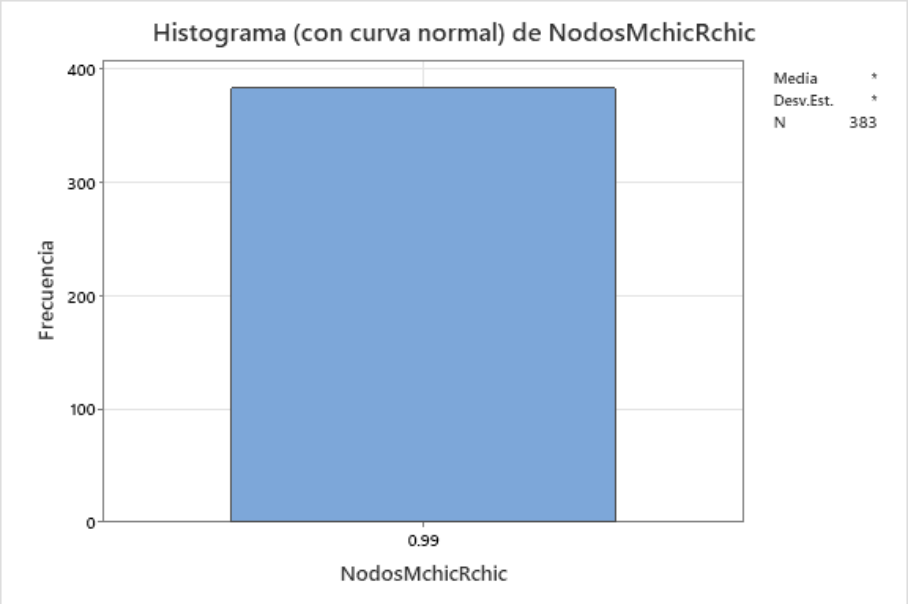


Ilustración 28-4: Histograma (con curva normal) de Nodos entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

* NOTA * La distribución no se pudo ajustar. El número de filas de datos distintas en NodosMchicRchic debe ser mayor que o igual al número de parámetros de distribución estimados.

En la Figura 28 del histograma de comparación entre los nodos de Mchic y Rchic no se presenta una curva normal ya que los nodos son mayormente parecidos.

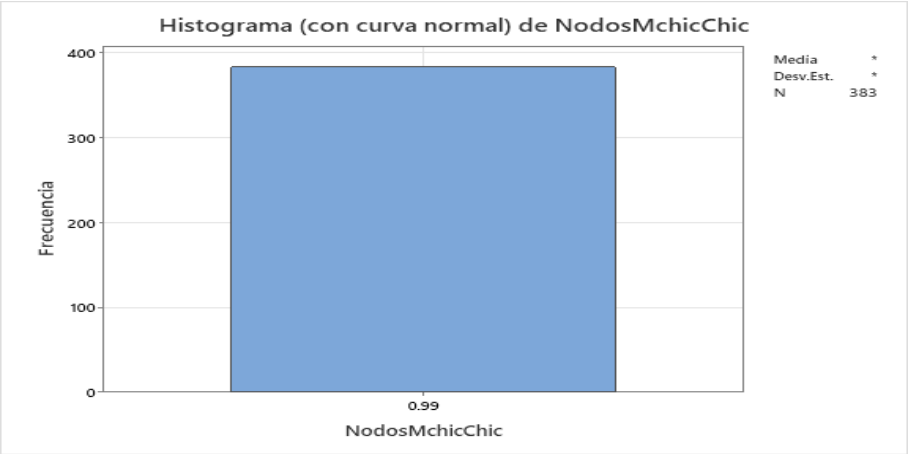


Ilustración 29-4: Histograma (con curva normal) de Nodos entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

* NOTA * La distribución no se pudo ajustar. El número de filas de datos distintas en NodosMchicChic debe ser mayor que o igual al número de parámetros de distribución estimados.

En la Figura 29 del histograma de comparación entre los nodos de Mchic y Chic no se presenta una curva normal ya que los nodos son mayormente parecidos.

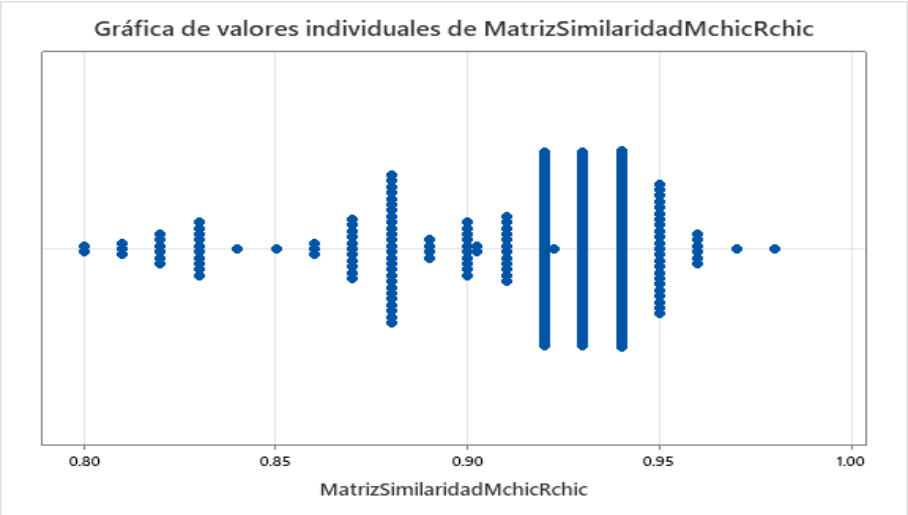


Ilustración 30-4: Gráfica de valores individuales de Matriz Similaridad entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 30 se presenta la Gráfica de valores individuales de la comparación entre la Matriz de Similaridad Mchic y Rchic donde observamos detalladamente cada una de las bases con datos binarios con su respectivo valor de similaridad, dando un total de 383 valores individuales.

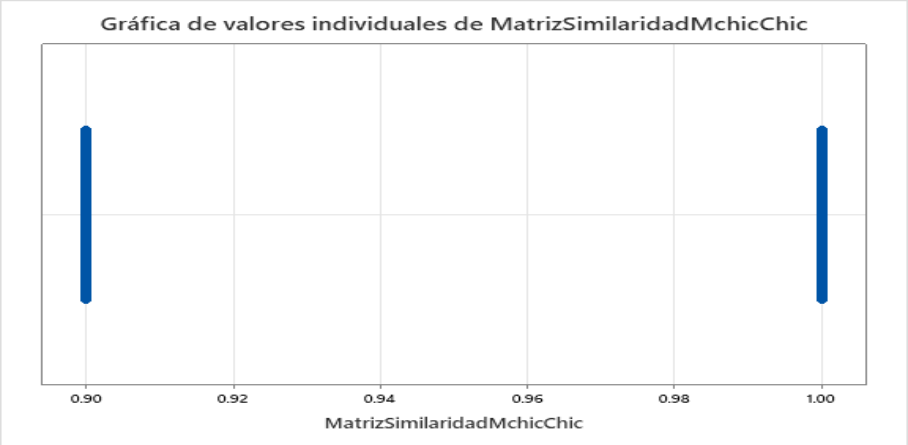


Ilustración 31-4: Gráfica de valores individuales de Matriz Similaridad entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 31 se presenta la Gráfica de valores individuales de la comparación entre la Matriz de Similaridad Mchic y Chic donde observamos que hay 218 puntos que representan a cada base de datos binarios y estas tienen el valor de similaridad de 0.90 y los 165 puntos restantes que también representan a cada base de datos binarios tienen el valor de similaridad de 1.00.

