UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

EAP. DE INGENIERÍA DE SISTEMAS

Implementación de una herramienta de análisis de riesgo de crédito basado en el modelo de rating de crédito, algoritmos genéticos y clustering jerárquico aglomerativo

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Henry Marcos RAMOS MARTINEZ

ASESOR

Marco Antonio SOBREVILLA CABEZUDO

Lima - Perú

2017

FICHA CATALOGRÁFICA

RAMOS MARTINEZ, Henry Marcos

IMPLEMENTACIÓN DE UNA HERRAMIENTA DE ANÁLISIS DE RIESGO DE CRÉDITO BASADO EN EL MODELO DE RATING DE CRÉDITO, ALGORITMOS GENÉTICOS Y CLUSTERING JERÁRQUICO AGLOMERATIVO

Ingenierías, Diseño y aplicación de nuevas tecnologías, Inteligencia Artificial
(Lima, Perú 2017)

Tesis, Facultad de Ingeniería de Sistemas e Informática, Pregrado, Universidad Nacional Mayor de San Marcos

Formato 21 x 29.7 cm, Páginas 93

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERIA DE SISTEMAS E INFORMÁTICA
ESCUELA ACADEMICO PROFESIONAL DE INGENIERÍA DE SISTEMAS

IMPLEMENTACIÓN DE UNA HERRAMIENTA DE ANÁLISIS DE RIESGO DE CRÉDITO BASADO EN EL MODELO DE RATING DE CRÉDITO, ALGORITMOS GENÉTICOS Y CLUSTERING JERÁRQUICO AGLOMERATIVO

Autor: RAMOS MARTINEZ, Henry Marcos

Asesor: SOBREVILLA CABEZUDO, Marco Antonio

Título: Tesis, para optar el Título Profesional de Ingeniero de Sistemas

Fecha: Septiembre del 2017

RESUMEN

En el presente trabajo de investigación se propone un método para generar modelos de clasificación de riesgo de crédito de acuerdo a la metodología de *rating* de crédito. La implementación de esta metodología requiere construir dos grandes bloques de análisis: (1) la construcción de un modelo de puntuaciones, y (2) la construcción de un modelo de agrupación de clases de riesgo. Para construir el modelo de *rating*, este trabajo propone el uso de dos técnicas de la inteligencia artificial: (1) el uso de algoritmos genéticos para determinar el modelo de puntuaciones óptimo, y (2) el uso de *clustering* jerárquico aglomerativo para la segmentación de los grupos de riesgo. Los resultados de la experimentación mostraron que la presente propuesta obtiene un buen indicador de poder de predicción (58.9%). Además, se comparó este modelo con el modelo de regresión logística (un conocido método de estimación estadística), teniendo la propuesta actual un mejor desempeño que el modelo logístico. Se concluye que las técnicas de inteligencia artificial usadas en este trabajo muestran un buen resultado para generar un modelo de *rating*, y tienen como ventaja la fácil interpretación de sus resultados por un experto humano.

Palabras Clave: *Rating* de crédito, Algoritmo Genético, *Clustering* Jerárquico Aglomerativo

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

SYSTEMS ENGINEERING AND INFORMATICS FACULTY
SYSTEMS ENGINEERING ACADEMIC PROFESSIONAL SCHOOL

A CREDIT RISK TOOL IMPLEMENTATION BASED ON THE CREDIT RATING MODEL, GENETIC ALGORITHMS AND AGGLOMERATIVE HIERARCHICAL CLUSTERING

Author: RAMOS MARTINEZ, Henry Marcos

Adviser: SOBREVILLA CABEZUDO, Marco Antonio

Title: Thesis, to apply for the Systems Engineer Professional Degree

Date: September 2017

ABSTRACT

This research work proposes a method for generating a credit risk classification model in accordance to the credit rating methodology. The implementation of this methodology requires the construction of two main analysis blocks: (1) the construction of a scoring model, and (2) the construction of a clustering model. For the credit rating model construction, this research work proposes the use of two artificial intelligence techniques: (1) genetic algorithms, to establish the optimal scoring model, and (2) agglomerative hierarchical clustering to determine the risk groups segmentation. The experiment results showed that this proposal have a good power statistics ratio (58.9%). In addition, this model was compared with the logistic regression model (a well-known statistical method), having the current proposal better performance than the logistic model. This paper concludes that the artificial intelligence techniques used in this research work show good performance at generating a credit rating model, and have the advantage of being easily interpreted by a human expert.

Keywords: Credit Rating, Genetic Algorithms, Agglomerative Hierarchical Clustering

JURADOS DE LA TESIS PARA OPTAR EL TÍTULO DE INGENIERO DE SISTEMAS

Autor: Henry Marcos Ramos Martinez

Tesis de Pregrado presentada a consideración del jurado revisor nombrado por la Escuela Académico Profesional de Ingeniería de Sistemas de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el grado académico de **Ingeniero de Sistemas**.

Aprobado por:

Dra. Luzmila Elisa Pró Concepción Presidente	Mg. Marcos Rivas Peña Miembro

Mg. Marco Antonio Sobrevilla Cabezudo
Asesor

Lima – Perú

2017

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

ÍNDICE DE ALGORITMOS

CAPÍTULO I
INTRODUCCIÓN

1.1 ANTECEDENTES

La ocupación más importante de una institución financiera es la actividad crediticia (Choudhry, 2012). Los servicios financieros, como los préstamos de dinero, han llegado a ser muy relevantes conforme a la evolución de la actividad bancaria (Samaniego, 2007). Pero si bien la actividad crediticia es la tarea que genera mayor rentabilidad a los esfuerzos de los bancos, es también la que le conlleva mayores riesgos. Y entre todos los riesgos que asume una empresa financiera, el principal es el riesgo de crédito (Choudhry, 2012).

El riesgo de crédito es la incertidumbre sobre la habilidad de un prestatario para hacer frente a su obligación de devolver una deuda. La probabilidad de incumplimiento refleja la evaluación de la posibilidad de que un prestatario incumpla con sus obligaciones contractuales de pago en un periodo de tiempo. La evaluación del riesgo de crédito consiste en lograr calcular dicha probabilidad de incumplimiento (Bhatia, 2006).

La gestión apropiada de los riesgos financieros es un asunto clave para reducir el riesgo de las ganancias de los bancos, así como para reducir el riesgo de que un banco se vuelva insolvente y sus depositantes no puedan ser reembolsados. Al ser la minimización del riesgo de crédito una de las tareas básicas de los bancos, éstos necesitan construir modelos confiables que detecten y predigan el incumplimiento de sus clientes de forma precisa (Van Gestel & Baesens, 2009).

Predecir y mitigar los eventos de incumplimiento financiero es el núcleo de una gestión apropiada del riesgo crediticio y esto puede ser ayudado a través del empleo de modelos cuantitativos adecuados, no obstante, sin excluir el juicio de un experto humano (Van Gestel & Baesens, 2009).

La literatura existente señala que una mala herramienta de evaluación de riesgo de crédito es la razón principal para que una empresa quiebre, por lo que es importante la definición de un modelo de evaluación correcto (Crook et al., 1994). Por este motivo, el riesgo de crédito ha sido por mucho tiempo de gran interés académico en las comunidades financieras y de negocios (Trujillo et al., 2014).

Para poder evaluar a las empresas, los modelos de análisis de riesgo toman como base la revisión de sus estados financieros. Los estados financieros consolidan la información económica-financiera de una empresa, permitiendo tener visibilidad sobre: las inversiones realizadas por la empresa; las obligaciones de la empresa y el monto financiado por los accionistas; el flujo de dinero de la empresa; y el nivel de liquidez y rentabilidad de ésta (Nakasone, 2001). Los estados financieros que manejan las empresas básicamente son el Balance general y el Estado de ganancias y pérdidas (Choudhry, 2012).

Una medida usual y recomendada para calcular el riesgo de una empresa es el uso de calificaciones, conocidas también como *ratings* financieros. Esta medida es útil para clasificar compañías según su riesgo de crédito, analizando sus estados financieros (García & García, 2010). El objetivo fundamental de una herramienta de *rating* es cuantificar y diferenciar el riesgo en una cartera de empresas a través de categorías de riesgo, prediciendo cuán probable es que un prestatario cumpla con sus obligaciones financieras (BBVA, 2005).

Un modelo de *rating* interno se desarrolla basado en datos históricos y el juicio de un experto de riesgo. La data histórica contiene toda la información disponible de los prestatarios de un banco en un momento en el pasado. Esto incluye la información cuantitativa, cualitativa y el comportamiento crediticio de sus deudores. Básicamente son dos bloques los que se definen (a partir de estas fuentes) para construir un modelo de *rating*: el desarrollo de un modelo de ordenación de las empresas basado en calificaciones, y la calibración del modelo a través de la asignación de grupos de *rating* en base a la homogeneidad de las puntuaciones obtenidas (Zaalberg, 2013).

Así como es importante para las instituciones financieras el manejo del riesgo de crédito a través de modelos de predicción de riesgo, también es importante para los estados (donde

operan dichos bancos) garantizar que estas instituciones manejen su riesgo de la manera correcta. En la mayoría de países, los servicios financieros y la banca son regulados por el estado. Las autoridades regulatorias y de supervisión otorgan permisos a las instituciones financieras para que puedan ofrecer sus servicios bancarios. Para poder recibir estos permisos, los bancos deben ser lo suficientemente seguros para sus clientes y para la economía del país. Si bien las autoridades supervisoras otorgan permisos de comercio a los bancos, también los mantienen en monitoreo y pueden intervenirlos e incluso retirarles el permiso de operación. La regulación bancaria tiene como objetivo reducir el riesgo de pago de los bancos, así como proteger los ahorros de los depositantes contra el riesgo excesivo que pueden tomar los bancos al otorgar créditos (Van Gestel & Baesens, 2009).

En el Perú, la Superintendencia de Banca, Seguros y AFP (SBS) es el organismo encargado de la regulación y supervisión de los Sistemas Financiero, de Seguros y del Sistema Privado de Pensiones (SPP). Su objetivo primordial es preservar los intereses de los depositantes, de los asegurados y de los afiliados al SPP ("Acerca de la SBS", s.f.).

La SBS mantiene dentro de sus objetivos la aplicación de metodologías para la medición y el análisis del riesgo crediticio. Para ello tiene como tarea el desarrollo de *ratings* financieros para los deudores comerciales del país, con el fin de supervisar y validar los modelos de riesgo que definen los bancos en el Perú.

1.2 DEFINICIÓN DEL PROBLEMA

1.2.1 PROBLEMA PRINCIPAL

El riesgo de crédito es un elemento que puede llevar a una institución financiera a la quiebra. La gestión de este riesgo es clave para minimizar las probabilidades de que un deudor incumpla sus obligaciones con la banca.

La definición de un modelo de alerta sirve a los bancos para la aproximación de los niveles de riesgo de una empresa, pero el diseño y construcción del mismo es complejo, requiriendo el manejo de mucha información financiera y un análisis preciso del riesgo de las instituciones evaluadas.

El modelo de *rating* es muy útil como herramienta para clasificar a un candidato a un préstamo, pero el desarrollo de este modelo tiene 2 bloques complejos de resolver: la construcción del modelo de puntuaciones, donde se deben determinar los ratios más relevantes para medir a un candidato y definir los pesos que debe tener cada ratio en el modelo; y la calibración del modelo, donde se deben definir los grupos de *rating* basándose en la homogeneidad de las puntuaciones de la data analizada, definiendo una probabilidad de incumplimiento a cada grupo de *rating* creado.

Ante esta problemática surge la pregunta, ¿es posible diseñar e implementar una solución que genere un modelo de clasificación del riesgo crediticio de clientes comerciales (de acuerdo al modelo de *rating* de crédito) utilizando técnicas de la inteligencia artificial?

1.2.2 PROBLEMAS ESPECÍFICOS

- ¿Es posible implementar un algoritmo genético que encuentre un modelo óptimo de puntuaciones de manera que éste valorice la capacidad crediticia de un solicitante de crédito?
- ¿Es posible implementar un algoritmo de *clustering* jerárquico que, basado en un modelo de puntuaciones, genere un esquema de clasificación de riesgo en grupos de *rating*?

1.3 DEFINICIÓN DE OBJETIVOS

1.3.1 OBJETIVO PRINCIPAL

Diseñar e implementar una solución, basada en la inteligencia artificial, que genere un modelo de clasificación del riesgo de crédito de clientes comerciales de acuerdo al modelo de *rating* de crédito.

1.3.2 OBJETIVOS ESPECÍFICOS

- Diseñar e implementar un algoritmo genético que permita hallar el modelo de puntuaciones óptimo, definiendo la relación de ratios más importantes y encontrando los pesos más adecuados para dichos ratios en el modelo de puntuación, de tal manera que éste valorice la capacidad crediticia de un prestatario.

- Diseñar e implementar un algoritmo de *clustering* jerárquico que, en base al resultado del algoritmo genético, clasifique a los deudores de acuerdo a la concentración de su riesgo y determine la cantidad final de grupos de riesgo del modelo.

1.4 JUSTIFICACIÓN

El propósito de la investigación se sustenta en lo importante que es a una entidad financiera tener conocimiento preciso y rápido para evaluar a un solicitante de crédito, reduciendo al mínimo el riesgo de que un posible prestatario incumpla con sus obligaciones futuras. Es por ello que la creación de un sistema que categorice el riesgo que existe en un deudor (para que la entidad tome una decisión más fina al aprobar un préstamo) es de gran relevancia en la administración del crédito en una empresa bancaria. Para ello se utilizarán técnicas de la inteligencia artificial buscando implementar en el sistema los criterios de evaluación de un analista, proveyendo así una alternativa distinta a la forma tradicional de construcción de modelos financieros.

El resultado de la investigación podría ser muy útil a las empresas del rubro financiero pues puede ayudar en el ahorro de tiempo en el proceso del análisis de riesgo, generando automáticamente el modelo de clasificación del riesgo de crédito, y también podría ayudar en la disminución de los errores humanos en dicho proceso.

1.5 ALCANCE

La solución se limita al desarrollo de una aplicación de escritorio que sirva de apoyo a instituciones financieras para el modelamiento del riesgo de sus deudores a través del modelo de *rating* de crédito.

La solución procesa deudores de tipo comerciales que presenten estados financieros anuales (balance y estado de ganancias y pérdidas). La solución no determinará el riesgo crediticio de personas naturales.

La solución no da la respuesta final de entregar crédito, sino que brinda un modelo de clasificación para que un cliente evaluado pueda ser ubicado en una categoría de riesgo.

La solución no prescinde de la aprobación final del experto analista de riesgos. Sólo sirve de apoyo al mismo.

La solución abarca la evaluación financiera cuantitativa del modelo de *rating* crediticio. No contempla la evaluación de variables cualitativas.

La solución será de apoyo para la SBS y será probada sobre un conjunto de datos de dicha institución, que contiene el comportamiento crediticio de 1459 comercios peruanos del año 2005.

1.6 ORGANIZACIÓN DE LA TESIS

El presente trabajo está organizado en 5 capítulos. El capítulo 1 (el presente) es donde se definen el problema, los objetivos, el alcance y justificación de la solución.

El capítulo 2 consiste en el marco teórico de la tesis y está compuesto de 3 grandes secciones: la primera trata acerca del modelo de *rating* de crédito; la segunda describe la teoría de los algoritmos genéticos, y la tercera explica la teoría de los algoritmos de *clustering* jerárquico aglomerativos.

El capítulo 3 presenta el estado del arte, donde se encuentran investigaciones relacionadas al presente trabajo de investigación.

El capítulo 4 presenta el aporte teórico y práctico de la solución, donde se detalla la propuesta a implementar y la justificación de las definiciones empleadas. Además, se presentan los resultados obtenidos y las características de las interfaces implementadas.

El capítulo 5 presenta las conclusiones del trabajo de investigación y la recomendación de trabajos futuros.

CAPÍTULO II
MARCO TEÓRICO

Este capítulo presenta el fundamento teórico para este estudio. El capítulo se divide en 3 partes. La primera describe el modelo de *rating* de crédito; la segunda revisa la teoría de algoritmos genéticos; y la tercera, la teoría de *clustering* jerárquico.

2.1 *RATING* DE CRÉDITO

2.1.1 INTRODUCCIÓN

La principal fuente de ingresos para la mayoría de bancos a nivel mundial es el ingreso por intereses. Los intereses se generan a través de la concesión de préstamos a personas o entidades comerciales (al receptor de un préstamo se le conoce como *contraparte*). Esta actividad crediticia se encuentra expuesta a riesgos, donde al riesgo de que una contraparte incumpla con las obligaciones del préstamo asumido, se le conoce como riesgo de crédito (Choudhry, 2012).

El riesgo de crédito es el riesgo más importante que enfrentan las instituciones financieras, y es por ello que éstas deben manejarlo de la manera más adecuada (Bhatia, 2006). Para lograrlo, los bancos utilizan *ratings* de crédito que clasifiquen a sus contrapartes. Estos *ratings* reflejan la capacidad crediticia de una contraparte y predicen cuán probable es que ésta cumpla con sus obligaciones financieras en su totalidad y a tiempo (Zaalberg, 2013). Una contraparte con una gran capacidad de cumplir sus compromisos financieros será calificada con un *rating* alto, mientras que una que esté al borde del incumplimiento tendrá un *rating* bajo.

Existen dos tipos de *ratings* de crédito: los *ratings* externos y los *ratings* internos. Los *ratings* externos son aquellos modelados y otorgados por una agencia independiente de

evaluación de *rating*, tales como Standard & Poor's[1], Moody's[2] y Fitch[3]. Por otro lado, en el enfoque de *rating* interno, los bancos desarrollan sus propios modelos para evaluar el riesgo de crédito de sus deudores.

Los *ratings* resultan de un análisis basado en información pública y privada, y contempla factores cuantitativos como también cualitativos. El análisis cuantitativo observa la estructura de la deuda de la contraparte, sus estados financieros, la hoja de balance y la información del sector; mientras que el análisis cualitativo analiza datos como la calidad de la gestión de la contraparte, su posición competitiva y el prospecto de crecimiento de la empresa (Van Gestel & Baesens, 2009).

2.1.2 MODELO DE RATING DE CRÉDITO

Un modelo de *rating* interno se desarrolla basado en 2 fuentes de información: el juicio de un experto de riesgo y datos históricos. La data histórica contiene muchos registros llamados *observaciones*. Una observación se define como la "radiografía" de toda la información disponible de una contraparte en un momento en el pasado. Esto incluye la información cuantitativa, cualitativa y el comportamiento crediticio de la contraparte en ese periodo (el comportamiento crediticio refleja si la contraparte cumplió o no una obligación). Esta información sirve como un conjunto de factores explicativos de la capacidad crediticia de una observación (Zaalberg, 2013).

La construcción de un modelo de *rating* consiste en 2 bloques principales: la construcción del modelo de ordenación y la calibración del modelo. La construcción del modelo de ordenación consiste en construir un modelo que ordene a las empresas mediante la asignación de puntuaciones (dichas puntuaciones se dan en función de la calidad crediticia). La calibración del modelo consiste en asignar categorías de *rating* en función al modelo de puntuaciones y calcular la probabilidad de incumplimiento de cada *rating* (BBVA, 2005).

2.1.3 PASOS PARA MODELAR EL RATING DE CRÉDITO

[1] Standard & Poor's. Disponible en https://www.standardandpoors.com. Accesado el 20/08/2016
[2] Moody's. Disponible en https://www.moodys.com. Accesado el 20/08/2016
[3] Fitch. Disponible en https://www.fitchratings.com. Accesado el 20/08/2016

En la Figura 1 se pueden distinguir cuatro pasos para el desarrollo de un modelo de *rating* (Zaalberg, 2013):

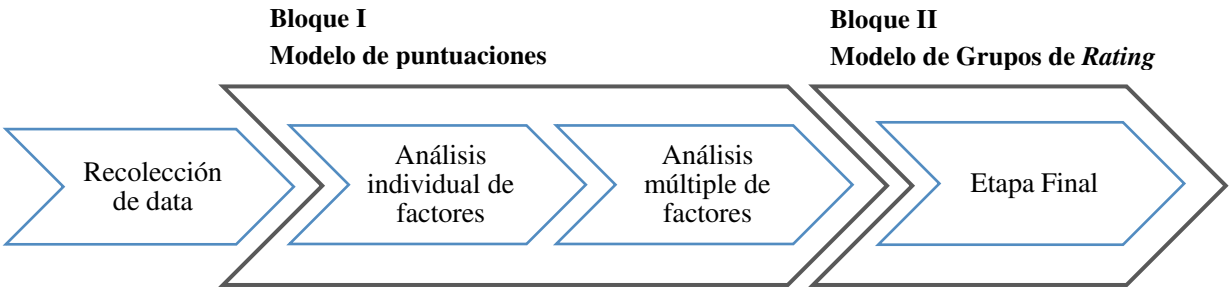


Figura 1: Pasos del desarrollo de un modelo de rating. Fuente: Adaptado de Zaalberg (2013)

El primer paso consiste en recoger la mayor cantidad de data posible, removiendo la información inútil y unificándola en un solo conjunto de datos. A cada característica que brinda esta información se le conoce también como *factor*.

El segundo paso consiste en realizar un análisis separado de cada factor para determinar el poder explicativo de dicho factor en el modelo. El objetivo del análisis individual de factores consiste en encontrar los factores más importantes que puedan explicar la capacidad crediticia de las contrapartidas.

En el tercer paso se toman los factores del paso anterior y se les determina un peso en la puntuación final. Una vez que se hayan definido los pesos por factor, se puede calcular la puntuación (también llamada *score*) para cada contrapartida. La puntuación final es la suma ponderada de las puntuaciones de cada factor.

Finalmente, en el cuarto paso los *scores* del modelo son relacionados a un grupo de riesgo (grupo de *rating*).

i. Recolección de data

El modelo de *rating* es desarrollado mediante un conjunto de observaciones. Una observación es una imagen de toda la información disponible de un banco en un momento determinado. Eso incluye la información cuantitativa, cualitativa y de comportamiento crediticio (Zaalberg, 2013).

Para la elaboración del modelo se recomienda dividir la muestra de datos en 2 partes. El 70% de las observaciones ser usan para el desarrollo del modelo, mientras que el resto se usa para la validación del modelo (Banasik et al., 1999).

ii. Análisis individual de factores

En el análisis individual de factores se prueba el poder explicativo que tiene cada factor cuantitativo y cualitativo por sí mismo. Eso significa que, por cada factor, se debe analizar el poder de predicción del modelo basado únicamente en dicho factor en específico. El objetivo principal del análisis individual de factores es encontrar una pequeña lista de los factores más importantes que expliquen la capacidad crediticia de las contrapartidas (Zaalberg, 2013).

El segundo objetivo de este análisis es transformar los valores de los factores en puntuaciones que sean fácilmente interpretables (por ejemplo, del 0 al 10). Esta transformación sirve como entrada para el siguiente paso del modelo de *rating*.

a) Transformación de los factores

Existen 3 razones para transformar los factores:

- Para que todos los factores pueden ser medidos bajo un mismo rango.
- Algunos factores tienen una distribución de valores que está relacionada inversamente a su capacidad crediticia. En el modelo de *score* final, los valores altos de estos factores deben corresponderse con puntuaciones bajas y viceversa.
- Los valores atípicos en la data pueden tener una influencia no justificada en los resultados de la puntuación. La aplicación de una transformación puede ayudar en la disminución de su efecto.

