

# **ESCUELA POLITÉCNICA NACIONAL**

## **FACULTAD DE CIENCIAS**

### **“HERRAMIENTAS PARA LA ADMINISTRACIÓN BANCARIA MEDIANTE LA UTILIZACIÓN DE MODELOS DE MIXTURA Y REDES BAYESIANAS”**

**PROYECTO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO  
EN CIENCIAS ECONÓMICAS Y FINANCIERAS**

**LOREIMA LOURDES CHIRIBOGA HERNÁNDEZ**  
loreima.chiriboga@gmail.com

**GINA KATHERYNE TORRES CUMBICUS**  
ginatorres17@gmail.com

**DIRECTOR: ING. DIEGO MALDONADO, M.Sc.**  
diego.maldonado6@gmail.com

**CO DIRECTOR: DR. JULIO MEDINA**  
julio.medina@epn.edu.ec

**Quito, Julio 2013**

## **DECLARACIÓN**

Nosotras, Loreima Lourdes Chiriboga Hernández y Gina Katheryne Torres Cumbicus, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que hemos consultado las referencias bibliográficas que se incluyen en este documento.

La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

---

**Loreima Lourdes  
Chiriboga Hernández**

---

**Gina Katheryne  
Torres Cumbicus**

## **CERTIFICACIÓN**

Certificamos que el presente trabajo fue desarrollado por Loreima Lourdes Chiriboga Hernández y Gina Katheryne Torres Cumbicus.

---

**Diego Maldonado, Msc.**  
**DIRECTOR DEL PROYECTO**

---

**Dr. Julio Medina**  
**CO DIRECTOR DEL PROYECTO**

## AGRADECIMIENTOS

*Gracias!!! ... una maravillosa palabra que utilizamos para expresar gratitud especial; pero a veces esa palabra no puede expresarlo todo. Gracias significa tanto...*

*Gracias a Dios y a la Virgen Dolorosa por tenerme en la palma de su mano, por permitirme concluir este gran sueño; por regalarme momentos felices y por darme valor, perseverancia y fortaleza para afrontar los momentos difíciles.*

*Gracias a mis padres, porque son el mejor regalo que la vida me pudo dar. Por su amor, apoyo incondicional y sabios consejos; por ser mi guía y ejemplo. Por depositar su confianza en mí sin dudar ni un solo instante. Sin ellos jamás hubiese podido conseguir lo que hasta ahora.*

*Gracias a mis hermanas por siempre estar a mi lado; por llenar mi vida de mágicos momentos y grandes alegrías.*

*Mi eterna gratitud a Diego, por ser un excelente director; pero sobretodo una gran persona. Gracias por sus palabras exactas en el momento oportuno, por su paciencia, tiempo dedicado, disponibilidad y generosidad para compartir su experiencia y amplio conocimiento.*

*Mi agradecimiento profundo a Gina por los gratos momentos compartidos, por brindarme su amistad; por el constante apoyo y optimismo.*

*Finalmente, son tantas las personas maravillosas a las que me gustaría agradecer, pero estas líneas no son suficientes para nombrar a todas, porque sería una gran lista. Sin embargo, quiero expresarles mi agradecimiento por su amistad, sugerencias, tiempo, apoyo, o simplemente por su gesto o palabra de ánimo.*

*Mil gracias a todos por su cariño y por inundar mi vida de luz, alegría y risas...*

*Merci Beaucoup!!!*

*Loreima*

## AGRADECIMIENTOS

*A mis padres, por ser los mejores. A Dios por dármelos.*

*A toda mi familia, por ser quienes siempre han estado a mi lado y de manera especial a mis hermanas Yuri y a mi Nena Cindy por su cariño y apoyo.*

*A Diego, por su tiempo, confianza, paciencia, disponibilidad para el desarrollo de esta investigación.*

*A quienes siempre han estado cerca de mí. En la Poli encontré amigos, amigas de los que duran para toda la vida. Lore, gracias por la paciencia ¡lo logramos!*

*Finalmente un recuerdo especial y gracias a quién ya no está en este mundo terrenal, pero que siempre ha estado acompañándome y regalándome luz cada día.*