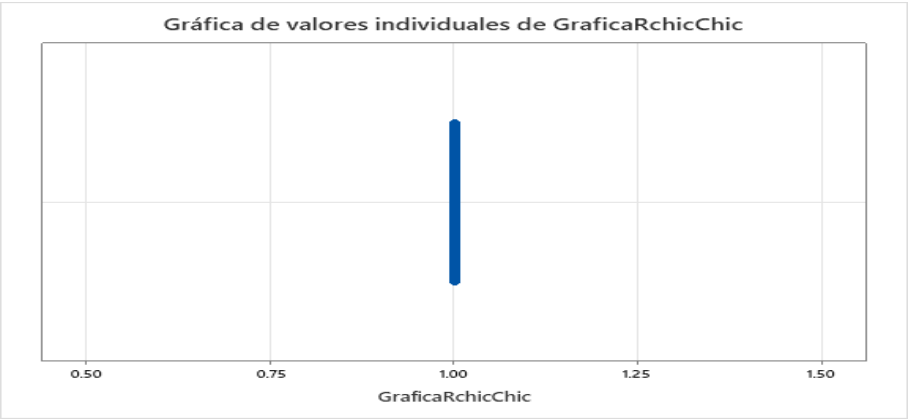


Ilustración 32-4: Gráfica de valores individuales de Gráfica entre RCHIC y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 32 se presenta la gráfica de valores individuales de la comparación entre la gráfica (dendograma o árbol de similaridad) de Rchic y Chic donde las 383 bases de datos binarios tienen un valor de similaridad de 1.00.

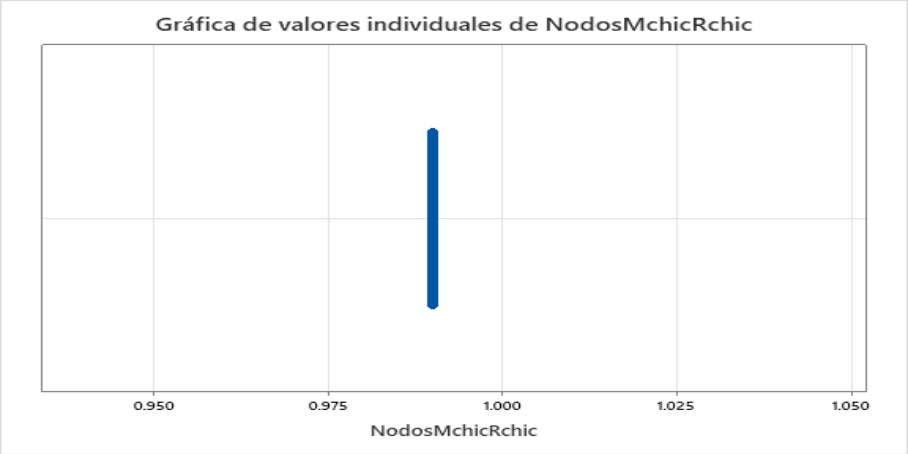


Ilustración 33-4: Gráfica de valores individuales de Nodos entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 33 se presenta la gráfica de valores individuales de la comparación entre los nodos de Mchic y Rchic donde las 383 bases de datos binarios tienen un valor de similaridad de 0.99.

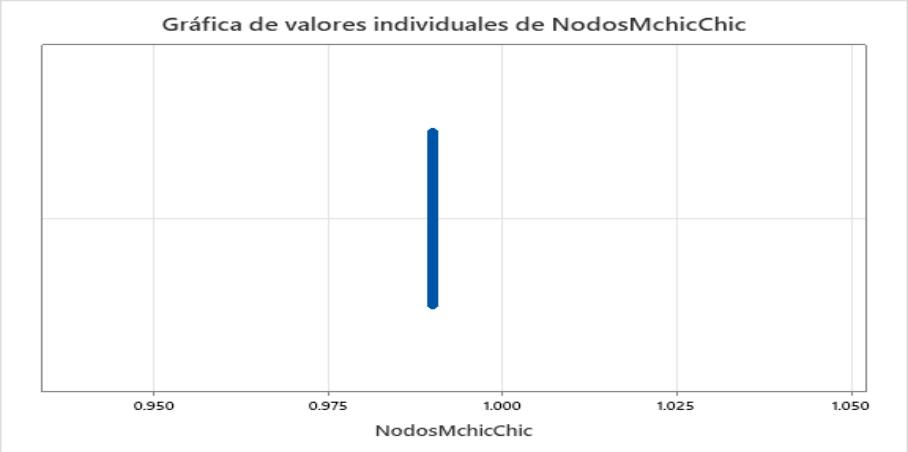


Ilustración 34-4: Gráfica de valores individuales de Nodos entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 34 se presenta la gráfica de valores individuales de la comparación entre los nodos de Mchic y Chic donde las 383 bases de datos binarios tienen un valor de similaridad de 0.99.

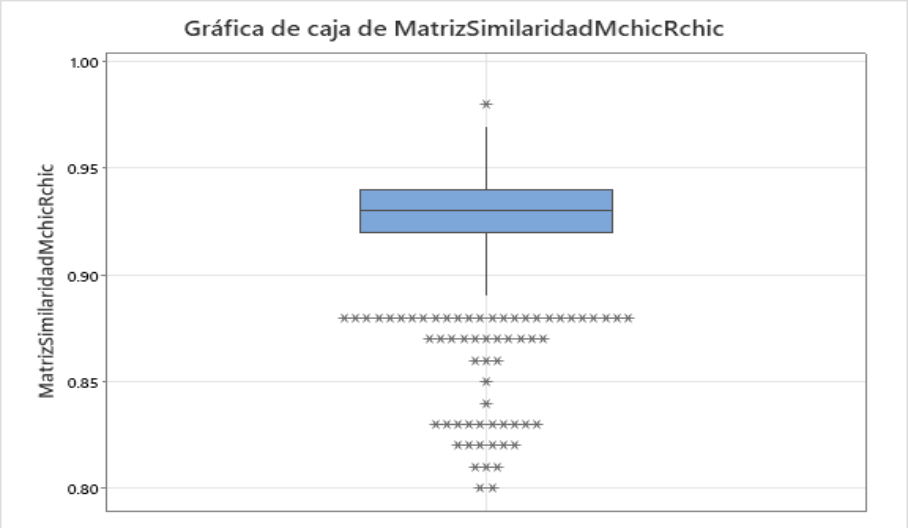


Ilustración 35-4: Gráfica de caja de Matriz Similaridad entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 35 que representa a la gráfica de caja de la comparación entre la Matriz de Similaridad de Mchic y Rchic se puede observar que el centro de los datos está en el valor 0.94, podemos observar un valor atípico entre el 0.95 y 1, otros valores atípicos estan en valor 0.8.

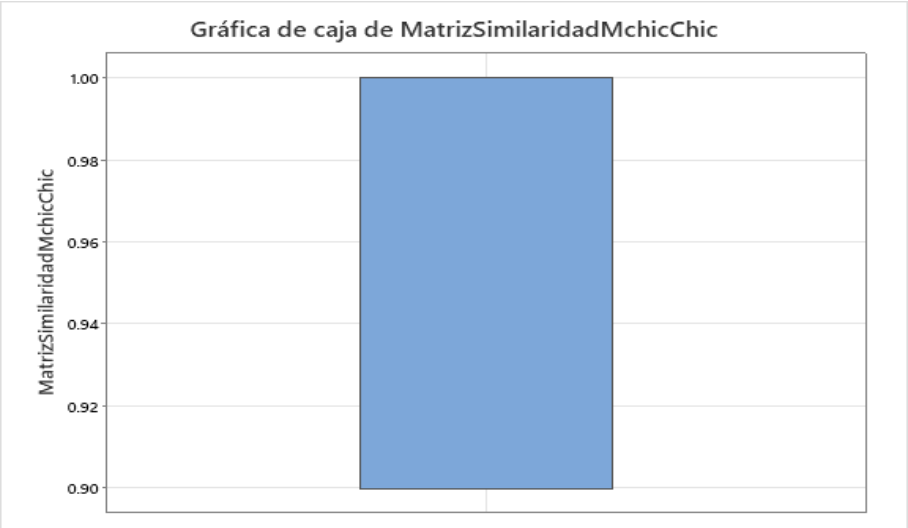


Ilustración 36-4: Gráfica de caja de Matriz Similaridad entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 36 que representa a la gráfica de caja de la comparación entre la Matriz de Similaridad de Mchic y Chic se puede observar que el centro de los datos está entre los valores 0.9 y 1.