Para la transformación de los factores se utilizan técnicas de *bucketing*. El *bucketing* divide todos los posibles valores de un factor en grupos de valores. Un grupo (o *bucket*) contiene aquellas observaciones cuyos valores del factor analizado se encuentran entre los límites del *bucket*. El *bucketing* tiene la ventaja de relacionar los valores de un factor con la capacidad crediticia que posee, lo que ayuda a que la evaluación del experto sea más fácil (Man, 2014).

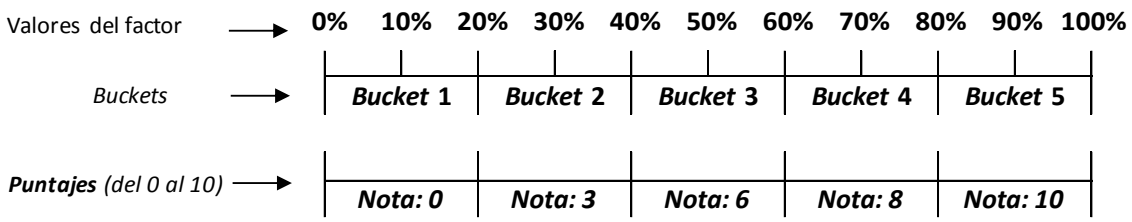


Figura 2: Ejemplo de bucketing con buckets de igual tamaño. Fuente: Adaptado de Man (2014)

En la Figura 2 se puede observar un ejemplo de *bucketing*. Este ejemplo transforma un factor (que en este caso toma valores desde el 0% hasta el 100%) en cinco partes o *buckets*. En este ejemplo cada *bucket* abarca un rango de valores de igual tamaño (los *buckets* han sido divididos de 20% en 20%). Por ejemplo, el "*Bucket* 1" abarca desde el valor 0% al valor 20% del factor analizado. Para cada *bucket* definido, se asigna un puntaje que transmita una valoración del riesgo crediticio que poseen los valores que comprenden dicho *bucket*. En este ejemplo, los puntajes del *bucketing* van desde el 0 hasta el 10. Por ejemplo, "el *Bucket* 1" tiene un puntaje de 0 porque los valores del 0% al 20% del factor analizado evidencian un mal comportamiento crediticio.

b) Poder de predicción

Una parte importante para el análisis individual de factores es la medida del poder de predicción de un factor individual. Esta medida se usa para la selección y transformación de los factores. La medida más popular es el poder estadístico (conocido también como *power statistic*) (Man, 2014).

El *power statistic* visualiza y cuantifica el poder predictivo de los factores individuales. La idea del modelo es que los peores valores de un factor correspondan con las observaciones de peor calidad.

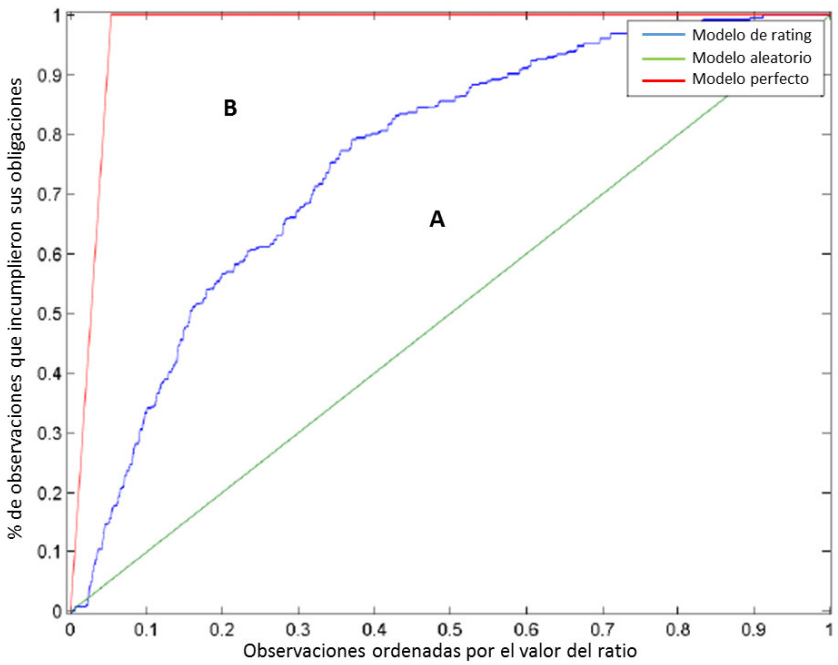


Figura 3: Gráfica de Power statistic. Fuente: Adaptado de Man (2014)

En la Figura 3, la línea roja (superior) representa el *modelo predictivo perfecto*. En el modelo perfecto todas las observaciones con mal comportamiento crediticio tienen asignados los peores valores del factor analizado. La línea verde (inferior) representa el *modelo aleatorio*. El modelo aleatorio representa un modelo que no guarda relación entre los valores de los factores con el número de observaciones "malas". La línea azul (intermedia) representa el *modelo de rating*.

Para construir esta gráfica, se clasifican todos los valores del factor de menor a mayor y se ubican en el eje-X. Luego se calcula el porcentaje acumulado de las observaciones "malas" y se ubican en el eje-Y.

Mientras más se acerque el modelo de *rating* (la línea azul) al modelo perfecto (la línea roja), mayor es el poder de predicción del factor. Mientras más se acerque el modelo de *rating* al modelo aleatorio (la línea verde), menor es el poder de predicción del factor.

El *power statistic* se calcula como el área entre la línea azul y la línea verde, dividido entre el área entre la línea roja y la línea verde. Mientras más alto sea el cociente, mayor es el poder de predicción. El cálculo del *power statistic* (PS) se resume en la Ecuación 1 (Man, 2014).

$$PS = \frac{\text{área } A}{\text{área } A + \text{área } B}$$

Ecuación 1: Cálculo del Power Statistic. Fuente: Man (2014)

iii. Análisis múltiple de factores

Durante el análisis múltiple de factores se determina cómo los factores individuales del paso anterior se incorporan en el modelo de *rating* final. El objetivo del análisis múltiple de factores es llegar a un modelo final basado en la combinación de los mejores factores explicativos, asignándoles pesos a estos factores y tomando en cuenta la redundancia entre ellos (Man, 2014). Con esta definición es posible calcular la contribución total de todos los factores de riesgo en un *score* final.

Luego del análisis individual y múltiple de factores, se calcula el *score* para cada observación del conjunto de datos. Mientras mayor sea el *score*, menor debe ser la probabilidad de incumplimiento de la contraparte (Zaalberg, 2013). El cálculo del *score* se ilustra en la Ecuación 2 (James, 2006):

$$score = c_1 v_1 + \cdots + c_N v_N,$$

Ecuación 2: Cálculo del score. Fuente: James (2006)

donde v_1 a v_N son los valores del puntaje de cada factor, y c_1 a c_N son los pesos asociados a cada factor. Estos valores se multiplican para llegar al *score* final.

Para validar el poder de predicción del modelo con el análisis de factores final, se calcula nuevamente la medida de *power statistic*. En este paso, para construir el indicador de *power statistic*, se clasifican todos los *scores* de las observaciones de menor a mayor y se ubican en el eje-X. Luego se calcula el porcentaje acumulado de las observaciones "malas" y se ubican en el eje-Y. Finalmente, se calcula el indicador de poder de predicción según lo descrito en la fórmula *PS* de la sección anterior (Man, 2014).

iv. Etapa final

El último paso consiste en agrupar a los *scores* (basándose en la homogeneidad de los *scores*) en grupos de *rating*, y finalmente asignarle una probabilidad de incumplimiento

(también conocida como *PD*) a cada grupo de *rating* (Zaalberg, 2013). Para asignar la PD a un grupo de *rating*, se cuentan las observaciones del grupo que cayeron en incumplimiento y se divide dicho número entre el número total de observaciones de dicho grupo de *rating*. En la Tabla 1 se muestra un ejemplo de agrupamiento de *ratings* con el cálculo de la PD correspondiente:

Grupo de *Rating*	**Criterio de agrupación**	**Observaciones**	**Observaciones que incumplieron**	**PD**
A	$Score = 10$	3	0	0 %
B	$8 \leq Score < 10$	11	1	9 %
C	$7 \leq Score < 8$	15	2	13 %
D	$6 \leq Score < 7$	26	5	19 %
E	$4 \leq Score < 6$	24	8	33 %
F	$0 \leq Score < 4$	21	10	48 %
Total		100	26	

Tabla 1: Ejemplo de grupos de rating. Fuente: Elaboración propia

En el ejemplo de la Tabla 1 se observa un modelo con 6 grupos de *rating* (de la A a la F), construidos con un conjunto de datos de 100 observaciones. La delimitación de estos grupos está en relación a los puntajes que han obtenido las observaciones con el modelo de puntuaciones. Por ejemplo, el grupo de *rating* "F" (que en este ejemplo representa al grupo de mayor riesgo) contiene las observaciones que hayan obtenido un puntaje entre 0 y 4 puntos. En este mismo ejemplo, se observa que el grupo "F" tiene una probabilidad de incumplimiento de 48%, ya que de las 21 observaciones que contiene, 10 cayeron en incumplimiento.

El poder de predicción final del modelo de *rating* se determina mediante el indicador *power statistic*. Si se observa que el modelo hallado se encuentra cercano al modelo perfecto, se puede concluir que el modelo explica adecuadamente el comportamiento (BBVA, 2005).

Como paso final de validación, el modelo debe ser testeado para determinar su robustez. Eso significa que se debe revisar si los pesos estimados no dependen mucho del conjunto de datos usado. Para ello se calcula nuevamente un modelo de *rating* pero con el conjunto de datos de validación. Si el resultado del modelo de validación no varía mucho respecto al modelo inicial, se puede concluir que el modelo hallado es suficientemente robusto (Zaalberg, 2013).

2.2 ALGORITMOS GENÉTICOS

2.2.1 INTRODUCCIÓN A LOS ALGORITMOS GENÉTICOS

Los algoritmos genéticos son una poderosa herramienta para solucionar problemas de búsqueda y optimización (Sivanandam & Deepa, 2008). Son algoritmos basados en la mecánica de selección natural y de la genética natural. Combinan la supervivencia del más apto entre estructuras de secuencias con un intercambio de información estructurado y aleatorizado (Goldberg, 1989).

2.2.2 PROCEDIMIENTO DE LOS ALGORITMOS GENÉTICOS

Un algoritmo genético maneja una población de posibles soluciones. Para estructurar dichas soluciones se deben definir dos partes importantes: la codificación de una solución en un cromosoma y la determinación de los operadores de reproducción. Los operadores de reproducción se aplican directamente en los cromosomas y se usan para realizar recombinaciones y mutaciones sobre las soluciones del problema (Sivanandam & Deepa, 2008).

En la reproducción de los cromosomas se realiza un proceso de selección, donde se compara cada individuo de la población a través de una función de aptitud. Cada cromosoma tiene asociado un valor correspondiente a la aptitud de la solución que representa. La aptitud corresponde a la evaluación de cuán buena es la solución del candidato representado. La solución óptima es la que tiene la aptitud máxima.

El proceso evolutivo de un algoritmo genético se observa en el Algoritmo 1:

1: Generar la población inicial
2: **Mientras** no se cumpla el criterio de terminación
3: Calcular la **aptitud** de cada individuo
4: **Seleccionar** individuos de la generación actual para reproducirlos
5: Crear descendencia mediante el **cruzamiento** de los seleccionados
6: **Mutar** la nueva descendencia, de acuerdo a una probabilidad
7: **Reemplazar** a los individuos antiguos con la nueva descendencia
8: Si el **criterio de terminación** está satisfecho, detener el algoritmo

9: Fin mientras

Algoritmo 1: Algoritmo Genético. Fuente: Sivanandam & Deepa (2008)

El algoritmo empieza generando una población inicial de cromosomas de forma aleatoria. Luego, el algoritmo genético recorre un proceso de iteraciones para hacer que la población evolucione. Cada iteración consiste de los siguientes pasos:

a) *Selección*: el primer paso consiste en seleccionar a los individuos para la reproducción. Esta selección se realiza aleatoriamente dependiendo de la aptitud relativa de los individuos, de tal manera que se escojan los mejores

b) *Reproducción*: en el segundo paso, la descendencia es engendrada por los individuos seleccionados. Para generar nuevos cromosomas, el algoritmo utiliza los operadores de cruzamiento y mutación.

c) *Reemplazo*: durante el último paso, los individuos de la población antigua son eliminados y reemplazados por los nuevos individuos

El algoritmo se detiene cuando la población converge hacia una solución óptima.

2.2.3. ELEMENTOS Y OPERADORES GENÉTICOS

i. **Representación:** consiste en codificar en una cadena de valores (llamada *cromosoma*) al conjunto de todos los parámetros (llamados también *genes*) que representan a la solución de un problema (Gestal et al., 2010).

ii. **Población:** En los algoritmos genéticos, los 2 aspectos más importantes a definir en la población son: el tamaño de la población y la generación de la población inicial (Sivanandam & Deepa, 2008).

El tamaño de la población dependerá de la complejidad del problema modelado. Su dimensionamiento es importante pues el tamaño de la población influye en el alcance de la exploración del espacio de búsqueda y también en el costo computacional.

En el caso de la generación de la población inicial, se frecuenta realizar una inicialización aleatoria. Pero también existen casos donde la inicialización de la población es realizada con algunas buenas soluciones ya conocidas (*siembra*).

iii. **Función de Aptitud:** En un algoritmo genético se debe poseer un método que indique si los individuos de la población representan o no buenas soluciones al problema planteado. De esto se encarga la función de aptitud, que establece una medida numérica de la bondad de una solución (Gestal et al., 2010).

iv. **Selección:** La selección es el proceso de escoger dos padres de la población para que sean cruzados. El propósito de la selección es escoger a los individuos más aptos esperando que su descendencia tenga mejores aptitudes. Este método escoge aleatoriamente cromosomas de la población de acuerdo a su función de aptitud. Mientras más alta sea la aptitud, más chances tendrá un individuo de ser seleccionado (Sivanandam & Deepa, 2008).

Los métodos de selección más usados son:

a) Selección aleatoria: Esta técnica selecciona aleatoriamente un padre de la población.

b) Selección por ruleta: En el proceso de selección por ruleta, el valor esperado de selección de un individuo es igual a su aptitud dividida entre la aptitud acumulada de toda la población. A cada individuo se le asigna un pedazo de la ruleta, donde el tamaño del sector es proporcional a la aptitud del individuo. Para elegir un individuo, se escoge un número aleatorio entre 0 y la suma total de aptitudes. El individuo elegido será aquel cuyos límites en la ruleta abarquen el número aleatorio generado.

c) Selección por torneo: En este método compiten N individuos aleatorios en un torneo (donde N es el total de los individuos que participan en el torneo. Este valor es predefinido en el algoritmo). El mejor individuo del torneo es aquel que tiene mejor aptitud. El torneo se repite hasta tener por lo menos 2 padres para generar una nueva descendencia.

- Elitismo: es un método que puede ser usado para eliminar la chance de que se pierdan buenas soluciones. Para aplicar el elitismo, los primeros mejores cromosomas son copiados a la nueva población, y el resto del proceso sigue tradicionalmente.

v. **Cruzamiento:** Cruzamiento es el proceso de tomar dos soluciones padres y producir un hijo a partir de ellos. El operador de cruzamiento es aplicado al conjunto de

padres con la esperanza de que se creen mejores descendencias (Sivanandam & Deepa, 2008). El cruzamiento es un operador que procede en tres pasos:

- El operador de selección toma 2 padres para la reproducción
- Se establece un punto de cruce aleatoriamente en un punto del cromosoma
- Se intercambian los valores de los cromosomas de acuerdo al punto de cruce

Las distintas técnicas de cruzamiento son: (Gestal et al., 2010)

a) *Cruzamiento de 1 punto*: Una vez seleccionados dos individuos se cortan sus cromosomas por un punto seleccionado aleatoriamente para generar dos segmentos diferenciados en cada uno de ellos: la cabeza y la cola. Se intercambian las colas entre los dos individuos para generar los nuevos descendientes.
b) *Cruzamiento de 2 puntos*: A diferencia del cruce de 1 punto, se realizan dos cortes en los padres. Para generar la descendencia se escoge el segmento central de uno de los padres y los segmentos laterales del otro padre. Generalizando, se pueden añadir más puntos de cruce dando lugar a algoritmos de cruce multipunto.
c) *Cruzamiento uniforme*: En este método, cada gen de la descendencia tiene las mismas probabilidades de pertenecer a uno u otro padre. La técnica implica la generación de una máscara de cruce con valores binarios para establecer el patrón de combinación de los genes de ambos padres.

   En la Figura 4 se muestra un ejemplo de cruzamiento uniforme. Si en una de las posiciones de la máscara hay un 1, el gen situado en esa posición en el primer padre será copiado en el descendiente. Si por el contrario hay un 0, el gen se copia del segundo padre.

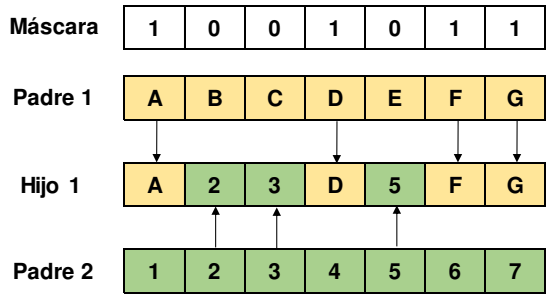


Figura 4: Ejemplo de cruce uniforme. Fuente: Gestal et al. (2010)

vi. **Mutación:** La mutación de un individuo provoca que alguno de sus genes varíe su valor de forma aleatoria. La mutación se aplica luego del cruzamiento, donde se muta un gen, con una probabilidad (Gestal et al., 2010).

vii. **Reemplazo:** Se han extraído dos padres de la población para reproducirse y han engendrado dos hijos, pero los cuatro no pueden retornar a la población, así que dos de ellos deben ser reemplazados. Existen dos tipos de métodos para el reemplazo: la actualización generacional y la de estado estable.

El esquema de actualización generacional consiste en producir "N" hijos desde una población de tamaño "N", para formar una nueva población en la siguiente generación. Esta nueva población de hijos reemplaza completamente a la anterior.

En la actualización de estado estable, los nuevos individuos son insertados en la población tan pronto son creados. El individuo reemplazado puede ser el peor padre o el peor miembro de la población (Sivanandam & Deepa, 2008).

viii. **Criterio de terminación del algoritmo genético:** Las siguientes son algunas condiciones de terminación: (Sivanandam & Deepa, 2008)

- *Máximo de generaciones*: el algoritmo se detiene cuando se ha alcanzado un número específico de generaciones.
- *Tiempo transcurrido*: el algoritmo se detiene cuando ha concluido un tiempo especificado.
- *Ningún cambio en la aptitud*: el algoritmo se detiene cuando no hay cambio en la mejor aptitud de la población en un número específico de generaciones.

2.3 *CLUSTERING* JERÁRQUICO AGLOMERATIVO

2.3.1 INTRODUCCIÓN AL ANÁLISIS DE CLUSTERS

El análisis de *clusters* es un proceso no supervisado que divide un conjunto de objetos en grupos homogéneos. El método de *data clustering* crea grupos de objetos (o *clusters*) de tal manera que los objetos en un *cluster* son muy similares, y los objetos en un *cluster* diferente son muy distintos (Gan et al., 2007).

El problema de *clustering* puede ser abordado usando una amplia variedad de métodos. Los algoritmos basados en distancias son unas de las técnicas más usadas para resolverlo (Aggarwal & Reddy, 2013).

2.3.2 ALGORITMOS BASADOS EN DISTANCIAS

Los algoritmos de *clustering* basados en distancias más estudiados son el *clustering* particional y el *clustering* jerárquico. Estos algoritmos han sido muy usados en un amplio rango de aplicaciones debido a su simplicidad y fácil implementación (Aggarwal & Reddy, 2013).

Los algoritmos de *clustering* particional tienen como objetivo descubrir agrupaciones al optimizar una función objetivo específica y mejorar iterativamente la calidad de las particiones. Estos algoritmos requieren ciertos parámetros de entrada para escoger los puntos *prototipo* que representan a cada *cluster*.

Por otro lado, los algoritmos de *clustering* jerárquico se aproximan al problema de *clustering* a través del desarrollo de una estructura de datos basada en un árbol binario, llamada dendograma. Una vez que el dendograma está construido, se puede escoger automáticamente el número correcto de *clusters* al dividir al árbol en diferentes niveles para obtener diferentes soluciones de *clustering*, sin necesidad de volver a procesar nuevamente el algoritmo de *clustering*. El *clustering* jerárquico puede ser logrado a través de dos diferentes maneras, llamadas *clustering* aglomerativo (o de "abajo hacia arriba") y *clustering* divisivo (o de "arriba hacia abajo").

Los métodos particionales necesitan ser provistos de un conjunto de semillas iniciales (o *clusters*), las que serán luego mejoradas iterativamente. Por otro lado, los métodos jerárquicos pueden empezar con cada dato individual y construir el *clustering* sin necesidad de un *input* previo.

2.3.3 CLUSTERING JERÁRQUICO AGLOMERATIVO

El *clustering* jerárquico aglomerativo inicia con cada objeto como un *cluster* individual. Luego repite el proceso de fusión del *cluster* más cercano de acuerdo a un criterio de similitud, hasta que la data esté agrupada en un solo *cluster* (Gan et al., 2007).

En el algoritmo 2 se explica el procedimiento del algoritmo. En la línea 1, se calcula la matriz de disimilitud para todos los puntos del conjunto de datos. La matriz de disimilitud es una matriz que contiene pares de índices que representan la proximidad de un conjunto de datos. En las líneas 3-4, se fusionan los pares de *clusters* más cercanos y se actualiza la matriz de disimilitud. Se quitan de la matriz de disimilitud las filas y columnas pertenecientes a los *clusters* antiguos y se añade a la matriz la fila y columna correspondiente al *cluster* nuevo. Y así sucesivamente, se llevan a cabo las operaciones de fusión en varias repeticiones actualizando la matriz de disimilitud. La línea 5 indica la condición de término del algoritmo (Aggarwal & Reddy, 2013).