*Gina*

## DEDICATORIA

*A Dios y a la Virgen Dolorosa por estar conmigo en cada paso que doy, porque con su infinito amor han permitido que mis sueños y anhelos se cumplan. Por la bendiciones recibidas al haber puesto en mi camino a excelentes personas que llenan mi vida de alegría y felicidad.*

*A mis padres, Anita y Ricardo por ser la luz de mi vida; por brindarme un hogar maravilloso; y porque me han enseñado que el triunfo no está en vencer siempre, sino en nunca rendirse.*

*A mis queridas hermanas Taty y Kelly por su comprensión y cariño; por darle color a mi vida y por siempre estar en los momentos más importantes.*

*Loreima*

*A mis padres, por su amor desde siempre.*

*Gina*

# Índice general

<b>1. INTRODUCCIÓN</b>	<b>7</b>
<b>2. EL UNIVERSO DE MODELOS DE RIESGO CREDITICIO</b>	<b>9</b>
2.1. MODELOS DE RIESGO DE CRÉDITO TRADICIONALES . . . . .	12
2.1.1. CREDIT SCORING . . . . .	12
2.1.2. SISTEMAS EXPERTOS . . . . .	14
2.1.3. REDES NEURONALES . . . . .	15
2.2. MODELOS PROPUESTOS POR EL ACUERDO DE BASILEA . . .	16
2.3. MODELOS CON UN ENFOQUE MODERNO . . . . .	17
2.3.1. MODELO KMV . . . . .	17
2.3.2. CREDITMETRICS . . . . .	18
<b>3. REDES BAYESIANAS: ANTECEDENTES Y DESARROLLO</b>	<b>20</b>
3.1. INTRODUCCIÓN . . . . .	20
3.2. APLICACIONES E IMPORTANCIA DE LAS REDES BAYESIANAS	
EN LA TOMA DE DECISIONES . . . . .	23
3.2.1. APLICACIONES . . . . .	23
3.2.2. IMPORTANCIA EN LA TOMA DE DECISIONES [31] . . . .	27
<b>4. FUNDAMENTACIÓN TEÓRICA</b>	<b>28</b>
4.1. INTRODUCCIÓN . . . . .	28
4.1.1. MODELO LINEAL MIXTO . . . . .	31
4.2. MODELOS LINEALES GENERALIZADOS (GLM) . . . . .	33
4.2.1. SUPUESTOS BÁSICOS . . . . .	34

4.2.2. ESTRUCTURA DEL MODELO . . . . .	35
4.3. MODELOS LINEALES GENERALIZADOS MIXTOS (GLMMs) <sup>1</sup> . .	37
4.3.1. INTRODUCCIÓN . . . . .	37
4.3.2. ESTRUCTURA DEL MODELO . . . . .	42
4.3.3. MODELOS DE MIXTURA DE BERNOULLI [43] . . . . .	44
4.3.4. MODELOS DE MIXTURA DE POISSON . . . . .	45
4.4. REDES BAYESIANAS . . . . .	46
4.4.1. INTRODUCCIÓN . . . . .	46
4.4.2. TEOREMA DE BAYES . . . . .	47
4.4.3. DEFINICIÓN DE RED BAYESIANA . . . . .	48
4.5. TIPOS DE REDES BAYESIANAS . . . . .	52
4.6. APRENDIZAJE DE REDES BAYESIANAS . . . . .	55
4.6.1. APRENDIZAJE PARAMÉTRICO . . . . .	55
4.6.2. APRENDIZAJE ESTRUCTURAL . . . . .	55
4.7. CLASIFICADORES BAYESIANOS . . . . .	56
4.7.1. NAIVE BAYES . . . . .	57
4.7.2. TREE ARGUMENTED NAIVE BAYES (TAN) . . . . .	58
4.7.3. VALIDACIÓN DE CLASIFICADORES . . . . .	60
4.7.4. SELECCIÓN DE VARIABLES . . . . .	64
4.8. ALGORITMO CHAID . . . . .	66
4.8.1. INTRODUCCIÓN . . . . .	66
4.8.2. METODOLOGÍA CHAID [55] . . . . .	67
4.8.3. DESCRIPCIÓN DE LA TÉCNICA CHAID[56] . . . . .	68
4.8.4. PRUEBA DE INDEPENDENCIA CHI - CUADRADA . . . . .	72
<b>5. METODOLOGÍA PARA LA CONSTRUCCIÓN DE UN SCORING</b>	
<b>DE APROBACIÓN</b>	<b>74</b>
5.1. INTRODUCCIÓN . . . . .	74
5.2. METODOLOGÍA . . . . .	75
5.2.1. SELECCIÓN DE LA MUESTRA DE ANÁLISIS . . . . .	75