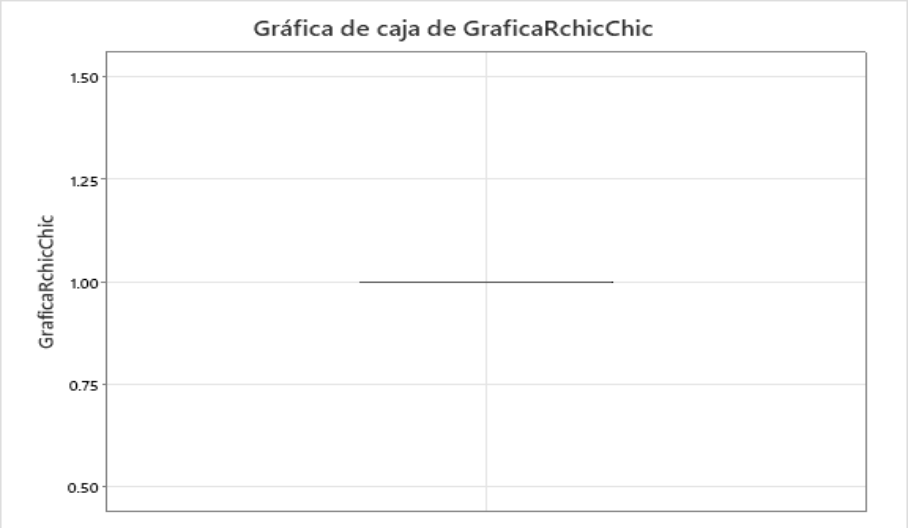


Ilustración 37-4: Gráfica de caja de Gráfica entre RCHIC y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 37 que representa a la gráfica de caja de la comparación entre la gráfica de Rchic y Chic se puede observar que el centro de los datos está en 1.

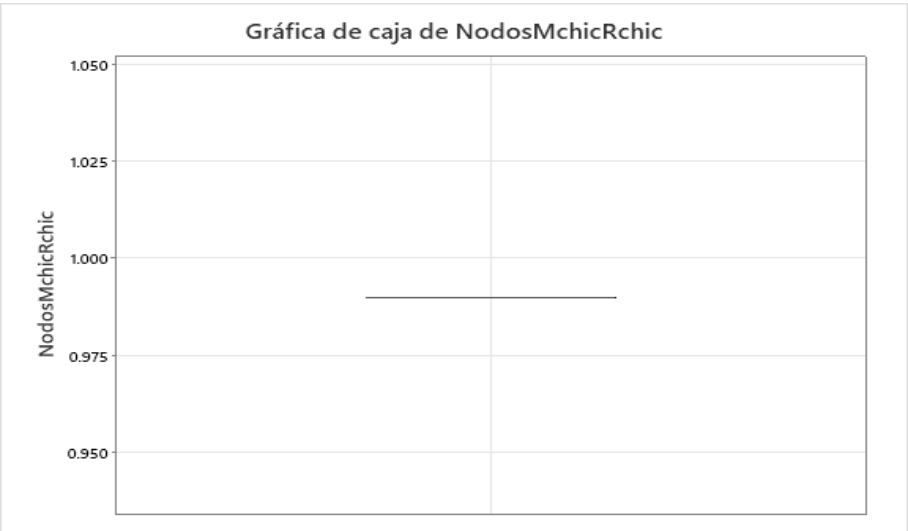


Ilustración 38-4: Gráfica de caja de Nodos entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 38 que representa a la gráfica de caja de la comparación entre los nodos de Mchic y

Rchic se puede observar que el centro de los datos está en el valor de 0.99.

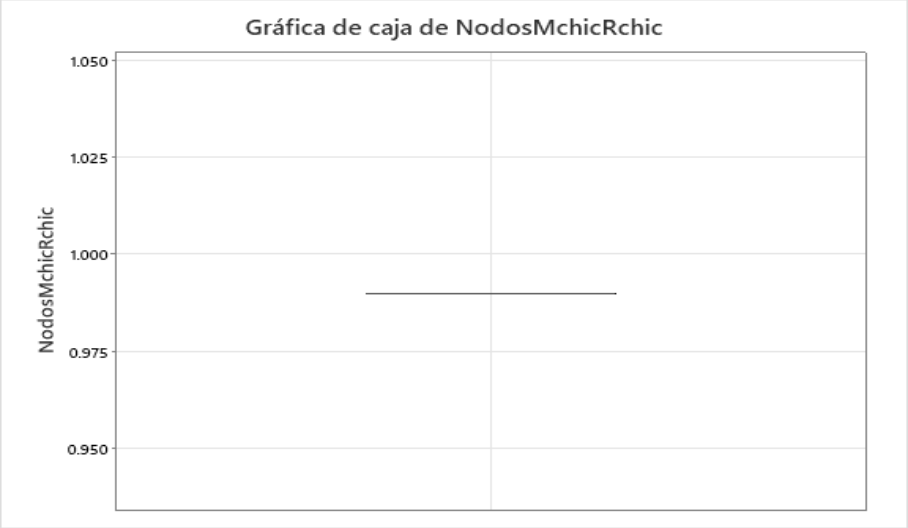


Ilustración 39-4: Gráfica de caja de Nodos entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 39 que representa a la gráfica de caja de la comparación entre los nodos de Mchic y Chic se puede observar que el centro de los datos está en el valor de 0.99.

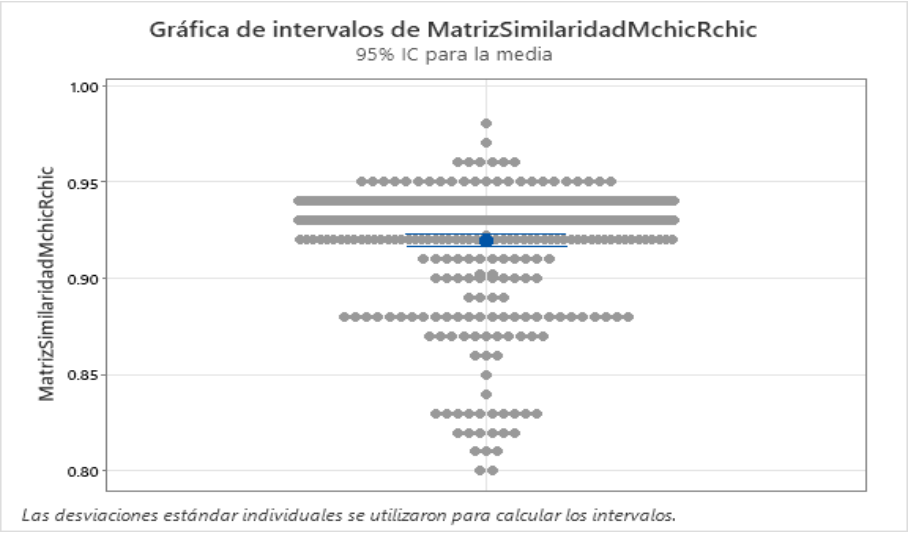


Ilustración 40-4: Gráfica de intervalos de Matriz Similaridad entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 40 se observa que para el para el 95 % de datos analizados entre la matriz de similaridad Mchic y Rchic existe cierta cantidad de datos que se encuentran en la zona de la media

que tiene un valor de 0.91942. También se tiene valores aislados que corresponden a la similaridad de 0.8 y 0.98.