1: Calcular la matriz de disimilitud entre cada uno de los puntos
2: **repetir**
3: Fusionar los *clusters* siendo $C_{aUb} = C_a \cup C_b$
4: Insertar una nueva fila y columna en la matriz, conteniendo las distancias entre el nuevo *cluster* C_{aUb} y el resto de *clusters*
5: **hasta que** Solo permanezca un único *cluster* máximo

Algoritmo 2: Clustering jerárquico aglomerativo. Fuente: Aggarwal & Reddy (2013)

Los algoritmos jerárquicos aglomerativos pueden ser subdivididos en diversos métodos de acuerdo a las diferentes medidas de distancia entre grupos. Los métodos más comunes se muestran en la Figura 5:

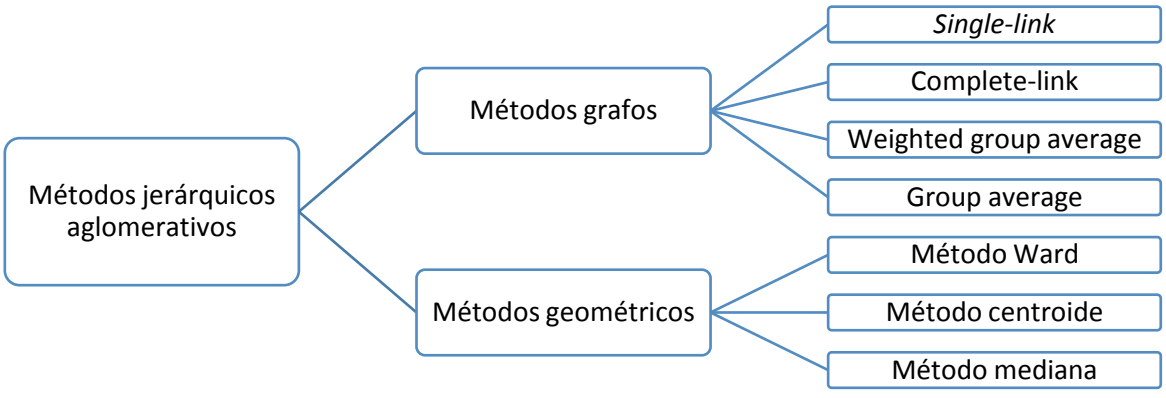


Figura 5: Métodos jerárquicos más comunes. Fuente: Adaptado de Gan et al. (2007)

i. Método de la distancia mínima (*Single-link*)

El método de la distancia mínima es uno de los métodos más simples de *clustering* jerárquico. Emplea la distancia del vecino más cercano para medir la disimilitud entre dos grupos (Gan et al., 2007).

Sean C_i, C_j y C_k tres grupos de puntos. Entonces la distancia entre C_k y (C_i U C_j) se obtiene de acuerdo a la Ecuación 3:

$$\boldsymbol{D\left(C_k, C_i \cup C_j\right) = \min\left\{D(C_k, C_i), D\left(C_k, C_j\right)\right\},}$$

Ecuación 3: Fórmula del método de la distancia mínima. Fuente: Gan et al. (2007)

donde $D(\cdot,\cdot)$ es la distancia entre dos *clusters*.

Se detalla un ejemplo del uso de este método en el ANEXO A.

ii. Método de la distancia máxima (*Complete-Link)*

A diferencia del método *single-link*, el método *complete-link* usa la distancia al vecino más alejado para medir la disimilitud entre dos grupos (Gan et al., 2007).

Sean C_i, C_j y C_k tres grupos de puntos. Entonces la distancia entre C_k y (C_i U C_j) se obtiene de acuerdo a la Ecuación 4:

$$\boldsymbol{D\left(C_k, C_i \cup C_j\right) = \max\left\{D(C_k, C_i), D\left(C_k, C_j\right)\right\},}$$

Ecuación 4: Fórmula del método de la distancia máxima. Fuente: Gan et al. (2007)

donde $D(\cdot,\cdot)$ es la distancia entre dos *clusters*.

Se detalla un ejemplo del uso de este método en el ANEXO A.

iii. Método de la distancia promedio no ponderada (*Weighted Group Average-Link)*

En el método de distancia promedio no ponderada, la distancia entre dos grupos se define como la media aritmética entre las distancias de los componentes de un *cluster* respecto a los del otro (Gan et al., 2007).

Sean C_i, C_j y C_k tres grupos de puntos. Entonces la distancia entre C_k y (C_i U C_j) se obtiene de acuerdo a la Ecuación 5:

$$\boldsymbol{D(C_k, C_i \cup C_j) = \frac{1}{2} D(C_k, C_i) + \frac{1}{2} D(C_k, C_j)}$$

Ecuación 5: Fórmula del método de la distancia promedio no ponderada. Fuente: Gan et al. (2007)

Se detalla un ejemplo del uso de este método en el ANEXO A.

iv. Método de la distancia promedio ponderada (*Group Average-Link)*

En el método de distancia promedio ponderada, la distancia entre dos grupos se define como el promedio ponderado de las distancias de los componentes de un *cluster* respecto a los del otro (Gan et al., 2007).

Sean C_1, C_2 y C_3 tres grupos de puntos. Entonces la distancia entre C_1 y (C_2 U C_3) se obtiene de acuerdo a la Ecuación 6:

$$\boldsymbol{D(C_1, C_2 \cup C_3) = \frac{1}{n_1(n_2 + n_3)} \cdot \left(\sum (C_1, C_2) + \sum (C_1, C_3) \right)}$$

Ecuación 6: Fórmula del método de la distancia promedio ponderada. Fuente: Gan et al. (2007)

donde:

$$\boldsymbol{\sum (C_i, C_j) = \sum_{x \in C_i, y \in C_j} d(x, y)}$$

y $\boldsymbol{n_i}$ es la cardinalidad. Ésta indica el total de puntos que contiene el *cluster* $\boldsymbol{C_i}$

Se detalla un ejemplo del uso de este método en el ANEXO A.

CAPÍTULO III

ESTADO DEL ARTE

En este capítulo se presentan los trabajos relacionados al presente trabajo de tesis. En la Sección 3.1 se muestra el método de investigación para explorar el estado del arte; en la Sección 3.2 se muestran 8 trabajos relacionados en del Estado del Arte; y en la Sección 3.3 se muestran las conclusiones de lo encontrado en los artículos investigados.

3.1 METODOLOGÍA DE INVESTIGACIÓN

Para la investigación de trabajos relacionados en el estado del arte, utilizó la metodología de revisión sistemática. Bajo este enfoque, se utilizó la siguiente pregunta de investigación:

¿Qué métodos/algoritmos/técnicas/enfoques se han usado para resolver el problema de rating de crédito?

El repositorio donde se realizó la búsqueda de trabajos relacionados fue *Google Scholar*[4]. Adicionalmente se tomaron los siguientes filtros para limitar la búsqueda:

- Sólo artículos que respondan a la pregunta de investigación
- Publicaciones desde el 2006 en adelante.
- Artículos con no menos de 5 hojas
- No considerar literatura gris

En la búsqueda se filtraron 30 resultados, de los que eligieron los 8 más representativos.

[4] Google Scholar. Disponible en https://scholar.google.com. Accesado el 20/08/2016

3.2 TRABAJOS RELACIONADOS

i. ***A study of Taiwan's issuer credit rating systems using support vector machines***

En el 2006, Chen & Shih (2006) realizaron un trabajo de investigación para desarrollar un modelo automático de clasificación de *rating* de crédito. Para ello se basaron en el uso de máquinas de vectores de soporte.

Al momento de la investigación, para resolver el problema de *rating* se habían utilizado técnicas como regresión lineal, análisis discriminante lineal multivariable, regresión probit, regresión logística y redes neuronales; pero las máquinas de vectores de soporte eran una técnica poco usada, siendo el objetivo de su trabajo probar la capacidad de este método para el problema de clasificación de *ratings* de crédito.

El conjunto de datos utilizado fue el de la corporación de *ratings* de Taiwan y del instituto de seguridad y futuro de Taiwan. Se utilizó data de tres años de historia de compañías bancarias.

Para comparar su modelo, los autores implementaron también una red neuronal *backpropagation*, concluyendo que la máquina de vectores de soporte arroja mejores resultados que el modelo de red neuronal. Los autores también concluyeron que el modelo arroja mejores resultados usando sólo el último año de historia del conjunto de datos, en lugar de usar tres años completos de información.

ii. ***Application of support vector machines to corporate credit rating prediction***

En este trabajo, Lee (2007) propuso un método alternativo a las técnicas tradicionales para la clasificación de *ratings* de crédito. Para ello aplicó máquinas de vectores de soporte intentando obtener un mejor poder de predicción en comparación a los métodos tradicionales de aprendizaje de máquina. Al momento de la investigación, las máquinas de vectores de soporte eran un método relativamente nuevo para abordar el problema de *ratings* de crédito.

El autor sostuvo que los métodos estadísticos tradicionalmente usados para el problema de *rating* de crédito son: regresión lineal, análisis discriminatorio multivariable, probit y

regresión logística; así como enfoques de inteligencia artificial, tales como: aprendizaje inductivo, redes neuronales y razonamiento basado en casos.

El conjunto de datos que usó fue obtenido del Servicio de información de Corea. Esta institución es una de las agencias de *rating* más prominentes de Corea. La muestra de datos histórica tenía la clasificación de *rating* de las compañías observadas.

Para comparar los resultados de la máquina de vectores de soporte, el autor implementó también una solución basada en una red neuronal *backpropagation*, análisis discriminatorio multivariable y razonamiento basado en casos. El trabajo concluyó que la máquina de vectores de soporte tuvo mucho mejores resultados que las otras técnicas de clasificación, respecto al poder de clasificación de la muestra evaluada. El autor también sostuvo que este método transforma problemas complejos en problemas más simples, y conducen al aprendizaje clasificatorio con una cantidad relativamente pequeña de data.

iii. *Modelling credit rating by fuzzy adaptive network*

En esta investigación, Jiao et al. (2007) propusieron un trabajo para desarrollar un modelo de *rating* de crédito tomando en cuenta el problema de variables lingüísticas y/o imprecisas en la data de entrada para la generación del modelo. Para ello utilizaron una red adaptativa difusa, donde combinaron la lógica difusa y las redes neuronales. Los autores sostuvieron que los enfoques tradicionales se basan solamente en los conceptos estadísticos generales, pero no pueden representar las expresiones lingüísticas ambiguas y por ello la representación de la data se simplifica. Para lograr tener tanto las habilidades lingüísticas y de aprendizaje, los autores propusieron usar la red adaptativa difusa, la cual es una combinación que aprovecha las habilidades de las redes neuronales y la lógica difusa, para el modelamiento de la capacidad crediticia de las compañías financieras.

El conjunto de datos que utilizaron fue una pequeña muestra de data de *ratings* de crédito del Comité de bancos de la ciudad de Taipei. La data tenía tanto variables cuantitativas como cualitativas. Por ejemplo, una variable cualitativa que aborda la lógica difusa en esta data es la experiencia del administrador, donde se tienen valores como: *muy bueno, bueno, regular, malo* y *muy malo*.

Los autores concluyeron que, gracias a la capacidad de aprendizaje y manejo de la ambigüedad de este enfoque, se obtuvieron buenos resultados de predicción con una tasa de error baja.

iv. *Credit rating by hybrid machine learning techniques*

En esta investigación, Tsai & Chen (2010) propusieron un trabajo para examinar el desempeño de predicción de cuatro tipos de modelos de aprendizaje híbrido en el problema de *rating* de crédito. Los autores sostuvieron que el enfoque híbrido va ganando más relevancia por sobre enfoques individuales de aprendizaje en los problemas de clasificación y predicción. El enfoque del modelo híbrido busca preparar la data en el primer algoritmo para que al aplicar el segundo algoritmo se tengan mejores resultados.

Un modelo de clasificación híbrida se puede combinar usando dos técnicas de aprendizaje de máquina: (1) combinando dos técnicas de clasificación, (2) combinando dos técnicas de *clusterización*, (3) combinando una técnica de *clustering* seguida de una de clasificación, y (4) combinando una técnica de clasificación seguida de una de *clustering*. Para las técnicas de clasificación se usaron árboles de decisión, clasificación Bayesiana, regresión logística y redes neuronales. Para la *clusterización* se usaron el algoritmo *K-means* y el algoritmo esperanza-maximización.

El conjunto de datos que se utilizó fue información real de un banco comercial de Taiwan. Los autores concluyeron que el modelo "clasificación+clasificación" arroja mejores resultados en la precisión de predicción y es el que tiene menor tasa de error. Los algoritmos que usaron para este modelo fueron regresión logística y redes neuronales.

v. *Genetic Algorithm Optimization for Selecting the Best Architecture of a Multi-Layer Perceptron Neural Network: A Credit Scoring Case*

En esta investigación, Correa et al. (2011) propusieron un trabajo para mejorar el poder de predicción de una red neuronal multi-capa en el problema de *scoring* de crédito. Para ello diseñaron un algoritmo genético (AG) que optimice la arquitectura de la red neuronal. La función objetivo del AG fue la curva ROC y las variables de entrada fueron el número de capas

ocultas y unidades, la función de activación, el uso o no de *bias*, y la decisión de establecer una conexión entre la capa inicial y final de la red neuronal.

El conjunto de datos que utilizaron fue de 125,557 clientes de un banco con tarjetas de crédito en el mes de Junio del 2009.

Los autores concluyeron que este método encontró la arquitectura óptima de una red neuronal multi-capa. Este método obtuvo mejores resultados que la regresión logística y una red neuronal estándar del software *SAS Enterprise Miner*.

vi. *A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach*

En esta investigación, Kim & Ahn (2012) realizaron un trabajo para proponer un nuevo tipo de máquina de vectores de soporte multi-clase (MSVM por sus siglas en inglés) para abordar el problema de *ratings* de crédito. Los autores sostuvieron que las máquinas de vectores de soporte (SVM por sus siglas en inglés) fueron originalmente ideadas para clasificaciones binarias (es decir, determinar sólo dos clases). Asimismo, comentaron que, para extender el alcance de este algoritmo a una clasificación de más de dos clases, se idearon los MSVMs, donde la técnica más conocida en este algoritmo es descomponer un problema multi-clase en varios sub-problemas binarios, de modo que se resuelvan a través de varios SVMs.

En este trabajo los autores propusieron un nuevo tipo de MSVM que toma en cuenta las características ordinales de un problema de clasificación ordinal manejando múltiples clases de tipo ordinal. Para ello propusieron el enfoque de máquina de vectores de soporte multi-clase ordinal (OMSVM por sus siglas en inglés). Este método es un algoritmo híbrido que combina un MSVM junto con el enfoque de particionamiento de pares ordinales (OPP por sus siglas en inglés). El propósito del algoritmo de OPP es considerar el orden de las clases mientras combina varios clasificadores binarios.

Los autores comentaron que el problema de *rating* de crédito es un problema de clasificación multi-clase ordinal, por lo que diseñaron el enfoque de OMSVM para poder mejorar el desempeño de los clasificadores bajo el enfoque de máquinas de vectores de soporte.

El conjunto de datos que se utilizó fue el de la Corporación de Información Nacional y Evaluación de Crédito de Corea. Utilizaron datos de compañías del 2002 con una clasificación de *rating* histórica.

Los autores concluyeron que, en el problema de clasificación multi-clase ordinal, el modelo de OMSVM obtiene mejores resultados que otros enfoques de MSVM, así como mejores resultados que otros enfoques de inteligencia artificial, tales como análisis discriminante múltiple, regresión multinomial logística, razonamiento basado en casos, y redes neuronales.

vii. *Feature selection in corporate credit rating prediction*

Hajek & Michalak (2013) propusieron un trabajo de investigación para preseleccionar las variables que intervienen en el desarrollo de un modelo de *rating* de crédito. Para ello utilizaron métodos de selección de variables (*feature selection*) como los métodos de filtro, métodos *wrapper*, y métodos integrados.

El experimento se realizó sobre conjuntos de datos de compañías de EE.UU y Europa. La base de EE.UU se obtuvo de las bases de datos de *Value Line* y *S&P*, y la base de Europa se obtuvo de las bases de datos de *Bloomberg* y *Capital IQ*. Los conjuntos de datos ya tenían clasificaciones de *rating* históricas de las compañías observadas.

Para la clasificación final de *ratings* los autores experimentaron con diversos métodos de clasificación: redes neuronales, máquinas de vectores de soporte, clasificador bayesiano ingenuo, clasificador discriminante lineal y el clasificador de la media más cercana.

Los autores concluyeron que realizar el método de *feature selection* previo a la clasificación de *ratings*, reduce la complejidad de la clasificación y mejora la precisión del proceso de clasificación. Asimismo, indicaron que la preselección de variables tuvo mejores resultados con dos tipos de algoritmos del método *wrapper* en comparación a los otros métodos de *feature selection*. Los 2 tipos son: la selección mixta de variables (MFS, por sus siglas en inglés) y la selección individual de variables (IFS, por sus siglas en inglés).

viii. ***Credit Rating Using Type-2 Fuzzy Neural Networks***

En su investigación, Abiyev (2014) propuso un trabajo para diseñar un modelo de *rating* de crédito. Para ello se basó en el uso de una red neuronal difusa tipo-2. El autor indicó que los atributos de un cliente de crédito se caracterizan por tener un alto grado de incertidumbre y falta de claridad en su información. Además, sustentó que el uso de un sistema difuso tipo-2 con redes neuronales logra tener buenos resultados en la problemática, pues la integración de un sistema neuronal y difuso tiene características de auto-aprendizaje y reduce el modelamiento de una data muy compleja.

El experimento se realizó sobre conjuntos de datos de crédito de Australia y Japón, obtenidos del repositorio UCI[5]. Para hacer un análisis comparativo, el autor contrastó los resultados de su propuesta versus modelos basados en máquinas de vectores de soporte y redes neuronales, obteniendo mejores resultados con su modelo en la precisión de la evaluación crediticia.

3.3 CONCLUSIONES DEL ESTADO DEL ARTE

Para resolver el problema de *rating* de crédito, muchos trabajos en el estado del arte han utilizado técnicas estadísticas y de inteligencia artificial, tales como regresión lineal, regresión logística, análisis discriminante múltiple, redes neuronales y máquinas de vectores de soporte. Para los casos mostrados, se utilizó data histórica de distintas agencias de evaluación de *ratings* de crédito, donde en todos los casos tuvo como variables de entrada variables cualitativas y cuantitativas y la clasificación de *rating* que una agencia especializada asignó previamente.

En estos casos fue poco estudiado el desarrollo de un modelo de *rating* sin el uso de un *rating* previo y basándose en la metodología de *rating* descrita en el capítulo 2: (a) desarrollar un modelo de puntuaciones en base a la selección/ponderación de las variables de entrada, y (b) posteriormente la clasificación de los *scores* en grupos de *rating*.

[5] Repositorio de data UCI. Disponible en http://archive.ics.uci.edu. Accesado el 20/08/2016

La mayoría utilizó los métodos de inteligencia artificial para aprender a clasificar *ratings* a partir de una clase de *rating* previamente conocida en la información histórica.

De lo observado en los trabajos relacionados, se concluye que los métodos de inteligencia artificial obtienen mejores resultados que los métodos tradicionales para abordar el problema de riesgo de crédito. También se concluye que el problema de *rating* de crédito ha sido ampliamente abordado a través de métodos de auto aprendizaje tomando como *input* las clasificaciones de *rating* definidas previamente por una agencia de *rating*. Como aporte del presente trabajo de investigación, abordaremos el problema de *rating* de crédito, pero bajo el enfoque de la metodología de *rating* donde no conocemos una clasificación previa de la información histórica.

CAPÍTULO IV
APORTE TEÓRICO Y PRÁCTICO

En este capítulo se presentan el aporte teórico y práctico de este trabajo de investigación. En la sección 4.1 se muestra el criterio de elección de los algoritmos empleados en la solución del problema; en la sección 4.2 se muestra el diseño de los métodos de inteligencia artificial usados para abordar el problema de *rating*; en la sección 4.3 se muestran los resultados de la experimentación realizada para construir el modelo; y en la sección 4.4 se muestra la interfaz gráfica implementada.

4.1 BENCHMARK DE LA SOLUCIÓN

El primer paso en esta tesis consiste en elegir los métodos de inteligencia artificial que ayuden a construir el modelo de puntuaciones óptimo y la agrupación de las clases de *rating*. Para ello, en la Tabla 2 se presentan los métodos más utilizados en el estado del arte para este tipo de problema, y 5 criterios de comparación entre ellos:

	Algoritmos	**Criterio 1**	**Criterio 2**	**Criterio 3**	**Criterio 4**	**Criterio 5**
Métodos individuales	Redes neuronales (multicapa, backpropagation)	X	X			
	Sistemas expertos (basado en reglas, en casos)	X				
	Máquinas de vectores de soporte	X				
Sistemas híbridos	Red neuronal MC + lógica difusa	X	X			
	Regresión logistica + Redes neuronales MC	X	X			
	Algoritmos genéticos + Red Neuronal MC	X	X	X		
Propuesta	**Sistema híbrido**: Algoritmo genético + *Cluster.*	X	X	X	X	X

Nota: la "X" indica que el criterio se cumple para el algoritmo evaluado

Tabla 2: Comparación de los métodos de Inteligencia Artificial más utilizados en el problema de rating. Fuente: Elaboración propia

Donde los criterios de comparación son los siguientes:

- *Naturaleza de Clasificación* (Criterio 1): se refiere a la capacidad del método de abordar problemas de clasificación

- *Naturaleza de exploración en un espacio de búsqueda grande* (Criterio 2): se refiere a la capacidad del método de abordar problemas de optimización

- *Capacidad de aprendizaje no supervisado* (Criterio 3): se refiere a la capacidad del método de aprender sin conocer los resultados esperados en la base de aprendizaje

- *Facilidad de implementación y adaptación al problema* (Criterio 4): se refiere a la sencillez del método para ser implementado y su flexibilidad para ser adaptado al problema de esta tesis

- *Facilidad de la lectura del resultado por un experto humano* (Criterio 5): se refiere a la capacidad del método de entregar un formato de solución que un experto humano pueda interpretar fácilmente, no siendo la solución una "caja negra" para él

Para abordar el problema de *rating* en este trabajo de investigación, se optó por utilizar un método híbrido que consiste en el uso de algoritmos genéticos y *clustering*. Los trabajos relacionados utilizan en su mayoría técnicas de clasificación bajo un aprendizaje supervisado, pero la naturaleza del problema de *rating* (bajo la metodología descrita en el capítulo 2) requiere que el algoritmo aprenda sin el conocimiento previo de los resultados deseados (características que sí cubren los métodos propuestos). Además, el resultado del método propuesto es fácilmente interpretable por un experto humano gracias al formato del entregable, a diferencia de la solución de otros trabajos actuales cuyo entregable resulta ser una "caja negra" para el experto, pudiendo generar incertidumbre en él.

Para elegir los métodos más adecuados de algoritmos genéticos y de *clustering*, se usó la metodología del proceso de jerarquía analítica (AHP por sus siglas en inglés). Como resultado de este análisis, se determinó que se utilizarían los métodos de "algoritmo genético simple" y de "*clustering* jerárquico aglomerativo". El detalle de este análisis se muestra en el ANEXO B.

En primer lugar, la metodología de *rating* requiere implementar un método de optimización para hallar el modelo de puntuaciones óptimo, el cual será abordado en esta tesis por el uso de algoritmos genéticos.

Para el algoritmo de agrupación se eligió el método de *clustering* jerárquico aglomerativo, ya que los métodos basados en distancias son uno de los más utilizados en los problemas de *clustering* (Aggarwal & Reddy, 2013). Además, se eligió este método pues no requiere el conocimiento previo del número de *clusters* que componen la solución (característica del problema de *rating*), a diferencia de otros métodos basados en distancias que sí lo requieren.