---

<sup>1</sup>Por sus siglas en inglés



---

<b>ÍNDICE GENERAL</b>	<b>3</b>
-----------------------	----------

---

5.2.2. SELECCIÓN DE CLIENTES BUENOS Y MALOS . . . . .	78
5.2.3. SELECCIÓN DE GRUPOS HOMOGÉNEOS . . . . .	80
5.2.4. SELECCIÓN DEL MODELO . . . . .	85
5.2.5. VALIDACIÓN DEL MODELO . . . . .	88
5.2.6. PERFIL DE CLIENTES . . . . .	91
5.2.7. CONSISTENCIA DE NEGOCIO DEL MODELO . . . . .	92
<b>6. CONCLUSIONES Y RECOMENDACIONES</b>	<b>96</b>
<b>7. ANEXOS</b>	<b>99</b>
7.1. SELECCIÓN DE LA MUESTRA DE CLIENTES BUENOS . . . . .	99
7.2. ÍNDICE ADELANTADO DE LA ECONOMÍA ECUATORIANA (IACCE)100	
7.2.1. INTRODUCCIÓN . . . . .	100
7.3. INDICADOR ADELANTADO DEL CICLO ECONÓMICO PARA EL ECUADOR . . . . .	103
7.3.1. ASPECTOS METODOLÓGICOS PRELIMINARES . . . . .	103
7.3.2. ELABORACIÓN DEL ÍNDICE MENSUAL ADELANTADO DEL CICLO ECONÓMICO PARA EL ECUADOR . . . . .	109
7.4. PROBABILIDADES DEL ÍNDICE ADELANTADO DE LA ECONOMÍA ECUATORIANA . . . . .	114
7.5. DESCRIPCIÓN DE LAS VARIABLES DISCRETAS A UTILIZAR EN EL MODELO . . . . .	115
7.6. GRUPOS HOMOGÉNEOS OBTENIDOS MEDIANTE METODOLOGÍA CHAID . . . . .	118

# Índice de cuadros

5.1. Número de clientes Buenos Malos . . . . .	79
5.2. Variables seleccionadas mediante Clasificadores Bayesianos . . . . .	82
5.3. Resultados de Modelos de Clasificadores Bayesianos, expresados en Ma- triz de Confusión. . . . .	83
5.4. Resultados de Modelos de Clasificadores Bayesianos, en términos de la sensibilidad, la especificidad, valor de la predicción de clientes buenos y clientes malos. . . . .	83
5.5. Cuadro resumen de los tres grupos crediticios homogéneos de una cartera de consumo creados a partir del método CHAID . . . . .	84
5.6. Cuadro acumulado de los grupos crediticios homogéneos de una cartera de consumo creados a partir del método CHAID . . . . .	85
5.7. Cuadro resumen de los tres grupos crediticios homogéneos de una cartera crediticia de consumo creados a partir del método CHAID. . . . .	85

# Índice de figuras

4.1. Ejemplo de Red Bayesiana . . . . .	51
4.2. Probabilidades condicionales ejemplo red bayesiana . . . . .	51
4.3. Estructura Naive Bayes . . . . .	57
4.4. Estructura de Tree Augmented Naive Bayes . . . . .	58
4.5. Criterio Inicial - CHAID . . . . .	68
4.6. Criterio 2 - CHAID . . . . .	69
5.1. Tasa de morosidad y créditos colocados . . . . .	76
5.2. Prueba CUSUM recursiva para el índice de tasa de morosidad . . . . .	77
5.3. Red Inicial para seleccionar variables mediante algoritmo TAN . . . . .	81
5.4. Red Final para seleccionar variables mediante algoritmo TAN . . . . .	81
5.5. Curvas ROC . . . . .	84
5.6. Residuos de los modelos desarrollados. . . . .	89
5.7. Perfilamiento de clientes en función de la tasa de morosidad . . . . .	92
5.8. Modelo “Pico”: % Morosidad - puntaje score . . . . .	92
5.9. Modelo “Valle”: % Morosidad - puntaje score . . . . .	93
5.10. Evolución PIB trimestral periodo 2006.I-2008.IV. Fuente: Banco Central del Ecuador . . . . .	94
5.11. Consistencia de Negocio del Modelo - Análisis Fase de Auge . . . . .	95
5.12. Consistencia de Negocio del Modelo -Análisis Fase de Contracción . . . . .	95
7.1. Períodos del Ciclo Económico - PIB . . . . .	112
7.2. IACCE - Pico y Valle . . . . .	113