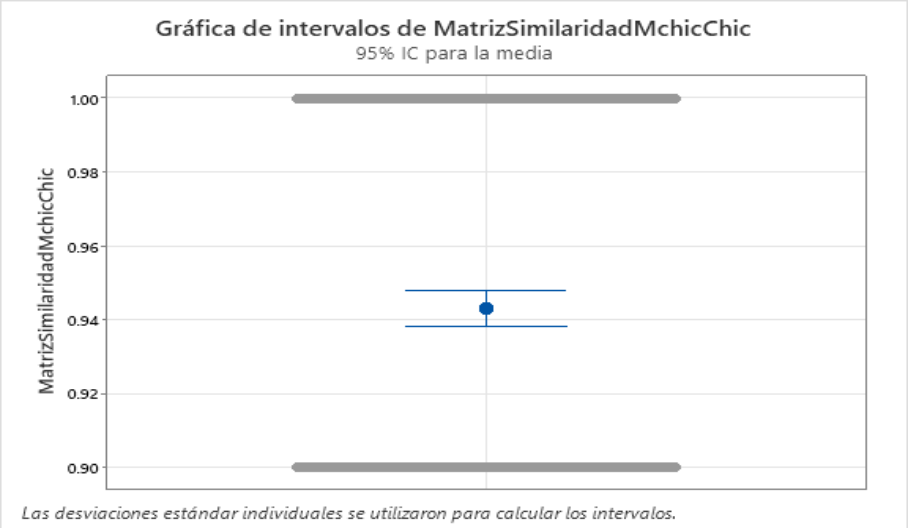


Ilustración 41-4: Gráfica de intervalos de Matriz Similaridad entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 41 se observa que para el para el 95% de datos analizados entre la matriz de similaridad Mchic y Chic se destaca el punto del valor de la media que es de 0.94308.

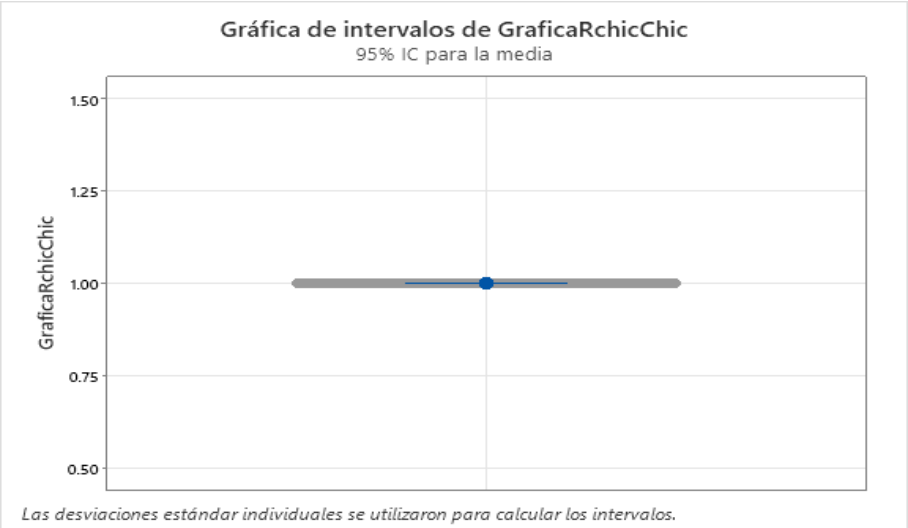


Ilustración 42-4: Gráfica de intervalos de gráfica RCHIC y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 42 se observa que para el para el 95 % de datos analizados entre la gráfica de similaridad Rchic y Chicse destaca el punto del valor de la media que es de 1.

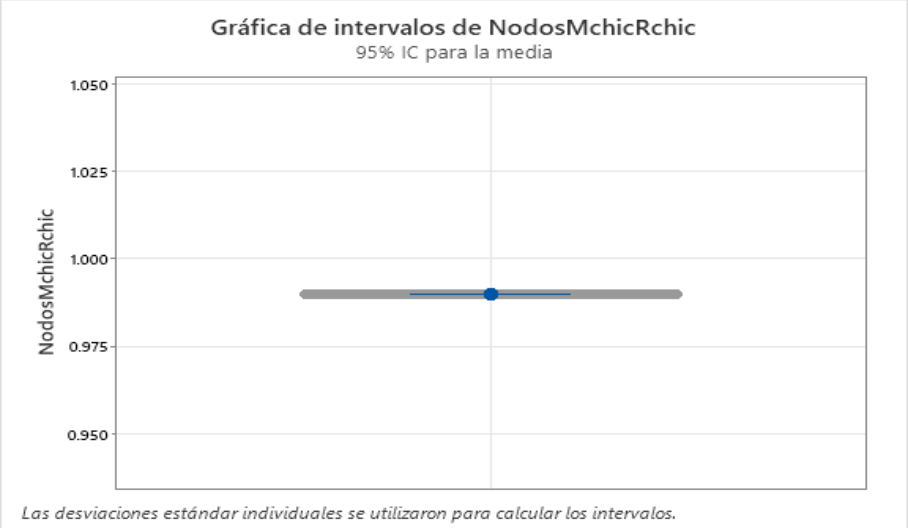


Ilustración 43-4: Gráfica de intervalos de Nodos entre mChic y RCHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 43 se observa que para el 95 % de datos analizados entre los nodos de Mchic y Rhic se destaca el punto del valor de la media que es de 0.99.

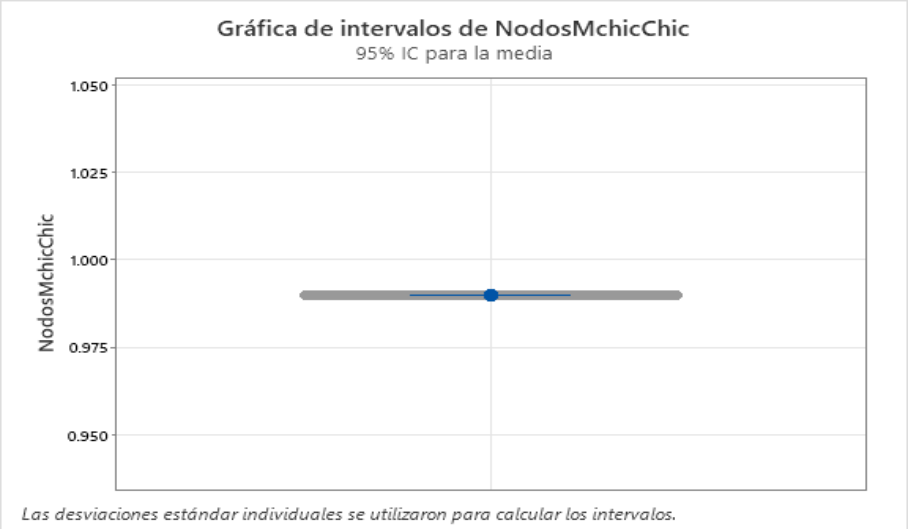


Ilustración 44-4: Gráfica de intervalos de Nodos entre mChic y CHIC.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 44 se observa que para el 95 % de datos analizados entre los nodos de Mchic y Rhic se

destaca el punto del valor de la media que es de 0.99.

4.3. Comprobación de la hipótesis

4.3.1. *Estadísticos descriptivos: matriz similaridad (mChic y RCHIC), matriz similaridad (mChic y CHIC), gráfica (RCHIC y CHIC), nodos (mChic y RCHIC), nodos (mChic y CHIC)*

Tabla 2-4: Comparaciones estadísticas descriptivas

Estadísticas descriptivas					
Muestra	**N**	**Media**	**Desv.Est.**	**Error estándar de la media**	**Límite inferior de 95 % para μ**
Matriz Similaridad (mChic y RCHIC)	383	0.91942	0.03229	0.00165	0.91670
Matriz Similaridad (mChic y CHIC)	383	0.94308	0.04958	0.00253	0.93890
Gráfica (RCHIC y CHIC)	383	1.000	0.000	0.000	1.000
Nodos (mChic y RCHIC)	383	0.9900	0.000	0.000	0.9900
Nodos (mChic y CHIC)	383	0.9900	0.000	0.000	0.9900

Fuente: Elaboración propia.
Realizado por: Córdova, Anabel, 2023.