4.2 Diseño de la solución

Para generar un modelo de *rating*, la presente propuesta se basa en la implementación de la metodología de *rating* de crédito (detallada en el capítulo 2). Esta metodología requiere la solución de dos bloques: (1) la construcción de un modelo de puntuaciones, donde se deben determinar las variables más relevantes para medir a un candidato; y (2) la construcción de un modelo de agrupación, donde se deben definir grupos de riesgo basándose en el modelo de puntuaciones anterior. En la Figura 6 se muestra la estructura de la solución propuesta:

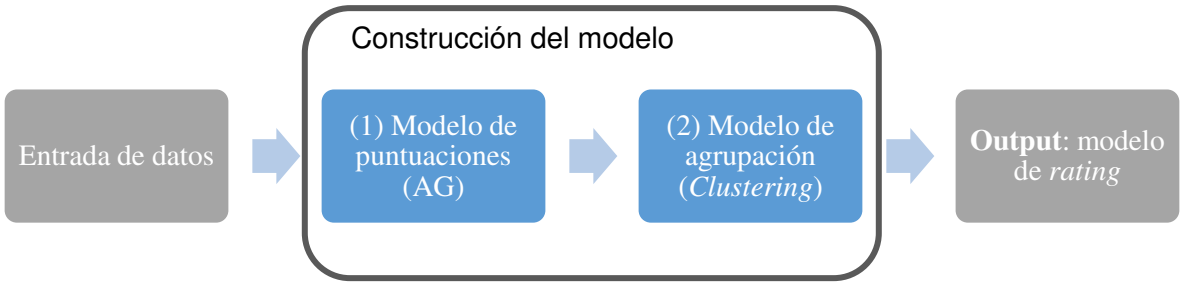


Figura 6: Arquitectura de la solución propuesta. Fuente: Elaboración propia

El primer bloque de construcción del modelo, que consiste en un problema de optimización de todas las variables y pesos para encontrar un modelo de puntuaciones óptimo, será abordado a través de la implementación de un algoritmo genético (AG). El diseño de este algoritmo se detalla en la sección 4.2.1. El segundo bloque del modelo, que consiste en identificar a los grupos de *rating* de acuerdo a la homogeneidad de sus puntuaciones, será abordado a través de un algoritmo de *clustering* jerárquico aglomerativo, cuyo diseño se detalla en la sección 4.2.2. Un ejemplo del proceso completo se muestra en el ANEXO C.

4.2.1 Diseño del Algoritmo Genético

En esta sección se muestra la configuración del problema de puntuaciones del modelo de *rating* a través de la definición de los elementos y parámetros de un algoritmo genético.

Previamente al diseño del algoritmo genético, se realizaron algunas definiciones necesarias para la construcción del modelo de puntuaciones. Estas definiciones corresponden al criterio utilizado para el *bucketing* del modelo de puntuaciones. El criterio utilizado en este caso corresponde a la recomendación del experto analista de riesgos (E.Bastante, comunicación personal, 19 de Octubre de 2010).

Para el *bucketing* se consideraron 6 *buckets*, y para los 5 puntos de corte de estos *buckets* se consideraron los siguientes criterios:

- El primer punto de corte corresponde al valor del percentil 5 de la muestra del ratio representado.
- El quinto punto de corte corresponde al valor del percentil 95 de la muestra del ratio representado.
- Para el resto de puntos de corte, se consideran *buckets* de igual tamaño desde el percentil 5 hasta el percentil 95.

Las puntuaciones que se asignaron a los *buckets* fueron del 0.0 al 10.0.

Por ejemplo, en la data de muestra, los valores del ratio "Patrimonio / Pasivos" (ver sección 4.3.1 para el detalle del conjunto de datos) abarca desde -0.91 hasta 59.98. Para este ratio, en la data el valor del percentil 5 es igual a 0.06 y el valor del percentil 95 es 1.87. Ya que por definición los *buckets* intermedios son de igual tamaño, en este ejemplo cada *bucket* tendrá un rango de 0.45. Bajo la definición dada sobre el *bucketing*, la distribución de este ratio se ilustra en la Figura 7:

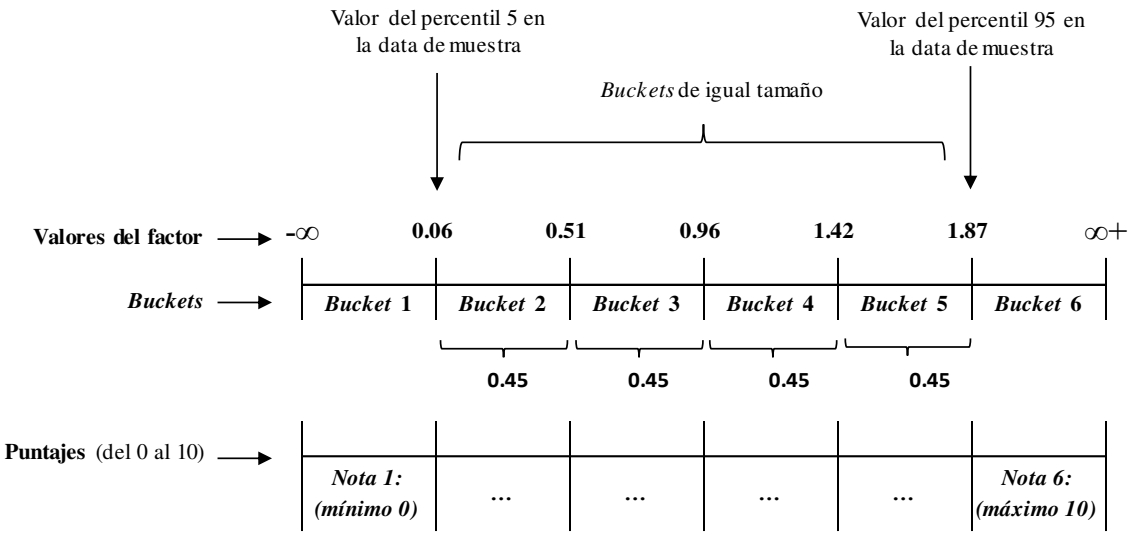


Figura 7: Ejemplo de la definición de bucketing. Fuente: Elaboración propia

A continuación, se describen los elementos y operadores genéticos usados en el algoritmo genético:

i. *Representación del cromosoma*: Para representar un cromosoma se definió una estructura segmentada en "N" secciones, donde cada sección contiene la información de un ratio (N es el número de ratios disponibles en la data de entrada). A su vez, cada ratio se compone de 3 bloques de información, que en total suman 8 variables:

(a) Bloque N°1: este bloque se compone de una variable que representa la participación del ratio en el modelo final. Esta variable contiene valores binarios, donde "0" significa que el ratio no participa en el modelo de puntuaciones, y "1" significa lo contrario.

(b) Bloque N°2: este bloque se compone de una variable que representa el peso del ratio en el modelo. Esta variable contiene valores enteros del 1 al 3, donde 1 significa importancia baja; 2, importancia media; y 3, importancia alta.

(c) Bloque N°3 (*bucketing*): este bloque se compone de seis variables que representan la calificación que tiene cada rango en el *bucketing* definido. Estas variables contienen valores reales del 0.0 al 10.0.

En la Figura 8 se muestra una representación gráfica de esta definición:

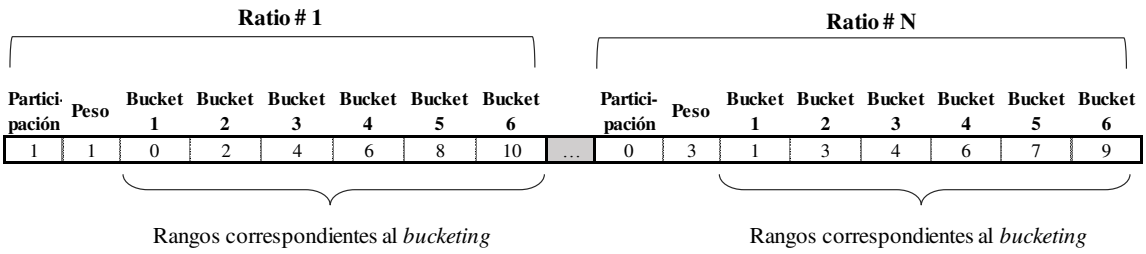


Figura 8: Representación gráfica del cromosoma planteado. Fuente: Elaboración propia

La puntuación obtenida a través del cromosoma planteado, se puede resumir en la Ecuación 7 (ecuación adaptada de la Ecuación 2):

$$\boldsymbol{score} = \sum_{i=1}^{N} \boldsymbol{Participación_i . Ponderación_i . Puntaje_i} ,$$

Ecuación 7: Cálculo del score según los parámetros del cromosoma propuesto. Fuente: Elaboración propia

donde la *Ponderación* $_i$ es igual a: $\frac{\boldsymbol{Peso_i}}{\sum_{j=1}^{N} \boldsymbol{Peso_j}}$. Los ratios que no participan en el modelo (es decir, cuyo valor en el cromosoma es 0) no se consideran en el cálculo de la ponderación.

ii. *Generación de la Población Inicial*: La población se generó de manera aleatoria con siembra (es decir, se introdujeron algunos cromosomas con "buenas características" para garantizar que algunos tengan un buen material genético). El tamaño de la población fue de 100 individuos y 5 de ellos fueron "sembrados". Estos criterios se usaron siguiendo la recomendación de la literatura (Sivanandam & Deepa, 2008).

Para la generación de valores aleatorios en la sección de *bucketing* del cromosoma, se consideró que la secuencia de las puntuaciones a través de los *buckets* guarde coherencia ascendente en correspondencia al incremento de los valores del ratio (es decir, a mayor valor del ratio, mayor calificación tendrán los *buckets* superiores, o viceversa). Para ello se estableció una regla donde el valor de un *bucket* superior debe ser siempre mayor al inferior, o viceversa. El detalle sobre la proporcionalidad de los ratios en el conjunto de datos se detalla en la sección 4.3.1. La Figura 9 ejemplifica este comportamiento en el cromosoma:

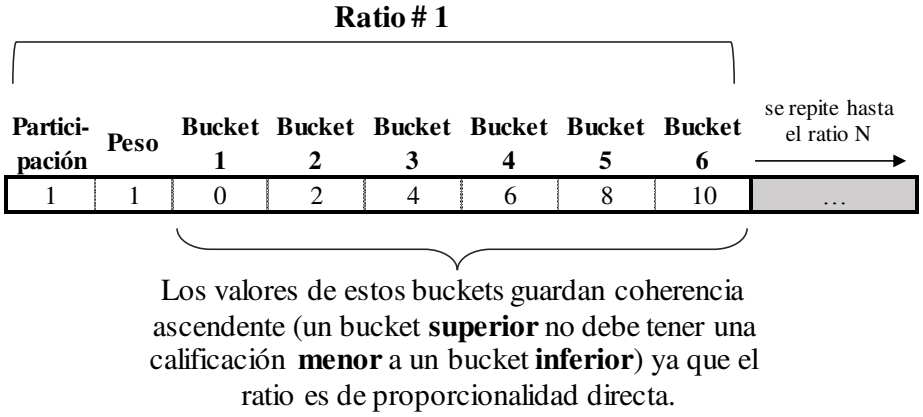


Figura 9: Ejemplo de coherencia en la asignación de valores del cromosoma. Fuente: Elaboración propia

iii. *Función de Aptitud*: Para la función de aptitud se utilizó el indicador de poder de predicción o *power prediction* (para más detalle, revisar la sección 2.1.3). Este indicador representa la capacidad de predicción que tiene el modelo de puntuaciones.

iv. *Selección*: La selección de los individuos se realizó a través de la selección por torneo. La literatura indica que este tipo de selección es muy usado y tiene mejor desempeño que otros métodos de selección (Sivanandam & Deepa, 2008).

v. *Cruzamiento*: El cruzamiento fue de tipo uniforme, donde los puntos de cruzamiento se generaron de manera aleatoria. La tasa de cruzamiento fue de 0.7, valor recomendado por la literatura (Man et al., 1996). Para el cruzamiento en la sección de *bucketing* del cromosoma se garantizó que la secuencia de calificaciones a través de los *buckets* guarde coherencia ascendente en correspondencia al incremento de los valores del ratio (es decir, a mayor valor del ratio, mayor calificación tendrán los *buckets* superiores, o viceversa). El detalle de la proporcionalidad de los ratios en el conjunto de datos se detalla en la sección 4.3.1. La Figura 10 muestra un ejemplo de cruzamiento:

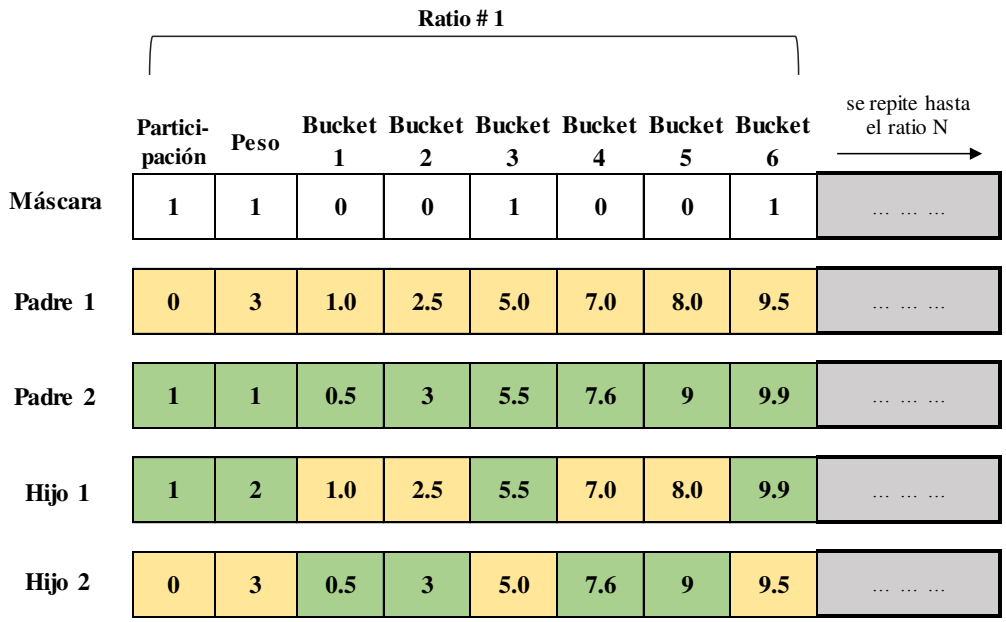


Figura 10: Ejemplo de cruzamiento uniforme en el ratio 1 del cromosoma. Fuente: Elaboración propia

En este ejemplo se observa el comportamiento del cruzamiento en la sección del ratio 1. El mismo comportamiento se repite en las siguientes secciones del cromosoma hasta el ratio N. En la ilustración se observa que el "Hijo 1" toma los valores del "Padre 1" cuando los valores del vector "máscara" tiene valores "0" y toma los valores del "Padre 2" cuando los valores del vector "máscara" tiene valores "1". El caso inverso sucede con el "Hijo 2".

Para garantizar la coherencia de la proporcionalidad en la sección del *bucketing*, se aplicó una regla que altere dinámicamente el vector "máscara" de tal manera que cuando un gen del "hijo" no cumpla con la proporcionalidad, se muta el valor de la máscara. Este comportamiento se muestra en la Figura 11, donde se observa que se muta el gen del *bucket* 5 para guardar la proporcionalidad en el "Hijo 2":

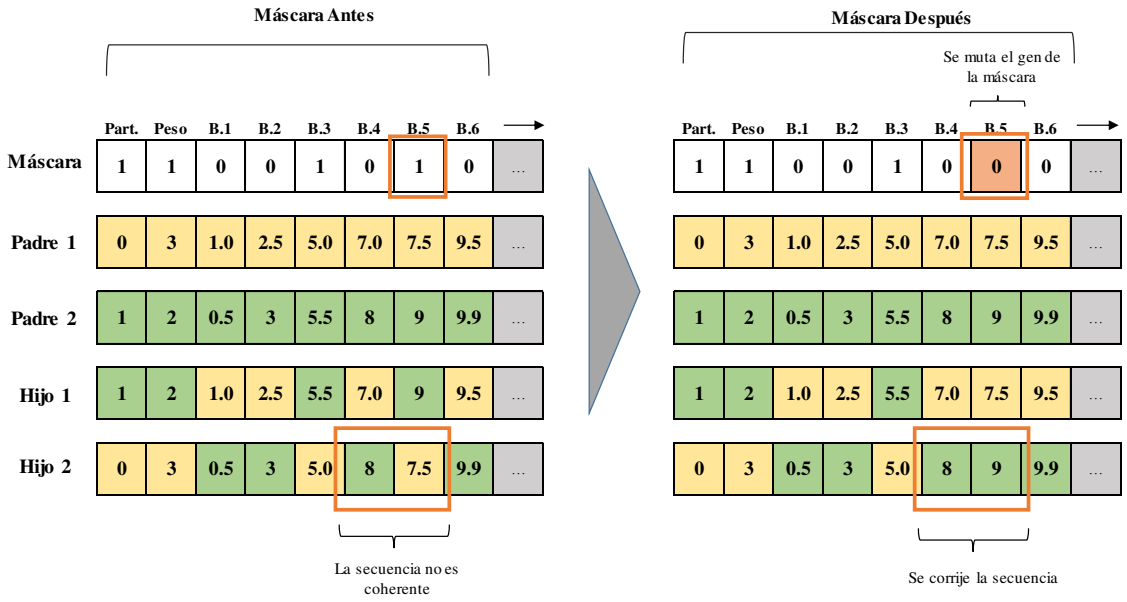


Figura 11: Ejemplo de mutación de los genes de la máscara para guardar la coherencia en los genes del bucketing de la descendencia. Fuente: Elaboración propia

vi. *Mutación*: Se aplicó mutación con una tasa de mutación de 0.01, según lo recomendado por la literatura (Gestal et al., 2010).

vii. *Reemplazo*: El tipo de reemplazo fue de actualización generacional (toda la población cambia de generación en generación). Para no perder buenas soluciones de la generación reemplazada, se utilizó el método de elitismo.

viii. *Criterio de terminación del algoritmo genético*: El algoritmo se detuvo cuando se alcanzaron las 1000 generaciones. Esto debido a que en la fase de experimentación se observó que con generaciones mayores a 1000 iteraciones se mantienen los mismos resultados.

4.2.2 Diseño del Algoritmo de Clustering Jerárquico Aglomerativo

En esta sección se muestran las definiciones consideradas para aplicar el algoritmo de *clustering* en la segunda etapa del modelo de *rating*: el agrupamiento de las puntuaciones de acuerdo a su homogeneidad.

Para aplicar el algoritmo de *clustering* jerárquico aglomerativo se deben realizar 4 definiciones: (i) establecer los *clusters* iniciales, (ii) establecer el criterio de distancia para

calcular la matriz inicial de disimilitud, (iii) establecer método de distancia para la fusión de los *clusters*, y (iv) definir el número de *clusters* final.

i. *Clusters* iniciales: Los *clusters* iniciales fueron los *scores* de cada observación de la data de entrada, ordenados de manera ascendente. El algoritmo los agrupó de tal manera que se identifiquen las aglomeraciones de *scores* por su cercanía. En la Figura 12 se muestra un ejemplo gráfico de la ubicación de los *clusters* iniciales, donde cada punto representa un *cluster* inicial. Cada *cluster* representa a cada observación del conjunto de datos y la información asociada a dichos *clusters* es su *score* respectivo.

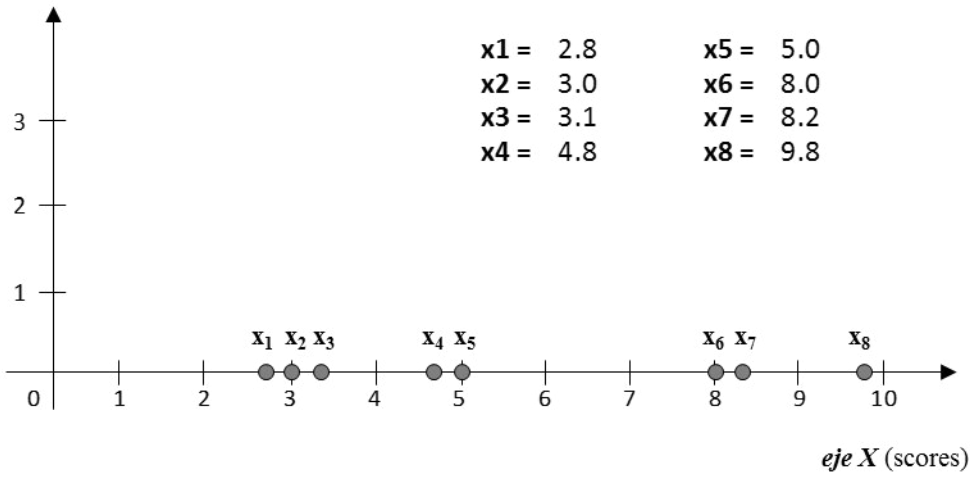


Figura 12: Ejemplo del diseño de los clusters iniciales. Fuente: Elaboración propia

ii. *Matriz de disimilitud inicial*: Para calcular la matriz inicial de disimilitud se usó la distancia Euclidiana entre los *scores*. Al ser la información unidimensional, la distancia consistirá en el valor absoluto de la resta simple de los *scores*. En la Figura 13 se muestra un ejemplo de matriz de disimilitud basada en el ejemplo de la Figura 12. Por ejemplo, la distancia inicial entre el *cluster* X_1 y el *cluster* X_2 es la resta de 3.0 - 2.8 = 0.2, lo que se representa en el cruce de la primera fila contra la segunda columna de la matriz de disimilitud.

	x1	x2	x3	x4	x5	x6	x7	x8
x1	0	0.2	0.3	2	2.2	5.2	5.4	7
x2	0.2	0	0.1	1.8	2	5	5.2	6.8
x3	0.3	0.1	0	1.7	1.9	4.9	5.1	6.7
x4	2	1.8	1.7	0	0.2	3.2	3.4	5
x5	2.2	2	1.9	0.2	0	3	3.2	4.8
x6	5.2	5	4.9	3.2	3	0	0.2	1.8
x7	5.4	5.2	5.1	3.4	3.2	0.2	0	1.6
x8	7	6.8	6.7	5	4.8	1.8	1.6	0

Figura 13: Ejemplo de matriz de disimilitud inicial. Fuente: Elaboración propia

iii. *Método de distancia para la fusión de clusters*: el método de distancia utilizado será la distancia promedio ponderada (*group average-link*). Este método hace que las distancias entre *clusters* se vean influenciadas por la concentración de cada uno de los puntos que componen a los *clusters*. Al utilizar este método para el modelo de *rating*, se busca que los *clusters* concentren en su centro los *scores* más comunes, de tal manera que la distribución de un grupo de *rating* contenga las casuísticas más repetidas de *scores* en su centro, y las casuísticas menos repetidas se encuentren en los extremos de la distribución. Al hacer esto, los grupos de *rating* estarían mejor delimitados y se evitaría el traslape entre ellos. Esto se representa en la Figura 14:

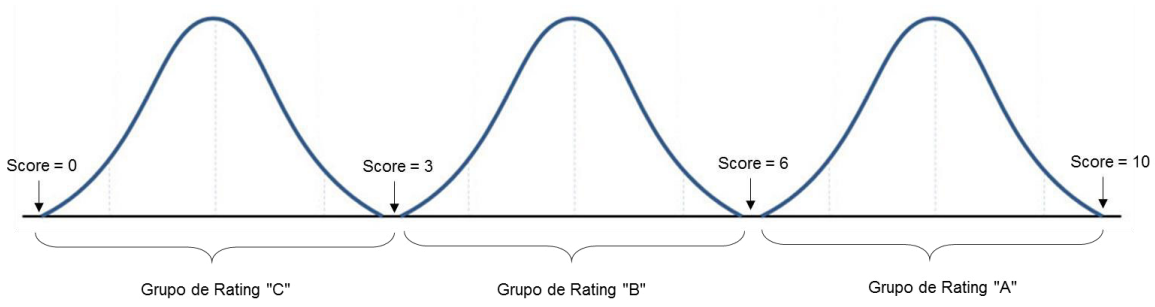


Figura 14: Objetivo de distribución de los grupos de rating a través del método de clustering utilizado. Fuente: Elaboración propia

iv. *Número de clusters final*: el algoritmo de *clustering* jerárquico aglomerativo no establece el número de *clusters* del modelo final, sino que construye todo el árbol de *clusters* permitiendo elegir el corte de *clusters* deseado. Para establecer este corte, construyó un algoritmo que recorra todos los cortes de *clusters*, desde un modelo con 2

clusters hasta 20 *clusters*, donde el criterio para elegir el número final de *clusters* es tomar el modelo de *clusters* que consiga maximizar el ratio de poder de predicción obtenido con dicho corte. Esto se representa en la Figura 15 en base al conjunto de datos de la Figura 12:

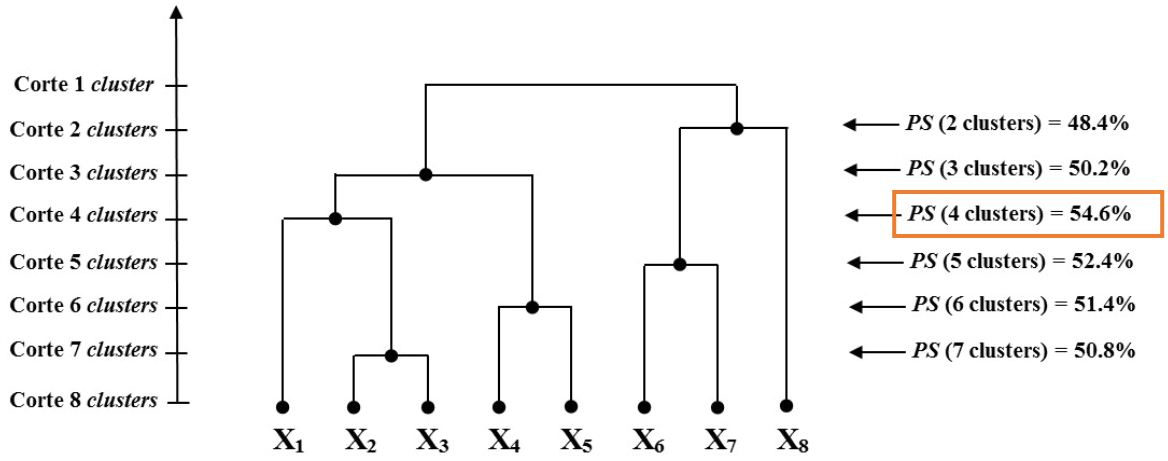


Figura 15: Ejemplo de dendograma para la elección del número de clusters final. Fuente: Elaboración propia

4.3 EXPERIMENTACIÓN

4.3.1 DATASET

El conjunto de datos utilizado para el presente trabajo de investigación fue otorgado por la Superintendencia de Banca, Seguros y AFP (SBS). La data es del año 2005 y está compuesta de 15 variables. Sólo se consideran para esta investigación variables de tipo cuantitativas. Para un modelo de *rating* también se consideran variables cualitativas, pero el conjunto de datos con el que se cuenta no tiene dicha información.

Las variables que conforman la información se presentan en la Tabla 3:

Tipo de Variable	N°	Nombre de la variable
Ratios de Solvencia	1	Patrimonio / Pasivos
	2	Pasivos / Activos
	3	EBITDA / Pasivo Corriente
	4	EBITDA / Pasivos
	5	EBITDA / Cargas Financieras
	6	Deuda / EBITDA
Ratios de Rentabilidad	7	Margen bruto / Ventas
	8	Utilidad Operativa / Ventas
	9	Utilidad Neta / Ventas
	10	Utilidad Operativa / Activos
	11	Utilidad Neta / Patrimonio
Ratios de Liquidez	12	Activo Corriente / Pasivo Corriente
	13	Pasivo Corriente / Pasivo
	14	Prueba ácida
Comportamiento de cumplimiento de deuda	15	Default (0: cumplió en el pago; 1: incumplió en el pago)

Tabla 3: Variables del conjunto de datos de entrada. Fuente: Elaboración propia

La información contiene 1459 casos de compañías a nivel nacional que tenían un préstamo con un banco peruano. En el 2005, de las 1459 empresas observadas, 35 incumplieron con el pago de la deuda asumida.

Para la generación y validación del modelo, se dividió la muestra en 2 partes: el 70% se destinó para la elaboración del modelo, mientras que el 30% restante se reservó para la validación de éste, según lo recomendado en la sección 2.1.3. Los registros de cada muestra se seleccionaron de manera aleatoria de la base de los 1459 casos.

En la Figura 16 se observa una muestra del conjunto de datos utilizado. En las primeras 14 columnas de la figura se observan los ratios descritos en la tabla anterior, y en la última columna se observa la columna "*Default*" donde se indica (con valores 1 ó 0) si la observación cumplió o no con el pago de su deuda en el periodo evaluado.

PAT/PT	PT/AT	EBITDA / Pas Corr	EBITDA / Pas Tot	EBITDA / Cargas Fin	Deuda / EBITDA	Margen bruto / ventas	UO / Ventas	UN / ventas	UO / activos	UN / Patrimonio	AC/PC	PC/PT	Prueba acida	default
-0.013	1.013	0.022	0.022	0.727	0.000	0.242	0.242	-0.665	0.008	-9.000	0.029	1.000	0.001	0
0.901	0.526	0.598	0.598	2.256	0.158	0.100	0.023	0.002	0.223	0.044	0.418	1.000	0.039	0
0.285	0.778	0.117	0.117	3.530	0.000	0.107	0.043	0.017	0.079	0.140	1.053	1.000	0.001	0
0.311	0.763	0.170	0.170	1.719	0.259	0.284	0.053	0.016	0.126	0.160	1.210	1.000	0.154	0
0.449	0.690	0.279	0.279	2.973	0.025	0.317	0.080	0.033	0.167	0.222	1.000	1.000	0.550	0
1.983	0.335	0.600	0.600	1.577	0.019	0.494	0.138	0.035	0.150	0.057	1.946	1.000	0.566	0
0.712	0.584	0.664	0.664	7.742	0.031	0.389	0.248	0.150	0.366	0.533	1.214	1.000	0.644	0
0.936	0.517	0.683	0.683	5.253	0.002	0.409	0.227	0.118	0.253	0.272	3.480	1.000	0.432	0
0.748	0.572	0.107	0.107	5.302	0.000	0.098	0.046	0.040	0.058	0.119	1.257	1.000	0.007	0
0.277	0.783	0.243	0.243	1.691	0.015	0.206	0.098	0.020	0.164	0.156	1.016	1.000	0.272	0
0.606	0.623	0.207	0.207	1.544	0.000	0.072	0.029	0.004	0.086	0.033	0.168	1.000	0.036	0
0.453	0.688	0.125	0.125	0.703	0.271	0.294	0.114	0.017	0.071	0.033	1.112	1.000	0.137	1

Figura 16: Muestra del conjunto de datos. Fuente: Elaboración propia

Al conjunto de datos se le ha aumentado un dato informativo sobre la proporcionalidad entre el valor del ratio y la calidad crediticia. Respecto a este comportamiento, se pueden clasificar a los ratios en dos condiciones: (a) cuando el valor del ratio es directamente proporcional a la calidad crediticia; y (b) cuando el valor del ratio es inversamente proporcional a la calidad crediticia.

Los ratios del primer tipo son los ratios N° 1, 3, 4, 5, 7, 8, 9, 10, 11, 12 y 14 de la Tabla 3. Los ratios del segundo tipo son los N° 2, 6 y 13.

En la Figura 17 se puede observar la representación de la proporcionalidad de los ratios a través de una fila (la fila 2) donde la proporcionalidad directa se representa con el signo "+" y la proporcionalidad inversa, con el signo "-". El algoritmo genético toma en cuenta esta condición al momento de asignar las calificaciones en la sección de *bucketing*.

	PAT/PT	PT/AT	EBITDA / Pas Corr	EBITDA / Pas Tot	EBITDA / Cargas Fin	Deuda / EBITDA	Margen bruto / ventas
→	+	-	+	+	+	-	+
	-0.013	1.013	0.022	0.022	0.727	0.000	0.242
	0.901	0.526	0.598	0.598	2.256	0.158	0.100
	0.285	0.778	0.117	0.117	3.530	0.000	0.107
	0.311	0.763	0.170	0.170	1.719	0.259	0.284
	0.449	0.690	0.279	0.279	2.973	0.025	0.317

Figura 17: Representación de la proporcionalidad de los ratios en el dataset. Fuente: Elaboración propia

4.3.2 BASELINE

Para comparar los resultados del modelo propuesto, se implementó adicionalmente una solución basada en el método de regresión logística. De acuerdo a la literatura, el modelo de regresión logística es uno de los más usados en el problema de riesgo de crédito (Tsai & Chen, 2010).

i. Introducción a la regresión logística:

La regresión logística es un enfoque matemático de modelamiento que se usa para describir la relación de una o más variables explicativas independientes con una variable dependiente dicotómica (es decir, la variable dependiente toma los valores verdadero o falso) (Kleinbaum & Klein, 2010). En las últimas décadas el modelo de regresión logística se ha convertido, en muchos campos, en el método estándar de análisis de este tipo de situaciones (Hosmer & Lemeshow, 2004).

En la Figura 18 se muestra la representación de la función logística, donde se describe la forma matemática sobre la cual se basa el modelo de regresión logística. Esta función, llamada ***f(z)***, es igual a 1 sobre 1 más ***e*** elevado a la menos *z* (Kleinbaum & Klein, 2010). Esto se muestra en la Ecuación 8:

$$\boldsymbol{f(z) = \frac{1}{1+e^{-z}}},$$

Ecuación 8: Función logística. Fuente: Kleinbaum & Klein (2010)

donde $\boldsymbol{z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$. Las *Xs* son las variables explicativas independientes, y α y los β_i son los términos constantes que representan parámetros desconocidos. Estos parámetros serán estimados en base a la data de las variables independientes y la variable dependiente de un grupo de observaciones.

En la Figura 18 se han trazado los valores de la función mientras ***z*** varía desde -∞ hasta +∞. Se observa, en el lado izquierdo de la figura, que mientras z se acerca a -∞, la función logística ***f(z)*** es igual a 0. Al lado derecho, cuando z se acerca a +∞, la función ***f(z)*** es igual a 1.

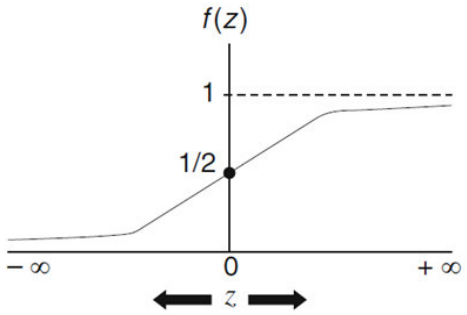


Figura 18: Ilustración de la función logística. Fuente: Kleinbaum & Klein (2010)

El hecho de que la función logística ***f(z)*** varíe entre 0 y 1 es la principal razón de que el modelo logístico sea tan popular. El modelo está diseñado para describir una probabilidad que siempre está en un número entre 0 y 1. En términos del análisis de riesgo de crédito, esta probabilidad indica el riesgo de que un individuo incumpla con un préstamo asumido.

Otra razón de la popularidad del uso de la regresión logística se deriva de la alargada *forma de S* de la función logística. Esta forma de la función indica que el efecto de *z* sobre el riesgo de un individuo es mínimo para un z pequeño, hasta que un umbral es alcanzado. El riesgo luego incrementa rápidamente en un cierto rango de valores de *z* intermedios, y luego se mantiene extremadamente alto alrededor de 1 una vez que *z* ha conseguido ser suficientemente grande. Se considera que este modelo en *forma de S* es ampliamente aplicable a la naturaleza multivariable de un problema de riesgo de crédito.

Usualmente se utiliza el *valor-p* para evaluar la significancia estadística de cada variable explicativa en el resultado del modelo logístico. El valor-p mide que una variable independiente tenga un aporte significativo al modelo, determinando si los cambios en estas variables predictoras tienen relación con los cambios en la variable de respuesta. Por convención, a las variables con valores-p que excedan a 0.05, se les considera que no son lo suficientemente fuertes como para predecir los cambios en la variable de respuesta (Thisted, 1998).

ii. Resultados del modelo con regresión logística:

Se implementó un modelo de predicción de riesgo de crédito usando el modelo de regresión logística para compararlo con el resultado obtenido con la propuesta de esta tesis.

Para construir el modelo de regresión logística, se utilizó el mismo conjunto de datos utilizado en esta investigación (ver sección 4.3.1). El modelo se construyó con el software estadístico **R**[6]. Para evaluar la significancia de las variables independientes en el modelo logístico, se utilizó la prueba de significancia del valor-p.

En el modelo de regresión logística, la variable dependiente fue la columna "default" y las variables explicativas fueron los 14 ratios del conjunto de datos. Tras aplicar la prueba de significancia a las variables, se obtuvo que solamente la variable "Ebitda/Cargas Financieras" es estadísticamente significativa en el modelo logístico. Los resultados se muestran en la Tabla 4, donde principalmente se muestran las constantes halladas (α y β) y el valor-P de la variable explicativa, donde se comprueba que la variable es significativa pues es menor a 0.05:

Coeficientes	Constantes	Desviación Estándar	Valor Z	Valor-P (Pr(>|z|))
(α) -> Intercepto	-3.7318737	0.1737191	-21.482	< 2e-16
(β) -> EBITDA/Cargas Financieras	-0.0005439	0.0002015	-2.698	0.00697

Tabla 4: Resultados de la regresión logística. Fuente: Elaboración Propia

Se aplicó el indicador de *power statistic* a este modelo, para determinar su poder de predicción. Ya que el modelo sólo depende de una variable, para graficar el indicador se ordenó de menor a mayor los valores del ratio "EBITDA/Cargas Financieras" de cada observación ubicándolos en el Eje-X, y en el Eje-Y se ubicaron los valores "Default" relacionados. Con esta medición, el indicador de *power statistic* de este modelo fue de 37.7%.

4.3.3 RESULTADOS

En la Tabla 5 se muestran los resultados del algoritmo genético para el modelo de puntuaciones del modelo de *rating*. Utilizando la muestra de entrenamiento, los resultados del algoritmo arrojan que el modelo se compone de 5 ratios (el N° 2, 5, 9, 11 y 13 de la Tabla 3). El algoritmo ha definido un peso de importancia en el modelo para cada ratio y las puntuaciones correspondientes para el *bucketing* (*buckets* del 1 al 6).

[6] Software R. Disponible en https://www.r-project.org. Accesado el 20/08/2016

N°Ratio	Ratio	Peso	*Bucket* 1	*Bucket* 2	*Bucket* 3	*Bucket* 4	*Bucket* 5	*Bucket* 6
2	Pasivos / Activos	3	6.46	6.03	5.88	5.86	5.04	0.58
5	EBITDA / Cargas Finan.	2	0.76	4.98	5.18	5.22	9.38	9.47
9	Utilidad Neta / Ventas	3	2.03	3.22	4.05	4.14	9.24	9.74
11	Utilidad Neta / Patrimonio	3	1.34	1.34	4.49	8.56	9.16	9.31
13	Pasivo Corriente / Pasivo	1	6.11	5.57	3.89	3.73	2.75	1.71

Tabla 5: Modelo de puntuaciones hallado por el algoritmo genético. Fuente: Elaboración propia

En la Tabla 6 se muestran los resultados obtenidos con el algoritmo de *clustering* jerárquico aglomerativo para hallar los grupos de *rating* del modelo. El algoritmo determinó 8 grupos de riesgo, definiendo los límites de los grupos en función a los *scores* del modelo de puntuaciones (ver columna "Nota mínima" y "Nota Máxima").

Grupo	Cantidad de Observaciones	Cantidad de incumplidos	% Probabilidad incumplimiento	Nota Mínima	Nota Máxima
A	81	0	0.00%	7.62	10.00
B	126	0	0.00%	6.53	7.62
C	426	4	0.94%	5.54	6.53
D	133	4	3.01%	4.70	5.54
E	171	11	6.43%	4.00	4.70
F	37	3	8.11%	3.02	4.00
G	22	2	9.09%	2.27	3.02
H	26	3	11.54%	0.00	2.27

Tabla 6: Modelo de grupos de rating hallado por el algoritmo de clustering. Fuente: Elaboración propia

El poder de predicción (*power statistic*) obtenido con estos modelos fue de 58.9%. El experto de riesgo indicó que un modelo de *rating* es bueno cuando el poder de predicción es mayor a 50% (Bastante, 2010), con lo que se observa que los modelos hallados por los algoritmos planteados son buenos y superaron los valores esperados en un análisis de *rating*.

Además, se probó la muestra de validación para evaluar la robustez del modelo generado. La prueba consistió en 2 pasos: primero, determinar el *score* y grupo de *rating* de cada registro de la muestra en base al modelo que construyó el algoritmo; y segundo, calcular el ratio de *power statistic* en base a la clasificación del paso 1 para analizar el performance de

clasificación de los individuos que logró el modelo. Con esto, se esperaba ver que el valor del ratio de *power statistic* de la muestra guarde la misma proporción que el valor de *power statistic* del modelo de entrenamiento. Se hizo el ejercicio y se obtuvo que este ratio obtuvo un valor de 59.4%, versus el 58.9% que se obtuvo en el modelo de entrenamiento, con lo que se observó que el modelo generado es robusto.

Adicionalmente, para hacer un análisis comparativo de la solución, se comparó con el método de regresión logística descrito en la sección anterior. Con el modelo logístico se obtuvo un poder de predicción de 37.7%, con lo que se observa que el modelo planteado en este trabajo de investigación es mejor al tradicional método logístico.

4.4 INTERFAZ GRÁFICA

La implementación de la solución se realizó sobre una interfaz gráfica desarrollada en Java para facilitar el ingreso de los datos de entrada y la visualización de los resultados. A continuación, se muestran las dos pantallas correspondientes a los dos módulos del modelo de *rating* planteado: (a) la pantalla donde se ejecuta el algoritmo genético y (b) la pantalla donde se ejecuta el algoritmo de *clustering* jerárquico; habiéndose diseñado para estas funcionalidades los respectivos casos de uso que se detallan en el ANEXO D.

(i) Módulo del modelo de puntuaciones: en esta pantalla se tienen 3 secciones: (A) la sección superior izquierda, donde se podrán cargar los datos de entrada y ejecutar el algoritmo genético; (B) la sección inferior izquierda, donde se visualiza el resultado del algoritmo con el modelo de puntuaciones; y (C) la sección derecha, donde se observa el resultado del poder de predicción a través de la gráfica de *power statistic*.

En la sección superior izquierda, el usuario tiene la opción “Cargar Datos” para seleccionar el archivo Excel que contiene el conjunto de datos que procesarán los algoritmos. Adicionalmente está la opción “Generar Solución”, que sirve para dar inicio al algoritmo genético. Así mismo, el usuario tiene la opción de cargar un modelo ya generado por el algoritmo genético en una ocasión previa. Una vez procesado el algoritmo genético, la interfaz muestra en la sección inferior izquierda los resultados del modelo de puntuaciones (con los ratios elegidos, los pesos de cada ratio y las puntuaciones del *bucketing*) y el índice de poder de predicción, y en la sección derecha se ilustra la figura de *power statistic* que

representa gráficamente el índice de poder de predicción del modelo. Finalmente, el usuario tiene la opción de guardar esta solución y/o seguir a la siguiente sección de generación del modelo de agrupación. En la Figura 19 se muestra un ejemplo de esta interfaz:

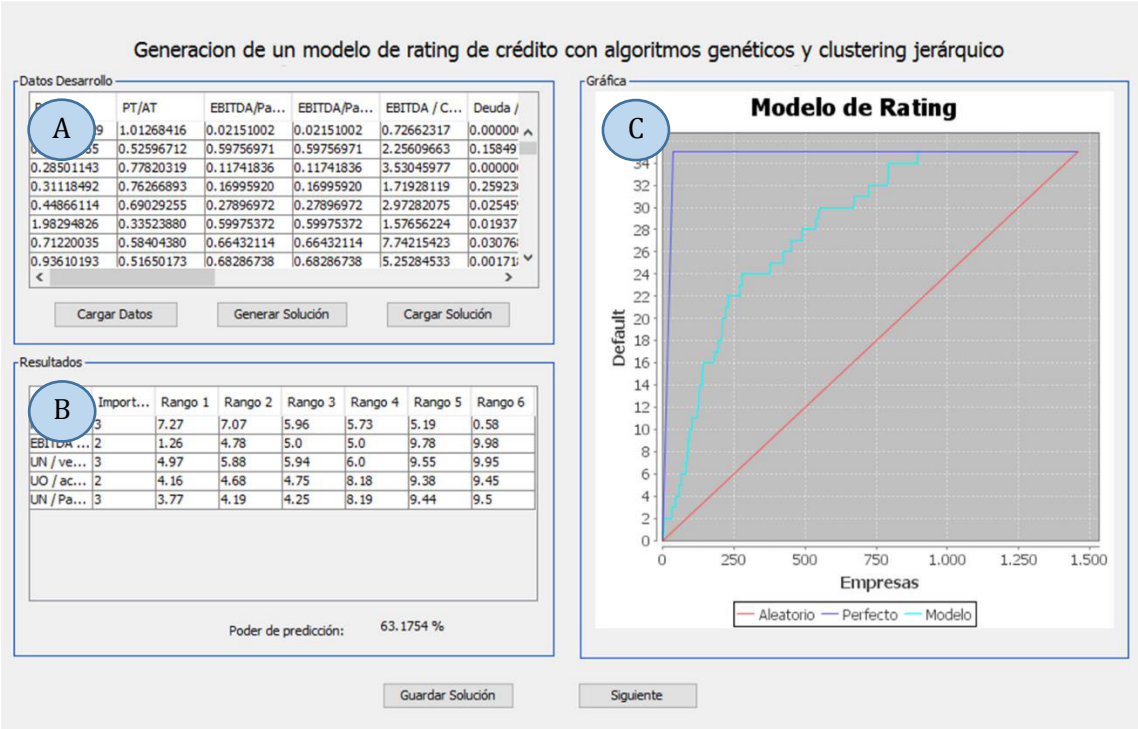


Figura 19: Pantalla N°1 para generar el modelo de puntuaciones. Fuente: Elaboración propia

(ii) Módulo de *clustering*: en esta interfaz se visualiza la continuación del primer módulo y toma como valores de entrada el modelo de puntuaciones generado por el algoritmo genético. Esta pantalla tiene 3 secciones: (A) la sección superior izquierda, donde se visualizan los grupos de *rating* generados por el algoritmo de *clustering*; (B) la sección inferior izquierda, donde se visualiza nuevamente el modelo de puntuaciones; y (C) la sección derecha, donde se observa el resultado del poder de predicción con el modelo de grupos de *rating* final.