---

7.3. Descripción de Variable Rango. . . . .	115
7.4. Descripción de la Variable: Estado Civil . . . . .	115
7.5. Descripción de la variable Nivel de Educación. . . . .	116
7.6. Descripción de la Variable Vivienda . . . . .	116
7.7. Descripción de la variable Género . . . . .	117
7.8. Descripción de la Variable Residencia del Cliente. . . . .	117
7.9. Selección de Grupos Homogéneos mediante metodología CHAID . . .	118

## RESUMEN

En el análisis de crédito tradicional, se asume que las pérdidas individuales de crédito son independientes unas de otras, es decir que un cliente incumpla no significa que otros también lo harán. Sin embargo, es importante analizar el comportamiento de la cartera en su conjunto con la finalidad de estudiar el riesgo al que está sujeta la institución.

Aunque existen factores que inciden en las probabilidades de incumplimiento, es difícil determinar una relación que indique cómo el incumplimiento de un crédito predispone el incumplimiento de otro. Así, cuando la situación económica del país es una recesión probablemente provocará un aumento en las probabilidades de incumplimiento. Por lo tanto, es necesario tomar en cuenta la dependencia existente entre las probabilidades de incumplimiento mediante un factor estocástico sujeto a distintos factores de riesgo.

Así, en este estudio se utilizó un modelo lineal generalizado mixto que parte del hecho de que los eventos de incumplimiento por parte de los deudores son eventos dependientes, es decir se rompe con el supuesto del modelo logit tradicional de independencia entre los individuos analizados. Por lo tanto, la probabilidad de incumplimiento del portafolio depende de un conjunto de factores económicos, dando como resultado que el incumplimiento entre los clientes sea condicionalmente independiente entre ellos.

Dentro de la modelización que se planteó anteriormente una parte crucial es la selección de variables explicativas es decir aquellas que definen al cliente como “bueno” o “malo”, dado que las variables escogidas deben ser suficientes para explicar a la variable dependiente, pero no deben ser demasiadas tal que compliquen el modelo. Se debe realizar un análisis de las variables en conjunto para conocer cada uno de sus atributos, el objeto final debe ser identificar el grado con el que pueden contribuir en el modelo para discriminar buenos y malos. Así en la presente investigación, se utilizará la técnica de minería de datos de redes bayesianas como herramienta para la selección de variables.

## ABSTRACT

In traditional credit analysis it is assumed that individual credit losses are independent from one another: one debtor falling in default does not mean others will as well. However, it is import to analyze the portfolio behavior as a whole to determine the risk that an institution is subject to.

Although there are factors that affect the probability of default, it is difficult to determine a relation that describes how default in one credit predisposes default in another. An economy being in recession will probably cause a raise in the probabilities of default. Therefore, it is necessary to consider the dependency among the probabilities of default by means of a stochastic factor subject to different risk aspects.

This study uses a generalized linear mixed model, inferred from the assumption that defaults from debtors are dependent events; dropping the assumption of independency between analyzed subjects from the traditional logit model. Therefore, the portfolio default probability depends on several economic factors; as a result, default among clients is conditionally independent.

A crucial part within the proposed modeling is the selection of explicative variables; those who define the client as “good” or “bad”. The chosen variables must be enough to explain the dependent variable, but not too many as to complicate the model.

A group variable analysis must be performed to get to know each of the variable’s attributes and to identify their contribution to the model in distinguishing between good and bad clients. This research uses a bayesian network for data mining technique as a tool for selecting such variables.