μ: media de población de Matriz Similaridad (M-chic y R-chic), Matriz Similaridad (M-chic y Chic), Gráfica (R-chic y Chic), Nodos (M-chic y R-chic), Nodos (M-chic y Chic).

En la tabla 8 de las comparaciones estadísticas descriptivas, se observa que al realizar el análisis, la comparación entre la matriz de similaridad Mchic y la matriz de similaridad Chic tiene un mayor promedio (media) de 0.94308, valor que se da del conteo total de las 383 bases que son el 100 % de las bases de la muestra, obteniendo una desviación estándar de 0.04958, con un error estándar (valor de separación de los datos con respecto a la media) de 0.00253 y el límite inferior de 95 % para μ es de 0.93890.

Por otro lado, la comparación entre la matriz de similaridad Mchic y la matriz de similaridad Rchic tiene un menor promedio (media) de 0.91942 valor que se da del conteo total de las 383 bases que son el 100 % de las bases de la muestra, obteniendo una desviación estándar de 0.03229, con un error estándar (valor de separación de los datos con respecto a la media) de 0.00165 y el límite inferior de 95 % para μ es de 0.91670.

A su vez, la compración entre las gráficas (dendograma o árbol de similaridad) de M chic y Rchic nos da un promedio (media) de 1,0000, valor que se da del conteo total de las 383 bases que son el 100 % de las bases de la muestra, obteniendo una desviación estándar de 0.000000, con un error estándar (valor de separación de los datos con respecto a la media) de 0.000 y el límite inferior de 95 % para μ es de 1.000.

Luego, la comparación entre los nodos que muestra Mchic con los nodos de Rchic y Chic tienen un promedio del 0.99000, valor que se da del conteo total de las 383 bases que son el 100 % de las bases de la muestra, obteniendo una desviación estándar de 0.000000 con un error estándar (valor de separación de los datos con respecto a la media) de 0.000 y el límite inferior de 95 % para μ es de 0.9900.

4.3.2. *T de una muestra: matriz Similaridad (mChic y RCHIC), Matriz Similaridad (mChic y CHIC), gráfica (RCHIC y CHIC), nodos (mChic y RCHIC), nodos (mChic y CHIC*

Prueba

Hipóptesis nula $H_0 : \mu \leqslant 0.9$
Hipótesis alterna $H_1 : \mu > 0.9$

Tabla 3-4: Comparaciones estadísticas descriptivas (T de una muestra)

Estadísticas descriptivas		
Muestra	**Valor T**	**Valor p**
Matriz Similaridad (mChic y RCHIC)	11.77	0.000
Matriz Similaridad (mChic y CHIC)	17.00	0.000
Gráfica (RCHIC y CHIC)	*	*
Nodos (mChic y RCHIC)	*	*
Nodos (mChic y CHIC)	*	*

Fuente: Elaboración propia.
Realizado por: Córdova, Anabel, 2023.

En la tabla 9 de las comparaciones estadísticas descriptivas, se observa que al realizar el análisis, la

comparación entre la matriz de similaridad Mchic y la matriz de similaridad Chic, el valor de la diferencia en relación con la variación en los datos de la muestra es de 17.00. La comparación entre la matriz de similaridad Mchic y la matriz de similaridad Rchic, el valor de la diferencia en relación con la variación en los datos de la muestra es de 11.77.

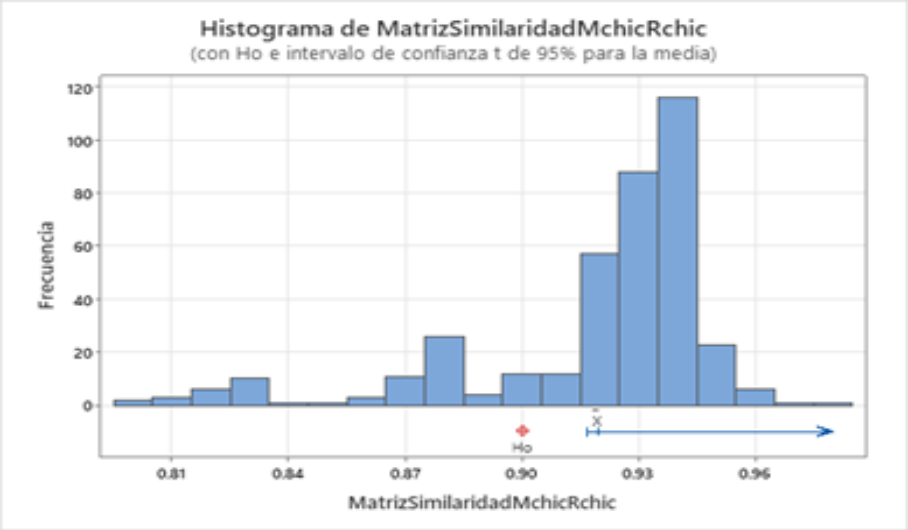


Ilustración 45-4: Histograma de Matriz Similaridad entre mChic y RCHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 45 se observa el histograma de la comparación entre la matriz de similaridad de Mchic y la matriz de similaridad de Rchic donde observamos que al menos una base de datos binarios tiene los siguientes valores de similaridad: 0.8, 0.84, 0.85, 0.97 y 0.98. De 2 a 10 bases de datos binarios tienen los siguientes valores de similaridad: 0.81, 0.82, 0.83, 0.86, 0.87, 0.90, 0.91, 0.96. De 20 a 60 bases de datos binarios tienen los siguientes valores de similaridad: 0.88, 0.92, 0.95. De 80 hasta cerca de 120 bases de datos binarios tienen los siguientes valores de similaridad: 0.93, 0.94. Por lo que podemos concluir que existe la moda en el valor 0.94 con 116 bases de datos binarios que contienen ete valor de similaridad. Hay que tener en cuenta que mientras más grande sea el número de variables de una base de datos binarios se va a tener mejor similaridad. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90%.

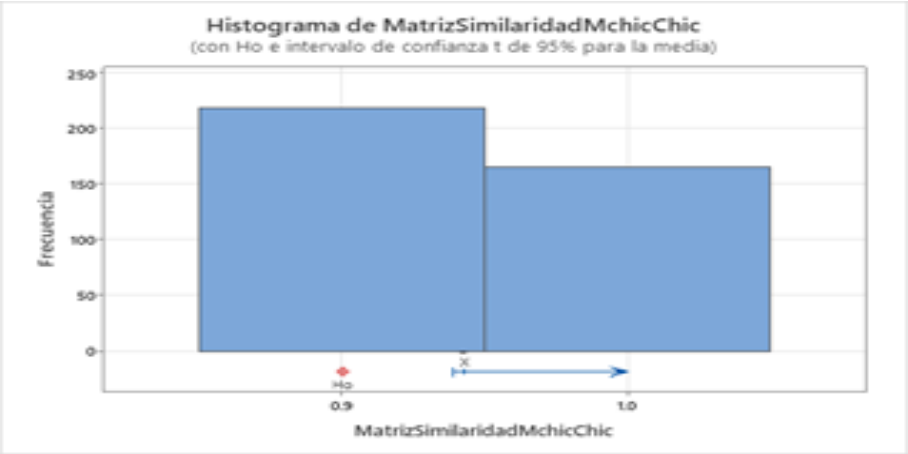


Ilustración 46-4: Histograma de Matriz Similaridad entre mChic y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 46 se observa el histograma de la comparación entre la matriz de similaridad de Mchic y la matriz de similaridad de Chic donde observamos que 218 bases de datos binarios tienen una similaridad de 0.9 y 165 bases de datos binarios tienen una similaridad de 1.0. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