Al iniciar este módulo, se ejecuta por defecto el algoritmo de *clustering* jerárquico, por lo que ya se muestran los resultados de los grupos de *rating* al inicio de esta interfaz. El número de grupos mostrado es aquel que maximiza el poder de predicción. El usuario también cuenta con una opción en la sección superior izquierda para elegir un corte distinto de número de *clusters*. Si el usuario desea guardar el modelo de *rating* hallado, puede

salvarlo con la opción “Guardar Modelo”. En la Figura 20 se muestra un ejemplo de la interfaz luego de ejecutarse el algoritmo de *clustering*:

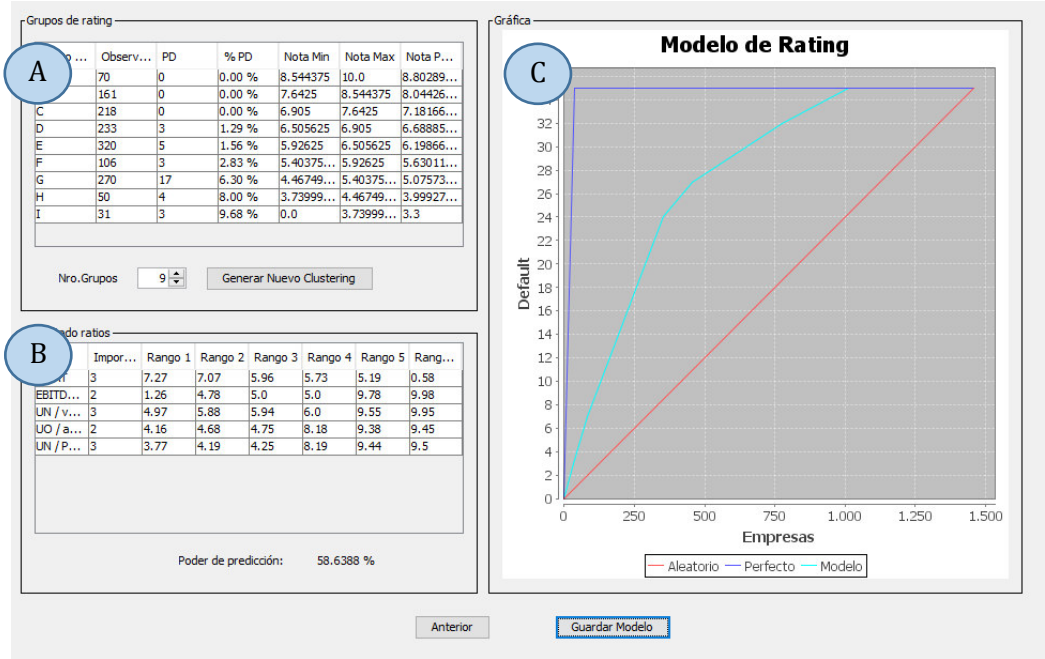


Figura 20: Pantalla N°2 para generar el modelo de agrupación. Fuente: Elaboración propia

4.5 RECURSOS UTILIZADOS PARA LA IMPLEMENTACIÓN:

Los recursos técnicos utilizados para la implementación de la solución fueron:

- El entorno de desarrollo **Java Development Kit** (JDK 8 u101) en conjunto con el entorno de desarrollo integrado (IDE) **Netbeans 8.1**.
- El API **JExcel.** Este API Java de código abierto permite leer y generar hojas de cálculo de Excel de forma dinámica.
- El API **JFreeChart**. Este API Java es una librería gráfica de código libre que facilita la generación de gráficos en aplicaciones Java.
- La librería **Commons Math** de Apache. Esta librería de código libre contiene componentes matemáticos y estadísticos adicionales al JDK 8.
- La librería **Hierarchical Clustering Java** de código libre con la implementación del algoritmo de *clustering* jerárquico aglomerativo.
- Una **PC** con procesador Intel Core i7, Memoria RAM de 12 GB y Sistema Operativo Windows 10 de 64 bits.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

En este capítulo se presentan las conclusiones del trabajo de tesis y los trabajos futuros derivados de la investigación.

5.1 CONCLUSIONES

La presente investigación propone la implementación de una herramienta de análisis de riesgo de crédito basada en el modelo de *rating* de crédito, para el cual se utilizan técnicas de la inteligencia artificial, como los algoritmos genéticos y algoritmos de *clustering* jerárquico aglomerativo, para hallar una solución al problema.

Se lograron adaptar los parámetros de ambas técnicas para hallar un modelo de puntuaciones y para determinar los grupos de riesgo de acuerdo a su homogeneidad. Adicionalmente, se lograron integrar ambas técnicas para poder determinar un modelo integral de clasificación de *ratings* de crédito, con un indicador de poder de predicción muy aceptable y con resultados coherentes, siendo esto validado por un experto de análisis de riesgos de crédito.

Con esto se concluye que sí es posible aplicar los algoritmos genéticos y los algoritmos de *clustering* jerárquico aglomerativo para hallar una solución al problema de análisis de riesgo basado en el modelo de *rating* de crédito. Adicionalmente, se concluye que el modelo obtenido con estos algoritmos puede ser fácilmente interpretado por un experto humano, a diferencia de otras técnicas de inteligencia artificial como las redes neuronales, donde el experto no tiene visibilidad sobre el razonamiento subyacente al modelo. Esto representa una ventaja frente a otros métodos pues el experto humano gana confianza en el método ya que entiende el razonamiento que hay detrás. Incluso el experto tendría la capacidad de modificar el entregable de este sistema, añadiendo al modelo final un input extra derivado de su experiencia, si así lo deseara.

También se concluye que la solución arroja modelos robustos. Se probó con una muestra de validación el poder de predicción de la solución, obteniendo un desempeño de clasificación muy similar entre el modelo de entrenamiento y el de validación (58.9% de *power statistic* versus 59.4%, respectivamente).

Además, se concluye que la solución planteada tiene mejor desempeño que el método de regresión logística, siendo la técnica logística una de las técnicas estadísticas más utilizadas en este tipo de problemas. La ventaja del modelo propuesto frente a la regresión logística es notable por su capacidad para trabajar con un conjunto de datos con pocas casuísticas de incumplimiento en las observaciones (es decir, sólo se tenían 35 casos de incumplimiento en un universo de 1459 observaciones). Se observó que el poco poder de predicción del modelo logístico se debió a los pocos casos de incumplimiento que tenía la información, ya que estos modelos estadísticos requieren de muchos de estos casos para poder calibrar su modelo (Bastante, 2010), concluyendo así que la presente propuesta puede llegar a buenos resultados aun a pesar de la carencia de estas casuísticas.

5.2 RECOMENDACIONES

En este trabajo de investigación se implementaron los 2 bloques correspondientes a la definición de un modelo de *rating* a través de 2 técnicas de la inteligencia artificial: (1) un modelo de puntuaciones óptimo a través del uso de algoritmos genéticos, y (2) un modelo de agrupación de grupos de riesgo a través de algoritmos de *clustering* jerárquico aglomerativo.

En futuras investigaciones se recomienda aplicar otras técnicas de optimización, así como otras técnicas de *clusterización*, y compararlas con el desempeño de la solución hallada en este trabajo.

Adicionalmente, se recomienda realizar una nueva investigación incluyendo variables cualitativas para medir el desempeño del algoritmo en esa situación. En este caso, también puede investigarse sobre el problema de ambigüedad de las variables cualitativas con alguna técnica de la inteligencia artificial como sistemas difusos, tal como se observó en algunos trabajos del estado del arte.

También se recomienda para un trabajo futuro la aplicación de alguna técnica automatizada de la inteligencia artificial para realizar la definición de los puntos de corte del *bucketing* en el modelo de puntuaciones. En este trabajo se aplicaron definiciones estándar sobre este tópico por recomendación del experto analista de riesgos, pero es posible realizar una investigación adicional para explorar valores óptimos para esta parametrización.

Adicionalmente, se recomienda investigar sobre un caso particular en la generación de un modelo de *rating*. Este caso se da cuando en la data de entrada no existen empresas que hayan incumplido con el pago de su deuda en el periodo evaluado (la columna *default* de la data de entrada en este trabajo de investigación). Este campo es indispensable para utilizar la metodología planteada en este trabajo de tesis, pero en algunas ocasiones puede no contarse con esta información. En aquellos casos se utiliza una metodología distinta para hallar el modelo de *rating* de crédito, pudiendo en un trabajo futuro determinar la técnica de inteligencia artificial más adecuada para resolver esta particularidad del problema.

REFERENCIAS BIBLIOGRÁFICAS

1. Abiyev, R. H. (2014). Credit rating using type-2 fuzzy neural networks. *Mathematical Problems in Engineering*, 2014.

2. Acerca de la SBS (s.f.). Recuperado el 22 de agosto del 2016, de http://www.sbs.gob.pe/principal/categoria/acerca-de-la-sbs/4/c-4

3. Aggarwal, C. C., & Reddy, C. K. (Eds.). (2013). *Data clustering: algorithms and applications*. CRC Press.

4. Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, 50(12), 1185-1190.

5. BBVA (2005). *Curso Intensivo de Riesgo de Crédito*. Lima, Perú

6. Bhatia, M. (2006). *Credit Risk Management and Basel II, An implementation Guide*. Risk books.

7. Chen, W. H., & Shih, J. Y. (2006). A study of Taiwan's issuer credit rating systems using support vector machines. *Expert Systems with Applications*, 30(3), 427-435.

8. Choudhry, M. (2012). *The Principles of Banking*. John Wiley & Sons.

9. Correa, A., Gonzalez, A., & Ladino, C. (2011). Genetic Algorithm Optimization for Selecting the Best Architecture of a Multi-Layer Perceptron Neural Network: A Credit Scoring Case. En *SAS Global Forum 2011 Data Mining and Text Analytics*.

10. Crook, J. N., Thomas, L. C., & Hamilton, R. (1994). Credit cards: haves, have-nots and cannot-haves. *Service Industries Journal*, 14(2), 204-215.

11. Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (Vol. 20). Siam.

12. García, M. L. S., & García, M. J. S. (2010). Modelos para medir el riesgo de crédito de la banca. *Cuadernos de Administración*, 23(40), 295-319.

13. Gestal, M., Rivero, D., Rabuñal, J. R., Dorado, J., & Pazos, A. (2010). Introducción a los Algoritmos genéticos y la Programación genética. *A Coruña, 2010*, 30-68.

14. Golberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Estados Unidos. Addison Wesley.

15. Hajek, P., & Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, 51, 72-84.

16. Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

17. James, C. F. I. (2006). System and method for providing search results with configurable scoring formula. *U.S. Patent No. 7,039,631*. Washington, DC: U.S. Patent and Trademark Office.

18. Jiao, Y., Syau, Y. R., & Lee, E. S. (2007). Modelling credit rating by fuzzy adaptive network. *Mathematical and Computer Modelling*, 45(5), 717-731.

19. Kim, K. J., & Ahn, H. (2012). A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach. *Computers & Operations Research*, 39(8), 1800-1811.

20. Kleinbaum, D. G., & Klein, M. (2010). *Survival analysis: a self-learning text*. Springer Science & Business Media.

21. Lee, Y. C. (2007). Application of support vector machines to corporate credit rating prediction. *Expert Systems with Applications*, 33(1), 67-74.

22. Man, K. F., Tang, K. S., & Kwong, S. (1996). Genetic algorithms: concepts and applications. *IEEE transactions on Industrial Electronics*, 43(5), 519-534.

23. Man, R. (2014). Survival analysis in credit scoring: A framework for PD estimation.

24. Nakasone, G. T. (2001). *Análisis de Estados Financieros para la Toma de Decisiones*. Fondo Editorial de la Pontificia Universidad Católica del Perú.

25. Samaniego, R. (2007). *El riesgo de crédito en el marco del acuerdo de Basilea II*. Delta Publicaciones.

26. Sivanandam, S. N., & Deepa, S. N. (2008). *Introduction to genetic algorithms*. Springer.

27. Standard & Poor's. (2011). Guide to credit rating essentials

28. Thisted, R. A. (1998). What is a P-value?. *Mayo, 25, 1-6.*

29. Trujillo, A., Samaniego, R., & Cardone, C. (2014). *Análisis del poder explicativo de los modelos de riesgo de crédito: Una aplicación a empresas no financieras europeas*. Editorial Universidad de Cantabria.

30. Tsai, C. F., & Chen, M. L. (2010). Credit rating by hybrid machine learning techniques. *Applied soft computing*, 10(2), 374-380.

31. Van Gestel, T., & Baesens, B. (2009). *Credit Risk Management: Basic concepts: Financial risk components, Rating analysis, models, economic and regulatory capital*. Oxford University Press.

32. Zaalberg, R. (2013). Quantifying uncertainty in credit rating model development.

ANEXO A

La siguiente sección muestra un ejemplo para cada método de *clustering* jerárquico aglomerativo detallado en la sección 2.3.3.

Para los ejemplos se usará el conjunto de datos mostrado en la Figura 21:

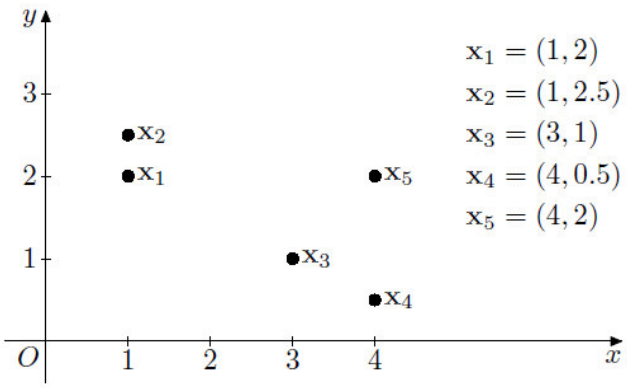


Figura 21: Ejemplo de un conjunto de datos bidimensional con cinco puntos. Fuente: Gan et al. (2007)

Para la data mostrada, se calcula la matriz de disimilitud descrita en la Figura 22. En esta matriz, la distancia entre dos puntos es calculada usando la distancia Euclidiana.

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{x}_1$	0	0.5	2.24	3.35	3
$\mathbf{x}_2$	0.5	0	2.5	3.61	3.04
$\mathbf{x}_3$	2.24	2.5	0	1.12	1.41
$\mathbf{x}_4$	3.35	3.61	1.12	0	1.5
$\mathbf{x}_5$	3	3.04	1.41	1.5	0

Figura 22: Matriz de disimilitud del conjunto de datos de la Figura 21. Fuente: Gan et al. (2007)

I. Método de la distancia mínima (*Single-link)*

Si se usa el algoritmo de *clustering* jerárquico de distancia mínima sobre este conjunto de datos, entonces $\mathbf{x}_1$ y $\mathbf{x}_2$ serán agrupados para formar un nuevo *cluster* en la primera etapa del algoritmo, ya que tienen la distancia más pequeña en la matriz de disimilitud.

Luego se calcularán las distancias entre el nuevo *cluster* $\{\mathbf{x}_1, \mathbf{x}_2\}$ y los puntos $\mathbf{x}_3$, $\mathbf{x}_4$ y $\mathbf{x}_5$.

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_3}) = \min\{d(\mathbf{x_1}, \mathbf{x_3}), d(\mathbf{x_2}, \mathbf{x_3})\} = 2.24$$

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_4}) = \min\{d(\mathbf{x_1}, \mathbf{x_4}), d(\mathbf{x_2}, \mathbf{x_4})\} = 3.35$$

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_5}) = \min\{d(\mathbf{x_1}, \mathbf{x_5}), d(\mathbf{x_2}, \mathbf{x_5})\} = 3.00$$

Luego de que $\mathbf{x}_1$ y $\mathbf{x}_2$ han sido fusionados, la matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.24	3.35	3
$\mathbf{x}_3$	2.24	0	1.12	1.41
$\mathbf{x}_4$	3.35	1.12	0	1.5
$\mathbf{x}_5$	3	1.41	1.5	0

En la segunda etapa del algoritmo, $\mathbf{x}_3$ y $\mathbf{x}_4$ se fusionan ya que tienen la menor distancia. Entonces las distancias entre el grupo $\{\mathbf{x}_3, \mathbf{x}_4\}$ y los grupos restantes se vuelven:

$$\boldsymbol{D}(\{\mathbf{x_3}, \mathbf{x_4}\}, \{\mathbf{x_1}, \mathbf{x_2}\}) = \mathbf{min}\{\boldsymbol{d}(\mathbf{x_1}, \mathbf{x_3}), \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_3}), \boldsymbol{d}(\mathbf{x_1}, \mathbf{x_4}), \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_4})\} = \mathbf{2.24}$$

$$\boldsymbol{D}(\{\mathbf{x_3}, \mathbf{x_4}\}, \mathbf{x_5}) = \mathbf{min}\{\boldsymbol{d}(\mathbf{x_3}, \mathbf{x_5}), \boldsymbol{d}(\mathbf{x_4}, \mathbf{x_5})\} = \mathbf{1.41}$$

Luego de que $\mathbf{x}_3$ y $\mathbf{x}_4$ han sido fusionados, la matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_3, \mathbf{x}_4\}$	$\mathbf{x}_5$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.24	3
$\{\mathbf{x}_3, \mathbf{x}_4\}$	2.24	0	1.41
$\mathbf{x}_5$	3	1.41	0

En la tercera etapa del algoritmo, $\{\mathbf{x}_3, \mathbf{x}_4\}$ y $\mathbf{x}_5$ se fusionan. La matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.24
$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$	2.24	0

En la quinta etapa, todos los puntos se fusionan en un solo *cluster*. El dendograma de este *clustering* se muestra en la Figura 23:

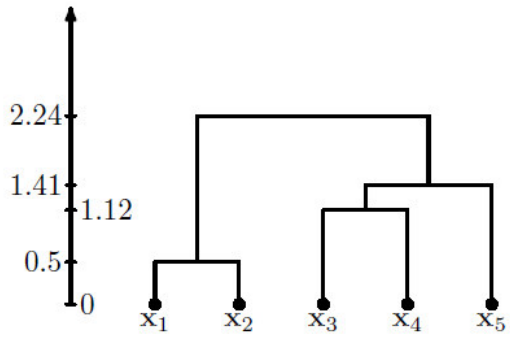


Figura 23: El dendograma resultante del método de distancia mínima. Fuente: Gan et al. (2007)

II. Método de la distancia máxima (*Complete-Link)*

Aplicando el método de distancia máxima a la tabla de disimilitud de la Figura 14, en la primera etapa se fusiona $\mathbf{x}_1$ y $\mathbf{x}_2$. Las distancias entre el grupo $\{\mathbf{x}_1, \mathbf{x}_2\}$ y los puntos restantes son:

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_3}) = \max\{d(\mathbf{x_1}, \mathbf{x_3}), d(\mathbf{x_2}, \mathbf{x_3})\} = 2.5$$

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_4}) = \max\{d(\mathbf{x_1}, \mathbf{x_4}), d(\mathbf{x_2}, \mathbf{x_4})\} = 3.61$$

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_5}) = \max\{d(\mathbf{x_1}, \mathbf{x_5}), d(\mathbf{x_2}, \mathbf{x_5})\} = 3.04$$

Luego de que $\mathbf{x}_1$ y $\mathbf{x}_2$ han sido fusionados, la matriz de disimilitud queda así:

	$\{x_1, x_2\}$	x_3	x_4	x_5
$\{x_1, x_2\}$	0	2.5	3.61	3.04
x_3	2.5	0	1.12	1.41
x_4	3.61	1.12	0	1.5
x_5	3.04	1.41	1.5	0

En la segunda etapa del algoritmo, x_3 y x_4 se fusionan, ya que tienen la menor distancia. Luego de que x_3 y x_4 se fusionan, las distancias entre el grupo $\{x_3, x_4\}$ y los grupos restantes se actualizan como sigue:

$$\boldsymbol{D(\{x_3, x_4\}, \{x_1, x_2\}) = \max\{d(x_1, x_3), d(x_2, x_3), d(x_1, x_4), d(x_2, x_4)\} = 3.61}$$

$$\boldsymbol{D(\{x_3, x_4\}, x_5) = \max\{d(x_3, x_5), d(x_4, x_5)\} = 1.5}$$

Luego de que x_3 y x_4 han sido fusionados, la matriz de disimilitud queda así:

	$\{x_1, x_2\}$	$\{x_3, x_4\}$	x_5
$\{x_1, x_2\}$	0	3.61	3.04
$\{x_3, x_4\}$	3.61	0	1.5
x_5	3.04	1.5	0

En la tercera etapa del algoritmo, $\{x_3, x_4\}$ y x_5 se fusionan ya que tienen la menor distancia. La distancia y la matriz de disimilitud quedan así:

$$\boldsymbol{D(\{x_1, x_2\}, \{x_3, x_4, x_5\})}$$
$$\boldsymbol{= \max\{d(x_1, x_3), d(x_1, x_4), d(x_1, x_5), d(x_2, x_3), d(x_2, x_4), d(x_2, x_5)\} = 3.61}$$

	$\{x_1, x_2\}$	$\{x_3, x_4, x_5\}$
$\{x_1, x_2\}$	0	3.61
$\{x_3, x_4, x_5\}$	3.61	0

El dendograma de este *clustering* se muestra en la Figura 24:

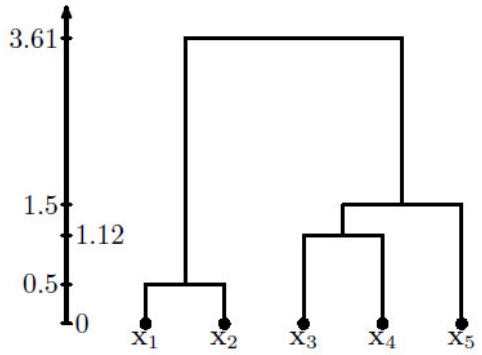


Figura 24: El dendograma resultante del método de distancia máxima. Fuente: Gan et al. (2007)

III. Método de la distancia promedio no ponderada (*Weighted Group Average-Link)*

Aplicando el método de distancia promedio no ponderada al conjunto de datos de la Figura 14, se tiene que en la primera etapa se fusiona $\mathbf{x}_1$ y $\mathbf{x}_2$. Tras fusionar $\mathbf{x}_1$ y $\mathbf{x}_2$, las distancias entre el grupo $\{\mathbf{x}_1, \mathbf{x}_2\}$ y los puntos restantes son:

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_3}) = \frac{1}{2} d(\mathbf{x_1}, \mathbf{x_3}) + \frac{1}{2} d(\mathbf{x_2}, \mathbf{x_3}) = 2.37$$

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_4}) = \frac{1}{2} d(\mathbf{x_1}, \mathbf{x_4}) + \frac{1}{2} d(\mathbf{x_2}, \mathbf{x_4}) = 3.48$$

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_5}) = \frac{1}{2} d(\mathbf{x_1}, \mathbf{x_5}) + \frac{1}{2} d(\mathbf{x_2}, \mathbf{x_5}) = 3.02$$

Luego de que $\mathbf{x}_1$ y $\mathbf{x}_2$ han sido fusionados, la matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.37	3.48	3.02
$\mathbf{x}_3$	2.37	0	1.12	1.41
$\mathbf{x}_4$	3.48	1.12	0	1.5
$\mathbf{x}_5$	3.02	1.41	1.5	0

En la segunda etapa del algoritmo, $\mathbf{x}_3$ y $\mathbf{x}_4$ se fusionan. Las distancias entre el grupo $\{\mathbf{x}_3, \mathbf{x}_4\}$ y los grupos restantes se actualizan como sigue:

$$\boldsymbol{D(\{\mathbf{x_1}, \mathbf{x_2}\}, \{\mathbf{x_3}, \mathbf{x_4}\}) = \frac{1}{2} d(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_3}) + \frac{1}{2} d(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_4}) = 2.93}$$

$$\boldsymbol{D(\{x_3, x_4\}, x_5) = \frac{1}{2} d(x_3, x_5) + \frac{1}{2} d(x_4, x_5) = 1.46}$$

Luego de que $\mathbf{x}_3$ y $\mathbf{x}_4$ han sido fusionados, la matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_3, \mathbf{x}_4\}$	$\mathbf{x}_5$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.93	3.02
$\{\mathbf{x}_3, \mathbf{x}_4\}$	2.93	0	1.46
$\mathbf{x}_5$	3.02	1.46	0

En la tercera etapa del algoritmo, $\{\mathbf{x}_3, \mathbf{x}_4\}$ y $\mathbf{x}_5$ se fusionan ya que tienen la menor distancia. La distancia entre $\{\mathbf{x}_1, \mathbf{x}_2\}$ y $\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ queda así:

$$\boldsymbol{D(\{x_1, x_2\}, \{x_3, x_4, x_5\}) = \frac{1}{2} d(\{x_1, x_2\}, \{x_3, x_4\}) + \frac{1}{2} d(\{x_1, x_2\}, x_5) = 2.98}$$

Y la matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.98
$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$	2.98	0

El dendograma de este *clustering* se muestra en la Figura 25:

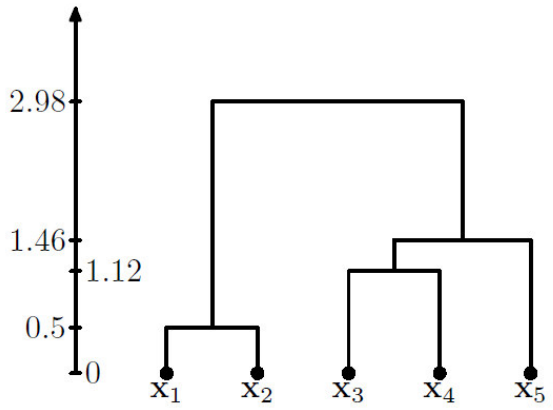


Figura 25: El dendograma resultante del método de distancia promedio no ponderada. Fuente: Gan et al. (2007)

IV. Método de la distancia promedio ponderada (*Group Average-Link)*

Aplicando el método de distancia promedio ponderada al conjunto de datos de la Figura 14, se tiene que en la primera etapa se fusiona $\mathbf{x}_1$ y $\mathbf{x}_2$. Tras fusionar $\mathbf{x}_1$ y $\mathbf{x}_2$, las distancias entre el grupo $\{\mathbf{x}_1, \mathbf{x}_2\}$ y los puntos restantes son:

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_3}) = \frac{1}{2} d(\mathbf{x_1}, \mathbf{x_3}) + \frac{1}{2} d(\mathbf{x_2}, \mathbf{x_3}) = 2.37$$

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_4}) = \frac{1}{2} d(\mathbf{x_1}, \mathbf{x_4}) + \frac{1}{2} d(\mathbf{x_2}, \mathbf{x_4}) = 3.48$$

$$D(\{\mathbf{x_1}, \mathbf{x_2}\}, \mathbf{x_5}) = \frac{1}{2} d(\mathbf{x_1}, \mathbf{x_5}) + \frac{1}{2} d(\mathbf{x_2}, \mathbf{x_5}) = 3.02$$

Luego de que $\mathbf{x}_1$ y $\mathbf{x}_2$ han sido fusionados, la matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.37	3.48	3.02
$\mathbf{x}_3$	2.37	0	1.12	1.41
$\mathbf{x}_4$	3.48	1.12	0	1.5
$\mathbf{x}_5$	3.02	1.41	1.5	0

En la segunda etapa del algoritmo, $\mathbf{x}_3$ y $\mathbf{x}_4$ se fusionan. Las distancias entre el grupo $\{\mathbf{x}_3, \mathbf{x}_4\}$ y los grupos restantes se actualizan como sigue:

$$\boldsymbol{D}(\{\mathbf{x_1}, \mathbf{x_2}\}, \{\mathbf{x_3}, \mathbf{x_4}\}) = \frac{\mathbf{1}}{\mathbf{4}} [\, \boldsymbol{d}(\mathbf{x_1}, \mathbf{x_3}) + \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_3}) + \boldsymbol{d}(\mathbf{x_1}, \mathbf{x_4}) + \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_4}) \,] = \mathbf{2.93}$$

$$\boldsymbol{D}(\{\mathbf{x_3}, \mathbf{x_4}\}, \mathbf{x_5}) = \frac{\mathbf{1}}{\mathbf{2}} \boldsymbol{d}(\mathbf{x_3}, \mathbf{x_5}) + \frac{\mathbf{1}}{\mathbf{2}} \boldsymbol{d}(\mathbf{x_4}, \mathbf{x_5}) = \mathbf{1.46}$$

Luego de que $\mathbf{x}_3$ y $\mathbf{x}_4$ han sido fusionados, la matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_3, \mathbf{x}_4\}$	$\mathbf{x}_5$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.93	3.02
$\{\mathbf{x}_3, \mathbf{x}_4\}$	2.93	0	1.46
$\mathbf{x}_5$	3.02	1.46	0

En la tercera etapa del algoritmo, $\{\mathbf{x}_3, \mathbf{x}_4\}$ y $\mathbf{x}_5$ se fusionan ya que tienen la menor distancia. La distancia entre $\{\mathbf{x}_1, \mathbf{x}_2\}$ y $\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ queda así:

$$\boldsymbol{D}(\{\mathbf{x_1}, \mathbf{x_2}\}, \{\mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}\}) = \frac{\mathbf{1}}{\mathbf{6}}[\boldsymbol{d}(\mathbf{x_1}, \mathbf{x_3}) + \boldsymbol{d}(\mathbf{x_1}, \mathbf{x_4}) + \boldsymbol{d}(\mathbf{x_1}, \mathbf{x_5}) + \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_3}) + \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_4}) + \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_5})] = \mathbf{2.96}$$

Y la matriz de disimilitud queda así:

	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$
$\{\mathbf{x}_1, \mathbf{x}_2\}$	0	2.96
$\{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$	2.96	0

El dendograma de este *clustering* se muestra en la Figura 26:

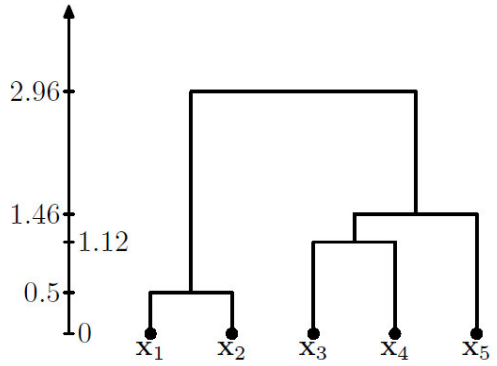


Figura 26: El dendograma resultante del método de distancia promedio ponderada. Fuente: Gan et al. (2007)

ANEXO B

La siguiente sección muestra el resultado del proceso de jerarquía analítica para determinar los métodos más adecuados de algoritmos genéticos y *clustering* que se utilizarán en la solución.

1) **Análisis de los métodos de algoritmos genéticos**

El objetivo de este análisis consiste en identificar el método de algoritmo genético (AG) más adecuado para la implementación de la solución propuesta en esta tesis. La siguiente tabla muestra los 3 métodos de algoritmos genéticos que se evaluaron en el análisis:

Método / Algoritmo	Descripción
AG simple (Alternativa 1)	Es la forma básica de algoritmo genético propuesta por Goldberg para solucionar problemas de búsqueda y optimización (Sivanandam & Deepa, 2008).
AG adaptable (Alternativa 2)	Es un tipo de algoritmo genético cuyos parámetros (como el tamaño de población, probabilidad de cruzamiento o mutación) pueden variar mientras el AG está en ejecución (Sivanandam & Deepa, 2008).
AG paralelo y distribuido (Alternativa 3)	Es un tipo de algoritmo genético (AG) que consiste en una ejecución paralela de varios AGs. Se diseñaron para reducir el tiempo de ejecución del algoritmo en problemas con un espacio de búsqueda muy grande (Sivanandam & Deepa, 2008).

Los criterios que se utilizaron para comparar estos métodos fueron los siguientes:

- *Eficacia* (Criterio 1): se refiere a la capacidad para encontrar la solución óptima global
- *Rendimiento* (Criterio 2): se refiere a la capacidad para encontrar la solución óptima de manera rápida
- *Simplicidad de implementación* (Criterio 3): se refiere a la simplicidad que tiene la técnica para implementarse de manera fácil y precisa, evitando así la excesiva complejidad en la construcción y ejecución

- *Acceso a la Literatura* (Criterio 4): esta característica se refiere a la cantidad de bibliografía disponible acerca de la técnica a implementar, de modo que sean accesibles las mejores fuentes de estudio.

A continuación, se calculan las matrices de comparación para establecer las prioridades por criterio según la metodología AHP:

Eficacia (Criterio 1)							
	Alternativa 1	Alternativa 2	Alternativa 3	Matriz Normalizada			Promedio
Alternativa 1	1	0.33	0.33	0.14	0.08	0.20	**0.14**
Alternativa 2	3	1	0.33	0.43	0.23	0.20	**0.29**
Alternativa 3	3	3	1	0.43	0.69	0.60	**0.57**
SUMA	7.00	4.33	1.67				

Rendimiento (Criterio 2)							
	Alternativa 1	Alternativa 2	Alternativa 3	Matriz Normalizada			Promedio
Alternativa 1	1	1	0.2	0.14	0.14	0.14	**0.14**
Alternativa 2	1	1	0.2	0.14	0.14	0.14	**0.14**
Alternativa 3	5	5	1	0.71	0.71	0.71	**0.71**
SUMA	7.00	7.00	1.40				

Simplicidad (Criterio 3)							
	Alternativa 1	Alternativa 2	Alternativa 3	Matriz Normalizada			Promedio
Alternativa 1	1	5	7	0.74	0.79	0.64	**0.72**
Alternativa 2	0.20	1	3	0.15	0.16	0.27	**0.19**
Alternativa 3	0.14	0.33	1	0.11	0.05	0.09	**0.08**
SUMA	1.34	6.33	11.00				

Acceso a la Literatura (Criterio 4)							
	Alternativa 1	Alternativa 2	Alternativa 3	Matriz Normalizada			Promedio
Alternativa 1	1	5	5	0.71	0.71	0.71	**0.71**
Alternativa 2	0.20	1	1	0.14	0.14	0.14	**0.14**
Alternativa 3	0.20	1.00	1	0.14	0.14	0.14	**0.14**
SUMA	1.40	7.00	7.00				

Matriz de comparación de criterios									
	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Matriz Normalizada				Promedio
Criterio 1	1	3	3	0.33	0.21	0.21	0.67	0.07	**0.29**
Criterio 2	0.33	1	0.14	0.33	0.07	0.07	0.03	0.07	**0.06**
Criterio 3	0.33	7	1	3	0.07	0.50	0.22	0.64	**0.36**
Criterio 4	3	3	0.33	1	0.64	0.21	0.07	0.21	**0.29**
SUMA	4.67	14.00	4.48	4.67					

Finalmente, en la siguiente tabla se observa que la "Alternativa 1" es la que obtiene una mayor ponderación (0.51). Por lo que el método elegido para la implementación de la solución es el algoritmo genético simple.

	Criterio 1	Criterio 2	Criterio 3	Criterio 4	Total
Alternativa 1	0.14	0.14	0.72	0.71	**0.51**
Alternativa 2	0.29	0.14	0.19	0.14	**0.20**
Alternativa 3	0.57	0.71	0.08	0.14	**0.28**
Ponderación de criterios	0.29	0.06	0.36	0.29	

2) Análisis de los métodos de *clustering*

El objetivo de este análisis consiste en identificar el método de *clustering* más adecuado para la implementación de la solución propuesta en esta tesis. La siguiente tabla muestra los 3 métodos de *clustering* que se evaluaron en el análisis:

Método	Descripción
Clustering k-means (Alternativa 1)	Este algoritmo agrupa datos numéricos, donde cada grupo tiene un "centro" llamado **media.** Cada elemento del conjunto de datos es asignado por proximidad a uno de los centros. Este algoritmo es de tipo particional y no jerárquico. El algoritmo requiere que el número de *clusters* sea definido como parámetro inicial (Aggarwal & Reddy, 2013).
Clustering Jerárquico aglomerativo (Alternativa 2)	Este algoritmo *clusteriza* sin necesidad de un parámetro predefinido a diferencia de los métodos particionales. Agrupa los elementos desde el nivel más desagregado hasta construir una jerarquía de *clusters* de "abajo hacia arriba" (Aggarwal & Reddy, 2013).
Clustering Jerárquico divisivo (Alternativa 3)	Este algoritmo *clusteriza* sin necesidad de un parámetro predefinido a diferencia de los métodos particionales. A diferencia del algoritmo aglomerativo, empieza con todos los objetos en un solo gran *cluster* "macro" y lo divide continuamente en 2 grupos generando una jerarquía de *clusters* de "arriba hacia abajo". Requiere un algoritmo adicional de bisección para la división de *clusters* (Aggarwal & Reddy, 2013).

Los criterios que se utilizaron para comparar estos métodos fueron los siguientes:

- *Ausencia de parámetros iniciales* (Criterio 1): se refiere a la capacidad del algoritmo de encontrar la solución sin necesidad de conocer de antemano el número de *clusters* final
- *Rendimiento* (Criterio 2): se refiere a la capacidad para encontrar la solución de manera rápida
- *Simplicidad de implementación* (Criterio 3): se refiere a la simplicidad que tiene la técnica para implementarse de manera fácil y precisa, evitando así la excesiva complejidad en la construcción y ejecución
- *Acceso a la Literatura* (Criterio 4): esta característica se refiere a la cantidad de bibliografía disponible acerca de la técnica a implementar, de modo que sean accesibles las mejores fuentes de estudio.

A continuación, se calculan las matrices de comparación para establecer las prioridades por criterio según la metodología AHP:

Ausencia de parámetros iniciales (Criterio 1)							
	Alternativa 1	**Alternativa 2**	**Alternativa 3**	**Matriz Normalizada**			**Promedio**
Alternativa 1	1	0.20	0.20	0.09	0.09	0.09	**0.09**
Alternativa 2	5	1	1.00	0.45	0.45	0.45	**0.45**
Alternativa 3	5	1	1	0.45	0.45	0.45	**0.45**
SUMA	11.00	2.20	2.20				

Rendimiento (Criterio 2)							
	Alternativa 1	**Alternativa 2**	**Alternativa 3**	**Matriz Normalizada**			**Promedio**
Alternativa 1	1	5	5	0.71	0.56	0.79	**0.69**
Alternativa 2	0.2	1	0.33	0.14	0.11	0.05	**0.10**
Alternativa 3	0.2	3	1	0.14	0.33	0.16	**0.21**
SUMA	1.40	9.00	6.33				

Simplicidad (Criterio 3)							
	Alternativa 1	**Alternativa 2**	**Alternativa 3**	**Matriz Normalizada**			**Promedio**
Alternativa 1	1	3	3	0.60	0.69	0.43	**0.57**
Alternativa 2	0.33	1	3	0.20	0.23	0.43	**0.29**
Alternativa 3	0.33	0.33	1	0.20	0.08	0.14	**0.14**
SUMA	1.67	4.33	7.00				

Acceso a la Literatura (Criterio 4)							
	Alternativa 1	**Alternativa 2**	**Alternativa 3**	**Matriz Normalizada**			**Promedio**
Alternativa 1	1	3	3	0.60	0.71	0.33	**0.55**
Alternativa 2	0.33	1	5	0.20	0.24	0.56	**0.33**
Alternativa 3	0.33	0.20	1	0.20	0.05	0.11	**0.12**
SUMA	1.67	4.20	9.00				

Matriz de comparación de criterios									
	Criterio 1	**Criterio 2**	**Criterio 3**	**Criterio 4**	**Matriz Normalizada**				**Promedio**
Criterio 1	1	5	5	5	0.63	0.54	0.42	0.75	**0.58**
Criterio 2	0.20	1	3	0.33	0.13	0.11	0.25	0.05	**0.13**
Criterio 3	0.20	0.33	1	0.33	0.13	0.04	0.08	0.05	**0.07**
Criterio 4	0.2	3	3	1	0.13	0.32	0.25	0.15	**0.21**
SUMA	1.60	9.33	12.00	6.67					

Finalmente, en la siguiente tabla se observa que la "Alternativa 2" es la que obtiene una mayor ponderación (0.51). Por lo que el método elegido para la implementación de la solución es el *clustering* jerárquico aglomerativo.

	Criterio 1	**Criterio 2**	**Criterio 3**	**Criterio 4**	**Total**
Alternativa 1	0.09	0.69	0.57	0.55	**0.30**
Alternativa 2	0.45	0.10	0.29	0.33	**0.37**
Alternativa 3	0.45	0.21	0.14	0.12	**0.33**
Ponderación de criterios	0.58	0.13	0.07	0.21	

ANEXO C

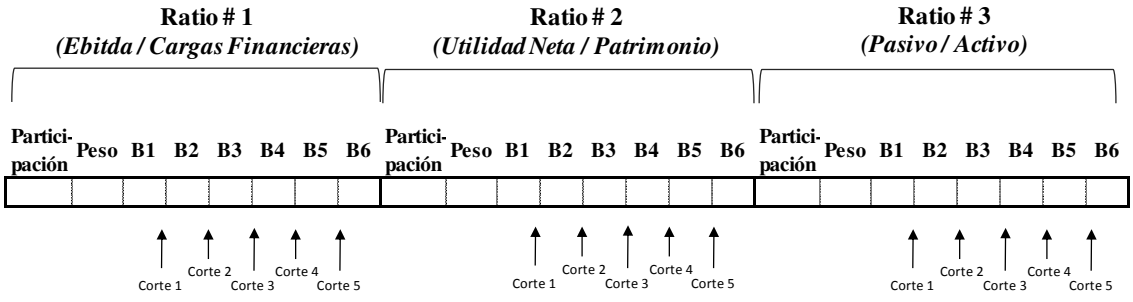
Esta sección muestra un ejemplo completo del procedimiento del algoritmo genético y *clustering* jerárquico.

Para esta ilustración, se usarán 3 variables de entrada, y un total de 5 registros de casos de entrada para mostrar de manera simple el proceso. Las variables a usar son:

Tipo de Variable	N°	Nombre de la variable	Proporcionalidad del ratio con la calidad crediticia
Ratios de Solvencia	1	EBITDA / Cargas Financieras	directa
Ratios de Rentabilidad	2	Utilidad Neta / Patrimonio	directa
Ratios de Solvencia	3	Pasivos / Activos	inversa
Comportamiento de cumplimiento de deuda	4	Default (0: cumplió en el pago; 1: incumplió en el pago)	

1. Procedimiento del Algoritmo genético

1.1 Representación del cromosoma: como la data de entrada tiene 3 variables predictoras, el cromosoma se compone de 3 bloques. Cada bloque tiene 8 genes que definirán la participación del ratio en el modelo, su ponderación y el *bucketing* del modelo. A continuación, se observa esta estructura:



Donde los puntos de corte para el *bucketing* se establecen en:

N°	Ratio	Corte 1	Corte 2	Corte 3	Corte 4	Corte 5
1	EBITDA / Cargas Fin	-0.4	2.9	6.1	9.4	12.6
2	UN / Patrimonio	-0.3	-0.1	0.0	0.2	0.4
3	PT/AT	0.2	0.4	0.6	0.7	0.9

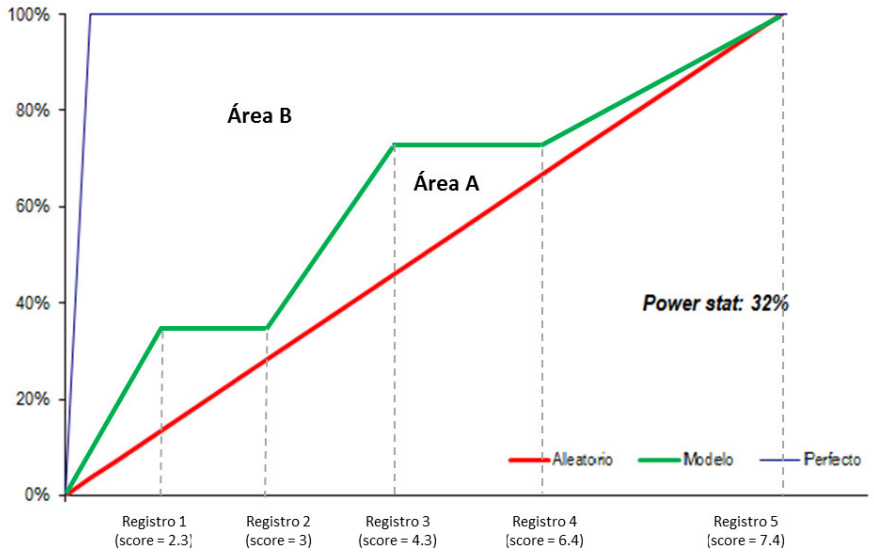
* De acuerdo al punto 4.2.1, el corte 1 para cada ratio es el percentil 5 de la muestra de datos; el corte 5 es el percentil 95 de la muestra de datos; y el resto de cortes divide en partes iguales el resto de *buckets*.

1.2 Generación de la Población Inicial: en este ejemplo, la población consta de 6 individuos (también llamados cromosomas). Se generan los valores de forma aleatoria de acuerdo al cromosoma definido en la sección anterior.