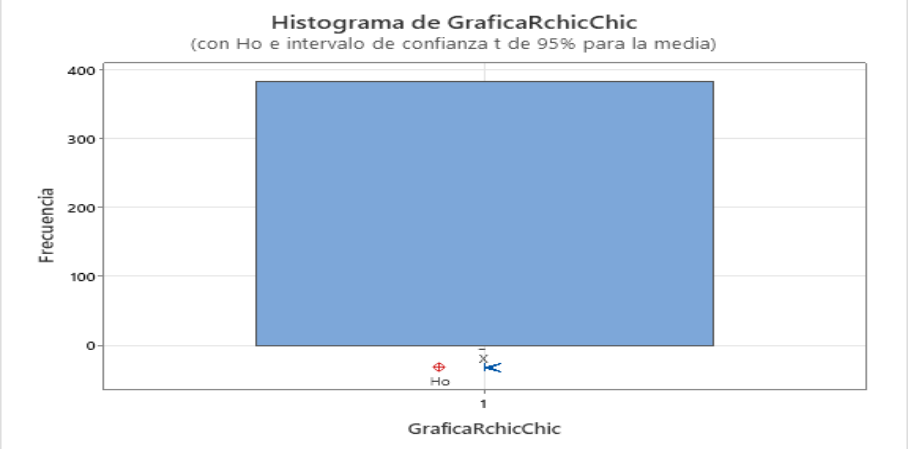


Ilustración 47-4: Histograma de Gráfica entre RCHIC y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 47 se observa el histograma de la comparación entre la gráfica (dendograma o árbol de similaridad) de Rchic y la gráfica (dendograma o árbol de similaridad) de CHIC donde observamos que todos los dendogramas de las 383 bases de datos binarios son iguales con un valor de 1. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

Ilustración 48-4: Histograma de Nodos entre mChic y RCHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 48 se observa el histograma de la comparación entre los nodos que muestra Mchic y los nodos que muestra Rchic donde observamos que todas las 383 bases de datos binarios tienen una similaridad de 0.99. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90%.

Ilustración 49-4: Histograma de Nodos entre mChic y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 49 se observa el histograma de la comparación entre los nodos que muestra Mchic y

los nodos que muestra Chic donde observamos que todas las 383 bases de datos binarios tienen una similaridad de 0.99. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

Ilustración 50-4: Gráfica de valores individuales de Matriz Similaridad mChic y RCHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 50 se presenta la Gráfica de valores individuales de la comparación entre la Matriz de Similaridad Mchic y Rchic donde observamos detalladamente cada una de las bases con datos binarios con su respectivo valor de similaridad, dando un total de 383 valores individuales. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

Ilustración 51-4: Gráfica de valores individuales de Matriz Similaridad mChic y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 51 se presenta la Gráfica de valores individuales de la comparación entre la Matriz de Similaridad Mchic y Chic donde observamos que hay 218 puntos que representan a cada base de datos binarios y estas tienen el valor de similaridad de 0.90 y los 165 puntos restantes que también representan a cada base de datos binarios tienen el valor de similaridad de 1.00. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

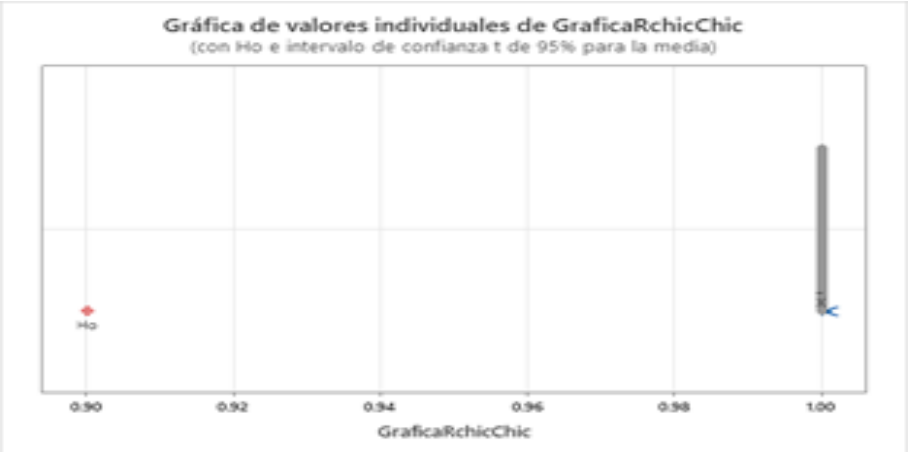


Ilustración 52-4: Gráfica de valores individuales de Gráfica entre RCHIC y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 52 se presenta la gráfica de valores individuales de la comparación entre la gráfica

(dendograma o árbol de similaridad) de Rchic y Chic donde las 383 bases de datos binarios tienen un valor de similaridad de 1.00. También se observa que se cumple la hipótesis pues se obtiene una similaridad del 100 % pues las gráfica son idénticas .

Ilustración 53-4: Gráfica de valores individuales de Nodos entre mChic y RCHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 53 se presenta la gráfica de valores individuales de la comparación entre los nodos de Mchic y Rchic donde las 383 bases de datos binarios tienen un valor de similaridad de 0.99. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

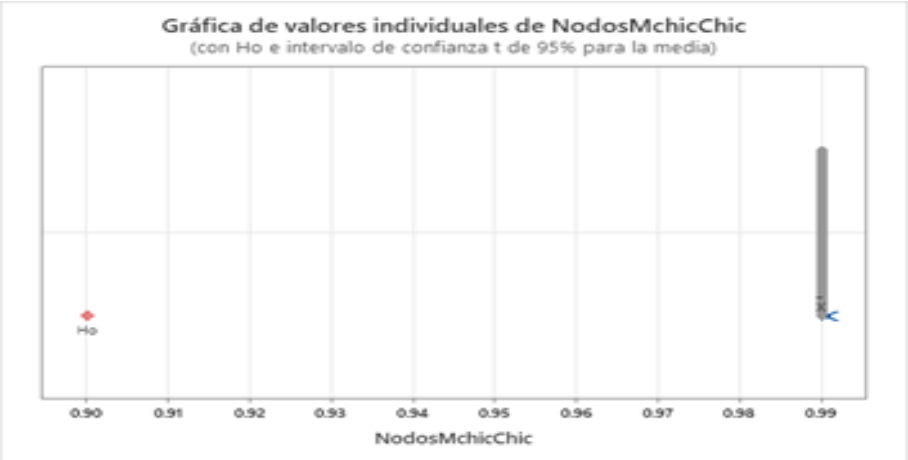


Ilustración 54-4: Gráfica de valores individuales de Nodos entre mChic y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 54 se presenta la gráfica de valores individuales de la comparación entre los nodos de Mchic y Chic donde las 383 bases de datos binarios tienen un valor de similaridad de 0.99. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

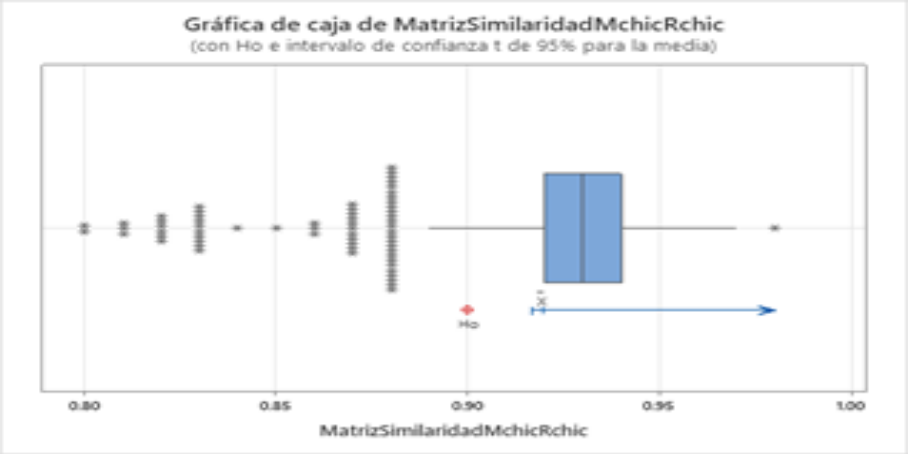


Ilustración 55-4: Gráfica de caja de Matriz Similaridad mChic y RCHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 55 que representa a la gráfica de caja de la comparación entre la Matriz de Similaridad de Mchic y Rchic se puede observar que el centro de los datos está en el valor 0.94, podemos observar un valor atípico entre el 0.95 y 1, otros valores atípicos estan en valor 0.8. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

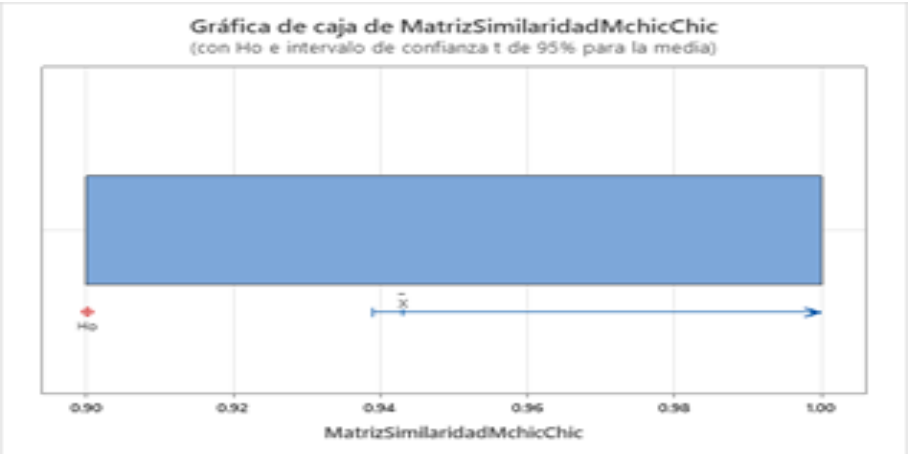


Ilustración 56-4: Gráfica de caja de Matriz Similaridad mChic y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 56 que representa a la gráfica de caja de la comparación entre la Matriz de Similaridad de Mchic y Chic se puede observar que el centro de los datos está entre los valores 0.9 y 1. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

Ilustración 57-4: Gráfica de caja de Gráfica entre RCHIC y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 57 que representa a la gráfica de caja de la comparación entre la gráfica de Rchic y Chic se puede observar que el centro de los datos está en 1. También se observa que se cumple la hipótesis pues se obtiene una similaridad del 100 % pues las gráfica son idénticas .

Ilustración 58-4: Gráfica de caja de Nodos mChic y RCHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

En la Figura 58 que representa a la gráfica de caja de la comparación entre los nodos de Mchic y Rchic se puede observar que el centro de los datos está en el valor de 0.99. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

Ilustración 59-4: Gráfica de caja de Nodos mChic y CHIC H0.

Fuente: Elaboración propia.

Realizado por: Córdova, Anabel, 2023.

* NOTA * Todos los valores de la columna son idénticos.

En la Figura 59 que representa a la gráfica de caja de la comparación entre los nodos de Mchic y Chic se puede observar que el centro de los datos está en el valor de 0.99. También se observa que se cumple la hipótesis pues se obtiene una similaridad mayor al 90 %.

CONCLUSIONES

El objetivo principal del presente proyecto de investigación es programar en el Software Matlab la similaridad de Lerman entre variables binarias y validarla mediante similaridad con RCHIC, por lo que, en primer lugar, se analizó y se estudió la formulación matemática de la similaridad de Lerman que es parte del Análisis Estadístco Implicativo; estas mismas fórmulas dieron paso a la elaboración del pseudocódigo de la similaridad de Lerman en el programa Matlab, la programación lleva como nombre **mCHIC**.

Con los resultados que se obtuvieron al momento de hacer la comparación del programa mCHIC con los programas CHIC y RCHIC, se llegó a la conclusión de que se obtuvo una similaridad de más del 90 % lo cual es bastante factible puesto que al principio se tomó como hipótesis que se obtendría una similaridad de al menos el 75 %. Hay que tomar en cuenta que se implementó el programa Chic para hacer la comparación cuando el objetivo solamente planteaba que se compararía solo con Rchic, es por eso que el porcentaje de similaridad aumentó.

Entonces sabiendo que se ha logrado más de lo que se esperaba, se puede decir que ahora existe una herramienta más con la cual el Análisis Estadístco Implicativo puede trabajar, en este caso sería en específico con la similaridad de Lerman. Se ha tomado en cuenta también que los matemáticos se sienten más familiarizados con el Software Matlab es por eso que se ha hecho la programación en un ambiente ya conocido por los matemáticos.

5. Relación con objetivos, hipótesis y problema

Con la creación del seudocódigo de la similaridad de Lerman entre variables binarios en el programa Matlab se cumplió el objetivo principal. Con el código hecho se procedió al análisis de las bases de muestra con el mismo obteniendo una similaridad mayor al 75 % cuando la hipótesis era tener al menos la similaridad del 75 %.

5.1. Sobre los objetivos específicos

Se propusieron 5 objetivos específicos en el Capítulo.- Problema de Investigación Sección 1.2.2 Objetivos específicos.

- El primer objetivo específico fue “Analizar la similaridad de Lerman” que se cumplió en el

Capítulo 2.- Marco Teórico, especificamente en la Sección 2.2 La similaridad de Lerman.

- El segundo objetivo específico fue “Determinar la formulación matemática de la La similaridad de Lerman” que se cumplió en el Capítulo 2.- Marco Teórico, especificamente en la Sección 2.5 Formulación matemática.

- El tercer objetivo específico fue “Elaborar el seudocódigo de la La similaridad de Lerman” que se cumplió en el Capítulo 4.- Marco de Análisis e Interpretación de resultados, especificamente en la Sección 4.1.1 Elaborar el seudocódigo de la La similaridad de Lerman.

- El cuarto objetivo específico fue “Programar la La similaridad de Lerman” que se cumplió en el Capítulo 4.- Marco de Análisis e Interpretación de resultados, especificamente en la Sección 4.1.3 Programación en MATLAB. Algoritmo del Análisi Estadístico Implicativo, Sección 4.1.4 Programación en MATLAB. Algoritmo del Análisi Estadístico Implicativo para una base de 7 variables y 20 filas, Sección 4.1.5 Programación en MATLAB. Algoritmo del Análisi Estadístico Implicativo para una base de 7 variables y 200 filas y la Sección 4.1.6 Programación en MATLAB, Algoritmo del Análisis Estadístico Implicativo para una bese aleatoria de m variables y n filas.

- El quinto objetivo específico fue “Estudiar la similaridad entre los resultados del programa Matlab (mCHIC) y RCHIC” que se cumplió en el Capítulo 4.- Marco de Análisis e Interpretación de resultados, especificamente en la Sección 4.2.1 Estadísticos descriptivos: Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC), la Sección 4.2.2 Gráfica de intervalos de Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC) y la Sección 4.3.1 T de una muestra: Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC).

6. Sobre el objetivo general

El objetivo general planteado fue “Programar en Matlab la similaridad de Lerman entre variables variables binarias y validarlas mediante la similaridad”. La programación de la similaridad de Lerman y su validacion se encuentran en las secciones 4.1.6, 4.2.1, 4.2.2 y 4.3.1.

- La Sección 4.1.6 Programación en MATLAB, Algoritmo del Análisis Estadístico Implicativo para una bese aleatoria de m variables y n filas, es la programación general de la similaridad de Lerman .