	Part.	Peso	B1	B2	B3	B4	B5	B6	Part.	Peso	B1	B2	B3	B4	B5	B6	Part.	Peso	B1	B2	B3	B4	B5	B6
Individuo 1	0	1	0.0	1.0	3.0	5.0	7.0	9.0	0	2	0.0	2.0	4.0	6.0	8.0	10.0	1	3	10.0	8.0	6.0	4.0	2.0	0.0
Individuo 2	1	3	0.9	4.9	5.5	6.8	8.0	9.8	0	1	2.9	4.5	5.3	6.5	7.9	8.1	0	2	7.5	7.3	6.8	4.4	3.2	2.6
Individuo 3	1	1	3.7	4.5	5.9	6.9	7.2	8.8	1	2	1.5	4.3	5.9	6.0	7.2	9.3	0	2	7.0	6.8	6.6	4.1	3.8	2.9
Individuo 4	0	3	0.6	3.2	3.3	5.8	8.1	8.2	0	3	3.5	3.6	5.8	7.4	7.8	8.5	0	3	9.1	8.4	8.4	2.1	1.3	0.1
Individuo 5	0	2	3.0	4.6	5.3	8.5	8.5	8.8	0	2	0.9	3.2	5.3	7.0	9.9	10.0	1	3	8.1	8.0	7.8	2.4	1.4	0.9
Individuo 6	0	3	2.4	4.8	5.2	8.2	9.6	9.9	1	3	0.3	3.0	4.8	5.6	8.1	9.1	0	3	8.5	8.0	5.6	4.3	2.8	1.4

1.3 Función de Aptitud: la función de aptitud consiste en el cálculo del *power statistic.* Se calcula este ratio para cada cromosoma (para el cálculo del Power Stat, ver sección 2.1.3). Como ejemplo, mostraremos el cálculo de la función de aptitud del primer individuo.

Siendo la base de entrada de 5 registros, se calcula el *score* de cada registro de acuerdo al modelo de *scoring* que establece el cromosoma 1. Se ordenan los registros de menor a mayor en el eje "X", y en el eje "Y" se establece el comportamiento de su deuda de forma acumulada (valor 0 si el registro cumplió o 1 si el registro incumplió). A continuación, se muestra la gráfica de *Power Statistic* para este ejemplo:



Dividiendo el área A entre la suma de las áreas A y B, se obtiene un *power statistic* de 32%. Lo mismo se repite para cada cromosoma de la población. Así se obtienen los siguientes valores de aptitud para la población inicial:

	Función de aptitud
Individuo 1	32%
Individuo 2	38%
Individuo 3	47%
Individuo 4	26%
Individuo 5	32%
Individuo 6	50%

1.4 Selección: en este trabajo se utiliza la selección por torneo. Se hacen 2 torneos para elegir 2 padres para el proceso de reproducción. Tanto en el primer como en el segundo torneo, se eligen aleatoriamente 2 individuos de la población que competirán para ser elegidos como padres. Gana el competidor con mayor valor de aptitud. A continuación, se muestran los valores de los 2 torneos de este ejemplo, en base a los individuos de la población inicial:

Torneo 1: Elección de 1° padre

	Individuo N°	**Aptitud**	
Aleatorio 1:	2	38%	**<- Ganador 1**
Aleatorio 2:	4	26%	

Torneo 2: Elección de 2° padre

	Individuo N°	**Aptitud**	
Aleatorio 1:	1	32%	
Aleatorio 2:	6	50%	**<- Ganador 2**

1.5 Cruzamiento: en el torneo se eligieron los individuos 2 y 6 de la población inicial para participar como padres. A continuación, de la fila 1 a la 5 se muestran el vector de cruzamiento (valores booleanos aleatorios), el padre 1 (individuo 2 de la población), el padre 2 (individuo 6 de la población) y los 2 hijos resultantes del cruce:

	Ratio 1								Ratio 2								Ratio 3							
	Part.	Peso	B.1	B.2	B.3	B.4	B.5	B.6	Part.	Peso	B.1	B.2	B.3	B.4	B.5	B.6	Part.	Peso	B.1	B.2	B.3	B.4	B.5	B.6
Máscara	1	1	0	0	1	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	1	0	1	0
Padre 1	1	3	0.9	4.9	5.5	6.8	8.0	9.8	0	1	2.9	4.5	5.3	6.5	7.9	8.1	0	2	7.5	7.3	6.8	4.4	3.2	2.6
Padre 2	0	3	2.4	4.8	5.2	8.2	9.6	9.9	1	3	0.3	3.0	4.8	5.6	8.1	9.1	0	3	8.5	8.0	5.6	4.3	2.8	1.4
Hijo 1	0	3	0.9	4.9	5.2	6.8	8.0	9.8	1	3	2.9	4.5	4.8	6.5	8.1	8.1	0	3	7.5	7.3	5.6	4.4	2.8	2.6
Hijo 2	1	3	2.4	4.8	5.5	8.2	9.6	9.9	0	1	0.3	3.0	5.3	5.6	7.9	9.1	0	2	8.5	8.0	6.8	4.3	3.2	1.4

1.6 Mutación: el siguiente paso para los hijos es la mutación de sus valores (de acuerdo a la probabilidad de mutación definida). En el ejemplo se mutaron los genes 1, 5, 10, 15, 17 y 21 del Hijo 1. El mismo proceso se realiza para el "Hijo 2"

	Ratio 1								Ratio 2								Ratio 3							
	Part.	Peso	B.1	B.2	B.3	B.4	B.5	B.6	Part.	Peso	B.1	B.2	B.3	B.4	B.5	B.6	Part.	Peso	B.1	B.2	B.3	B.4	B.5	B.6
Genes a mutar	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	0
Hijo 1 (antes de mutar)	0	3	0.9	4.9	5.17	6.8	8.0	9.8	1	3	2.9	4.5	4.8	6.5	8.09	8.1	0	3	7.5	7.3	5.61	4.4	2.81	2.6
Hijo 1 (mutado)	1	3	0.9	4.9	6	6.8	8.0	9.8	1	1	2.9	4.5	4.8	6.5	7	8.1	1	3	7.5	7.3	6	4.4	2.81	2.6

1.7 Reemplazo: en este trabajo, el proceso de reproducción se repite hasta completar el número de individuos de la población (a este tipo de reemplazo se le conoce como actualización generacional según la sección 2.2.3. En este ejemplo, se generan 6 hijos). Pero, además en este trabajo se usa **elitismo**, por lo que se rescata el mejor individuo de la población inicial y se inserta en la nueva población, dejando de lado al peor de los 6 hijos. En este ejemplo, el mejor individuo de la población inicial es el individuo 6. La nueva población se muestra a continuación:

	Part.	Peso	B1	B2	B3	B4	B5	B6	Part.	Peso	B1	B2	B3	B4	B5	B6	Part.	Peso	B1	B2	B3	B4	B5	B6	Función de aptitud
Hijo 1	1	3	0.9	4.9	6.0	6.8	8.0	9.8	1	1	2.9	4.5	4.8	6.5	7.0	8.1	1	3	7.5	7.3	6.0	4.4	2.8	2.6	32%
Hijo 2	0	3	2.4	4.8	5.5	8.2	9.6	10.0	1	2	0.3	3.0	5.3	6.0	7.9	9.1	1	2	8.5	8.0	5.0	4.3	3.2	1.4	38%
Hijo 3	1	1	3.7	4.5	5.9	6.9	7.2	8.8	1	2	1.5	4.3	5.9	6.0	7.2	9.3	0	2	7.0	6.8	6.6	4.1	3.8	2.9	47%
Hijo 4	0	3	0.6	3.2	3.3	5.8	8.1	8.2	0	3	3.5	3.6	5.8	7.4	7.8	8.5	0	3	9.1	8.4	8.4	2.1	1.3	0.1	26%
Hijo 5	0	2	3.0	4.6	5.3	8.5	8.5	8.8	0	2	0.9	3.2	5.3	7.0	9.9	10.0	1	3	8.1	8.0	7.8	2.4	1.4	0.9	32%
Individuo 6	0	3	2.4	4.8	5.2	8.2	9.6	9.9	1	3	0.3	3.0	4.8	5.6	8.1	9.1	0	3	8.5	8.0	5.6	4.3	2.8	1.4	50%

1.8 Criterio de terminación del algoritmo genético: el criterio de terminación en este ejemplo son 1000 generaciones, por lo que todo el proceso anterior se repite 1000 veces. Finalmente se llega a la última población, y a continuación se muestra que el mejor individuo de esta población final es el número 2 con 64% de valor de aptitud.

	Part.	Peso	B1	B2	B3	B4	B5	B6	Part.	Peso	B1	B2	B3	B4	B5	B6	Part.	Peso	B1	B2	B3	B4	B5	B6	Power Stat
Individuo final 1	0	3	0.9	5.2	6.3	7.2	8.4	9.5	1	2	3.1	4.7	5.1	6.8	7.4	8.5	1	3	7.9	7.7	6.3	4.6	3.0	2.7	58%
Individuo final 2	1	3	2.5	5.0	5.8	8.7	9.0	9.3	0	2	0.3	3.2	5.5	6.3	8.3	9.6	1	1	8.9	8.5	5.3	4.5	3.8	1.5	64%
Individuo final 3	1	1	3.9	4.7	6.2	7.2	7.6	9.2	1	3	1.5	4.6	6.2	6.3	7.6	9.8	0	2	7.4	7.2	6.9	4.3	3.6	2.6	57%
Individuo final 4	0	3	0.6	3.4	3.4	6.1	8.5	8.6	1	2	3.7	3.8	6.1	7.8	8.2	9.0	0	3	9.5	8.9	8.9	2.3	1.4	0.1	59%
Individuo final 5	1	2	3.1	4.9	5.6	8.9	9.0	9.3	0	2	1.0	3.4	5.6	7.4	8.0	9.4	1	3	8.6	8.4	8.2	2.5	1.4	0.9	60%
Individuo final 6	0	3	2.5	5.0	5.4	8.7	9.0	9.7	1	3	0.3	3.2	5.1	5.9	8.5	9.6	0	3	8.9	8.5	5.9	4.5	3.0	1.5	53%

Con lo que finalmente obtenemos el modelo de *scoring* óptimo, mostrado a continuación:

N°Ratio	Ratio	Peso	Rango 1	Rango 2	Rango 3	Rango 4	Rango 5	Rango 6
1	EBITDA / Cargas Fin	3	2.48	5.04	5.81	8.68	9.00	9.30
3	Pasivo / Activo Total	1	8.93	8.47	5.27	4.51	3.8	1.51

2. Algoritmo de *Clustering*

2.1 *Clusters* iniciales: para iniciar el proceso de *clustering*, primero deben de calcularse los *scores* de cada registro (de acuerdo al modelo de *scoring* definido con el algoritmo genético) y ordenarse de menor a mayor.

Como ejemplo, mostraremos el cálculo del score del primer individuo de acuerdo a la Ecuación 2: $\boldsymbol{score} = \sum_{i=1}^{N} \boldsymbol{Participación_i . Ponderación_i . Puntaje_i}$

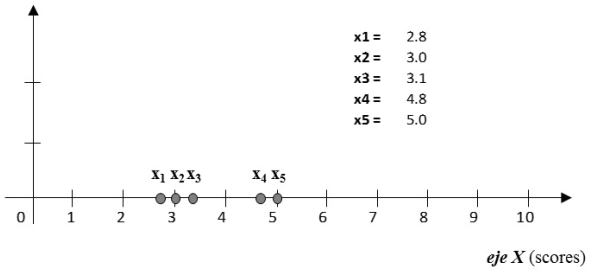
El primer individuo tiene como valores de entrada:

	EBITDA / Cargas Fin	UN / Patrimonio	Pasivo / Activo
Registro 1	-6.16	0.25	0.75

Ubicando los valores del registro 1 en los *buckets* del modelo de *scoring* se obtiene el score de este individuo. Como el ratio 2 no participa en el modelo, no entrará en el cálculo:

$$\boldsymbol{Score = 1 . \frac{3}{4} . 2.48 + 1 . \frac{1}{4} . 3.8 = 2.8}$$

A continuación, se muestran los *scores* de los 5 registros de la data de entrada. Cada *score* se constituye como un *cluster* inicial:



2.2 Matriz de disimilitud inicial: se calcula la matriz de disimilitud a través de la resta simple de los *scores* (de acuerdo a la definición de la sección 4.2.2). A continuación, se muestra el resultado:

	x1	x2	x3	x4	x5
x1	0	0.2	0.3	2	2.2
x2	0.2	0	0.1	1.8	2
x3	0.3	0.1	0	1.7	1.9
x4	2	1.8	1.7	0	0.2
x5	2.2	2	1.9	0.2	0

2.3 Método de distancia para la fusión de *clusters*: en este trabajo se utiliza el método de distancia promedio ponderada (group average-link) en la fusión de *clusters*. Aplicando este método al conjunto de datos, se tiene que en la primera etapa se fusiona $\mathbf{x}_2$ y $\mathbf{x}_3$. Tras fusionar $\mathbf{x}_2$ y $\mathbf{x}_3$, las distancias entre el grupo $\{\mathbf{x}_2, \mathbf{x}_3\}$ y los puntos restantes son:

$$D(\{\mathbf{x_2}, \mathbf{x_3}\}, \mathbf{x_1}) = \frac{1}{2} d(\mathbf{x_1}, \mathbf{x_2}) + \frac{1}{2} d(\mathbf{x}_1, \mathbf{x_3}) = 0.25$$

$$D(\{\mathbf{x_2}, \mathbf{x_3}\}, \mathbf{x_4}) = \frac{1}{2} d(\mathbf{x_2}, \mathbf{x_4}) + \frac{1}{2} d(\mathbf{x_3}, \mathbf{x_4}) = 1.75$$

$$D(\{\mathbf{x_2}, \mathbf{x_3}\}, \mathbf{x_5}) = \frac{1}{2} d(\mathbf{x_2}, \mathbf{x}_5) + \frac{1}{2} d(\mathbf{x_3}, \mathbf{x_5}) = 1.95$$

Luego de que $\mathbf{x}_2$ y $\mathbf{x}_3$ han sido fusionados, la matriz de disimilitud queda así:

	x1	{x2,x3}	x4	x5
x1	0	0.25	2	2.2
{x2,x3}	0.25	0	1.75	1.95
x4	2	1.75	0	0.2
x5	2.2	1.95	0.2	0

En la segunda etapa del algoritmo, $\mathbf{x}_4$ y $\mathbf{x}_5$ se fusionan. Las distancias entre el grupo $\{\mathbf{x}_4, \mathbf{x}_5\}$ y los grupos restantes se actualizan como sigue:

$$\boldsymbol{D}(\{\mathbf{x_4}, \mathbf{x_5}\}, \mathbf{x_1}) = \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{d}(\mathbf{x_4}, \mathbf{x_1}) + \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{d}(\mathbf{x_5}, \mathbf{x_1}) = \mathbf{2.1}$$

$$\boldsymbol{D}(\{\mathbf{x_2}, \mathbf{x_3}\}, \{\mathbf{x_4}, \mathbf{x_5}\}) = \frac{\mathbf{1}}{\mathbf{4}}[\, \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_4}) + \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_5}) + \boldsymbol{d}(\mathbf{x_3}, \mathbf{x_4}) + \boldsymbol{d}(\mathbf{x_3}, \mathbf{x_5})\,] = \mathbf{1.85}$$

Luego de que $\mathbf{x}_4$ y $\mathbf{x}_5$ han sido fusionados, la matriz de disimilitud queda así:

	x1	{x2,x3}	{x4,x5}
x1	0	0.25	2.1
{x2,x3}	0.25	0	1.85
{x4,x5}	2.1	1.85	0

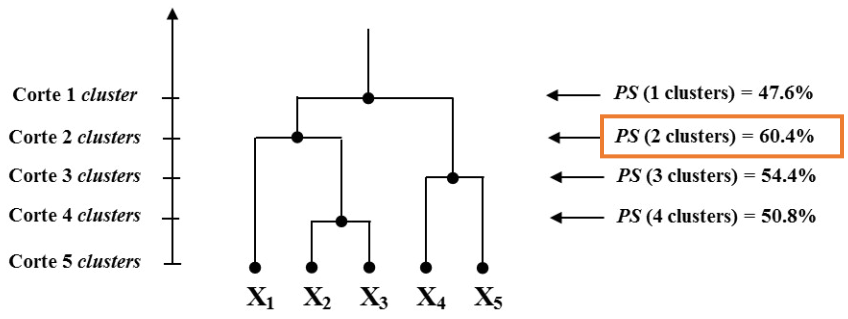
En la tercera etapa del algoritmo, $\{\mathbf{x}_2, \mathbf{x}_3\}$ y $\mathbf{x}_1$ se fusionan ya que tienen la menor distancia. La distancia entre $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ y $\{\mathbf{x}_4, \mathbf{x}_5\}$ queda así:

$$\boldsymbol{D}(\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}\}, \{\mathbf{x_4}, \mathbf{x_5}\}) = \frac{\mathbf{1}}{\mathbf{6}}[\boldsymbol{d}(\mathbf{x_1}, \mathbf{x_4}) + \boldsymbol{d}(\mathbf{x_1}, \mathbf{x_5}) + \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_4}) + \boldsymbol{d}(\mathbf{x_2}, \mathbf{x_5}) + \boldsymbol{d}(\mathbf{x_3}, \mathbf{x_4}) + \boldsymbol{d}(\mathbf{x_3}, \mathbf{x_5})] = \mathbf{1.93}$$

Y la matriz de disimilitud queda así:

	[x1,x2,x3]	[x4,x5]
[x1,x2,x3	0	1.93
[x4,x5]	1.93	0

El dendograma de este *clustering* se muestra a continuación:



2.4 Número de *clusters* final

Para la elección del corte final de clusterización, se calcula el ratio de *power statistic* para cada modelo de clusterización. En la figura anterior se observa que el corte de 2 *clusters* es el que obtiene el *power statistic* más elevado (60.4%). De acuerdo a ello, la siguiente tabla muestra el modelo final de *clusters*:

Grupo	Cantidad de Observaciones	Cantidad de incumplidos	% Probabilidad incumplimiento	Nota Mínima	Nota Máxima
A	2	0	0.00%	4.0	10.0
B	3	2	66.67%	0.0	4.0

ANEXO D

Esta sección muestra el diagrama de casos de uso del sistema en la Figura 27, así como la definición de cada caso de uso.

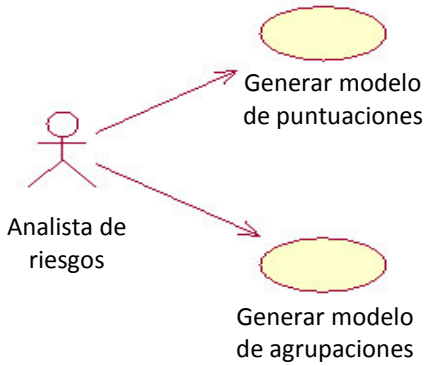


Figura 27. Diagrama de casos de uso del sistema. Fuente: Elaboración Propia

1) Caso de Uso: Generar Modelo de Puntuaciones

Caso de Uso	Generar Modelo de Puntuaciones
Objetivo	Generar el modelo de puntuaciones del modelo de *rating*
Actor	Analista de riesgos
Post-condición	Se obtiene el modelo de puntuaciones óptimo.
Flujo Principal	
1. El caso de uso se inicia cuando el usuario selecciona la opción "**Generar Modelo de *Rating***". 2. El sistema muestra la interfaz "**Generar Modelo de Puntuaciones**" con las siguientes opciones: "**Cargar Datos**", "**Generar solución**", "**Cargar Solución**", "**Guardar Solución**", "**Siguiente**", y también se muestra una tabla	

para los datos de entrada y una tabla para los datos de salida, así como un panel gráfico para mostrar el resultado gráficamente. 3. El usuario selecciona la opción "**Cargar Datos**" 4. El sistema muestra una interfaz donde se podrá elegir la ruta dónde se encuentra el archivo Excel con los datos para el entrenamiento del algoritmo genético. 5. El usuario ingresa la ruta del archivo Excel. 6. El sistema importa los datos del archivo y los muestra en una tabla. 7. El usuario selecciona la opción "**Generar solución**" 8. El sistema ejecuta un algoritmo genético para hallar la solución óptima. 9. El sistema muestra una tabla con el resultado del modelo de *rating*, la gráfica resultante del modelo y el Poder de predicción. 10. El usuario selecciona la opción "**Siguiente**" y el Caso de uso finaliza
Flujo Alternativo
1. En el paso 10, el usuario selecciona la opción "**Guardar Solución**" y el sistema almacena el modelo con sus parámetros en un archivo para su uso posterior. El caso usuario puede seguir con la opción "**Siguiente**", o caso contrario el caso de uso finaliza. 2. En el paso 7, el usuario selecciona la opción "**Cargar Solución**" y el sistema restaura los parámetros del modelo de puntuaciones y grafica los resultados en el panel gráfico. El usuario puede seguir con la opción "**Siguiente**", o caso contrario el caso de uso finaliza. 3. Se puede salir del caso de uso en cualquier momento.

2) Caso de Uso: Generar Modelo de Agrupaciones

Caso de Uso	Generar Modelo de Agrupaciones
Objetivo	Generar el modelo de agrupaciones del modelo de *rating*
Actor	Analista de riesgos

Pre-condición	Se ha generado previamente un modelo de puntuaciones
Post-condición	Se obtiene el modelo de grupos de riesgo, generándose por último el modelo de *rating* final

Flujo Principal

1. El caso de uso se inicia cuando el usuario selecciona la opción "**Siguiente**" de la interfaz "**Generar modelo de puntuaciones**".
2. El sistema muestra la interfaz "**Generar modelo de *ratings***" con las siguientes opciones: "**Guardar Modelo**", "**Generar Nuevo *Clustering***", "**Anterior**", una caja de texto "**Nro. Grupos**" y también una tabla para los datos del modelo de grupos de riesgo y una tabla con el modelo de puntuaciones.
3. Apenas se muestra la interfaz "**Generar modelo de *ratings***", el sistema debe ejecutar el algoritmo de *clustering* jerárquico aglomerativo para hallar el modelo de grupos de riesgo.
4. Una vez ha terminado el algoritmo, el sistema completa las tablas con los modelos de grupos de riesgo y el modelo de puntuaciones. Adicionalmente, el sistema muestra la gráfica resultante del modelo y el Poder de predicción
5. El usuario selecciona la opción "**Guardar Modelo**", donde el sistema almacena en un archivo el modelo de *rating* completo (el modelo de puntuaciones y el modelo de grupos de riesgo).
6. El usuario selecciona la opción "**Salir**" y el Caso de uso finaliza

Flujo Alternativo

1. En el paso 5, el usuario selecciona la opción "**Anterior**" y regresa a la interfaz "**Generar modelo de puntuaciones**".
2. En el paso 5, el usuario puede volver a ejecutar el algoritmo de *clustering* jerárquico pero con una cantidad diferente de grupos de riesgo. Para ello modifica el número de grupos en la caja de texto "**Nro.Grupos**" y selecciona la opción "**Generar Nuevo *Clustering***".
3. Se puede salir del caso de uso en cualquier momento.