- La Sección 4.2.1 Estadísticos descriptivos: Matriz similaridad (mCHIC y RCHIC), Matriz

similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC); siendo los análisis de las bases obtenidas en mCHIC, RCHIC y CHIC.

- La Sección 4.2.2 Gráfica de intervalos de Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC);siendo los análisis de las bases obtenidas en mCHIC, RCHIC y CHIC.

- la Sección 4.3.1 T de una muestra: Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC); siendo los análisis de las bases obtenidas en mCHIC, RCHIC y CHIC.

7. Sobre la hipótesis

La hipóptesis planteada fue "Programar en Matlab la similaridad de Lerman entre variables variables binarias y validarlas mediante la similaridad al 75 %", se puede comprobar que la similaridad de los resultados sobrepasan el 75 % y comprobando que el programa M-Chic responde correctamente en las secciones:

- La Sección 4.1.6 Programación en MATLAB, Algoritmo del Análisis Estadístico Implicativo para una bese aleatoria de *m* variables y *n* filas.

- La Sección 4.2.1 Estadísticos descriptivos: Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC).

- La Sección 4.2.2 Gráfica de intervalos de Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC)..

- la Sección 4.3.1 T de una muestra: Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC).

El problema planteado fue "Se puede programar similaridad de Lerman entre variables binarias en Matlab", se comprobó la programación de la similaridad de Lerman en las secciones:

- La Sección 4.1.6 Programación en MATLAB, Algoritmo del Análisis Estadístico Implicativo para una bese aleatoria de *m* variables y *n* filas.

- La Sección 4.2.1 Estadísticos descriptivos: Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos

(mCHIC y CHIC).

- La Sección 4.2.1 Estadísticos descriptivos: Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC).

- La Sección 4.2.2 Gráfica de intervalos de Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC).

- la Sección 4.3.1 T de una muestra: Matriz similaridad (mCHIC y RCHIC), Matriz similaridad (mCHIC y CHIC), Gráfica (RCHIC y CHIC), Nodos (mCHIC y RCHIC), Nodos (mCHIC y CHIC).

Pero al no existir un cluster que gráfique la Matriz de similaridad en Matlab no se pudo comprobar la similaridad de Gráficas con RCHIC.

RECOMENDACIONES

Con base de los resultados obtenidos en el presente proyecto de investigación, a continuación se enumeraran algunas recomendaciones.

- Trabajar en la mejora de la programación de la Similaridad de Lerman, hecha para esta tesis, con el fin de que al momento de que se muestran los datos de las matrices de los diferentes niveles de similaridad, los decimales tengan un portentajo de similaridad aún más alto al momento de comparar con el programa original CHIC.

- Buscar una manera de implementar a la programación de la Similaridad de Lerman el dendograma (árbol de similaridad), el cual debe alimentarse con los datos de los nodos significativos.

- Se sabe que la programación de la Similaridad de Lerman tiene un alto porcentaje de similaridad con el programa CHIC, los dos trabajan con la distribución de Poisson pero se recomienda al usuario saber exactamente con que distribución va a trabajar para que obtenga mejores resultados ya sea en mCHIC, en RCHIC o CHIC.

BIBLIOGRAFÍA

BARRAGAN-PAZMIÑO, B.M.; PAZMIÑO-MAJI, R.A. "Literatura Cientifica sobre Analisis Estadistico Implicativo: Un mapeo sistematico de la decada que transcurre". *CIENCIA DIGITAL* [en línea], 2018, vol. 2. [Consulta: 23 abril 2022]. ISSN 2602-8085. Disponible en: http://cienciadigital.org/revistacienciadigital2/index.php/CienciaDigital/issue/view/13

COUTRIER, R., et al. "Statistical implicative analysis for educational data sets: 2 analysis with RCHIC" [en lí nea], 2015. [consulta: 1 diciembre 2022]. ISBN 978-84-608-3627-8. Disponible en: https://gredos.usal.es/handle/10366/127757

COUTURIER, R.; PAZMIÑO, R. "Use of Statistical Implicative Analysis in Complement of Item Analysis". *International Journal of Information and Education Technology*, vol. 6, no. 1 (2016), ISSN 2010-3689.

GRAS, R., et al. "El Analisis Estadistico Implicativo (ASI) en respuesta a problemas que le dieron origen" . 2009 pp. 3-50.

GREGORI HUERTA, P., et al. "On the probability distribution of the classical Gras implication index between two binary random variables". 2014.

IURATO, G. "The Implicative Statistical Analysis: An Interdisciplinary Paradigm". [en lí nea], 2012. [consulta: 3 noviembre 2021]. Disponible en: https://hal.archives-ouvertes.fr/hal-00750049

LERMAN, CI. "Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering". [en lí nea],2016. [consulta: 17 enero 2022]. Disponible en: https://books.google.com/books/about/Foundations_and_Methods_in_Combinatorial.html?hl=es&id=rJzWCwAAQBAJ

LÓPEZ-ROLDÁN, P.; FACHELLI, S. 2021. "Analisis de clasificacion". [en lí nea], 2016. [consulta: 9 noviembre 2021]. Disponible en: https://www.mdx.cat/handle/10503/113211

MARTÍNEZ GÓMEZ, M.; MARÍ BENLLOCH, M.D. "Distribucion Binomial". 2010.

MEDEROS, Y.G., et al. 2015. "Analisis estadistico implicativo en la identificacion de factores de riesgo en pacientes con cancer de pulmon". *MediSan* [en línea], 2015, vol. 19(08), [Consulta: 7 noviembre 2021]. ISSN 1029-3019. Disponible en: https://www.medigraphic.com/cgi-bin/new/resumen.cgi?IDARTICULO=60754

ORUS, P., et al. "Teoria y aplicaciones del Analisis Estadistico Implicativo: primera aproximacion en lengua hispana". 2009.

PAZMIÑO-MAJI, R.A., et al. "El Analisis Estadistico Implicativo y la mejora del aprendizaje en el marco de las Analiticas de Aprendizaje: Un Mapeo Sistematico de Literatura". 2018.

PEREZ, M.G., et al. "Cuasi-implicacion estadistica y determinacion automatica de clases de equivalencia en imagenes de resonancia magnetica de cerebro". *Revista Politecnica* [en linea], 2014, vol. 34(1), [Consulta: 15 diciembre 2021]. ISSN 2477-8990. Disponible en: https://revistapolitecnica.epn.edu.ec/ojs2/index.php/revista_politecnica2/article/view/260

SAGARO DEL CAMPO, N.M., et al. . "¿Por que emplear el analisis estadistico implicativo en los estudios de causalidad en salud?" *Revista Cubana de Informatica Medica* [en linea], 2019, vol. 11(1), [Consulta: 6 noviembre 2021]. ISSN 1684-1859. Disponible en: http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1684-18592019000100088&lng=es&nrm=iso&tlng=en

VALLS PLA, X., et al. "Diseño de un paquete R para el analisis estadistico implicativo". [en lí nea], 2014. [consulta: 15 diciembre 2021]. Disponible en: http://repositori.uji.es/xmlui/handle/10234/107441

ZAMORA, L., et al. "Conceptos fundamentales del Analisis Estadistico Implicativo (ASI) y su soporte computacional CHIC". *Contribuciones al ASI*. [en lí nea], 2009, vol. 4. [Consulta: 1 noviembre 2021]. Disponible en: http://repositori.uji.es/xmlui/bitstream/handle/10234/125568/asi4esp_v18_libro.pdf

UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 29 / 06 / 2023

INFORMACIÓN DEL AUTOR/A (S)
Nombres – Apellidos: Anabel Dejaneira Córdova Ruiz
INFORMACIÓN INSTITUCIONAL
Facultad: Ciencias
Carrera: Matemática
Título a optar: Matemática
f. Analista de Biblioteca responsable: Ing. Rafael Inty Salto Hidalgo

1082-DBRA-UPT-2023