## Capítulo 1

# INTRODUCCIÓN

El sistema financiero juega un rol fundamental en el funcionamiento de la economía. Instituciones financieras sólidas y solventes permiten que los recursos financieros fluyan eficientemente desde los agentes superavitarios a los deficitarios permitiendo que se aprovechen las oportunidades de negocios y de consumo. La principal actividad generadora de ingresos de las instituciones bancarias, la representa las operaciones activas a través de los créditos otorgados a los solicitantes, entendidos como el capital adeudado, los rendimientos por cobrar y cualquier otro monto que represente derechos no recaudados por una operación crediticia. Sin embargo, en los últimos años, la globalización de la economía de mercado ha disminuido la estabilidad y seguridad del negocio financiero, multiplicando los riesgos a los que se enfrentan estas entidades.

Dadas las funciones de intermediación que cumple el sistema financiero en la economía, siempre existirán riesgos, específicamente, riesgo de crédito.

El riesgo de crédito es así uno de los más importantes que enfrenta el sistema financiero y por esta razón debe ser medido y monitoreado de manera eficiente y de acuerdo con las características de la economía y del momento del ciclo económico.

En cuanto a la medición del riesgo, hay una gran variedad de modelos de riesgo de crédito, los cuales pueden ser clasificados a partir de distintos criterios. De acuerdo con Elizondo (2004), la diversificación del riesgo en un portafolio de créditos se da de manera natural al tener un gran número de firmas o individuos; sin embargo si varios

individuos están afectados por los mismos factores de riesgo, esta diversificación puede ser insuficiente, créditos que pertenecen al mismo sector económico, al mismo lugar de destino pueden generar altas correlaciones entre sus probabilidades de incumplimiento. Así, los factores de riesgo se clasifican en *sistémicos*, que son los que afectan a un grupo de la cartera de crédito y en *no sistémico*, que son los que afectan particularmente a un crédito de la cartera. En la medida en que los factores de riesgo *sistémico* afectan a un gran grupo de créditos del portafolio, dicho portafolio experimenta una gran concentración de riesgo, pues cambios no deseados pueden llevar a pérdidas extremas.

Al respecto, se han planteado los modelos de mixtura (*GLMM Generalized linear mixed models*) para modelar la dependencia considerando que el incumplimiento de un deudor depende de un conjunto de factores económicos que son modelados estocásticamente; así, la dependencia entre incumplimientos se la obtiene de la dependencia que tiene cada incumplimiento con el mismo conjunto de factores económicos.

En este contexto, es necesario mencionar que para el desarrollo de estos modelos es fundamental cuantificar la dependencia contemporánea entre créditos incumplidos y no condicionar a una distribución en particular, para lo cual se utilizará la técnica de redes bayesianas que no realizan supuestos sobre las distribuciones marginales de las variables, ni tampoco sobre la relación de dependencia entre ellas, permitiendo que capture la dinámica real entre las variables explicativas del modelo.

La presente investigación busca desarrollar una herramienta para mejorar la gestión de riesgo a través de la aplicación de un modelo scoring mixto que permita incorporar factores de riesgo sistémico, considerando dependencias de los clientes asociados a una cartera de consumo de una institución financiera. Para ello, en el capítulo II y capítulo III se realizará una breve descripción de los modelos de riesgo de crédito y la aplicación e importancia de las redes bayesianas en riesgo de crédito respectivamente. El capítulo IV constará de la fundamentación teórica de los modelos lineales generalizados mixtos, redes bayesianas y algoritmo CHAID. En el capítulo V se desarrollará la aplicación a una cartera de consumo de una institución financiera ecuatoriana. Finalmente, en el capítulo VI se presentarán las conclusiones y recomendaciones.



## Capítulo 2

# EL UNIVERSO DE MODELOS DE RIESGO CREDITICIO

El sector bancario es uno de los principales motores económicos de un país, ya que actúa como colocador de recursos económicos entre agentes superavitarios y deficitarios, desempeñando un importante papel en la movilización de fondos de los depositantes y en la concesión de créditos, canalizando así, recursos para inversión y ahorro [1].

La principal actividad de las instituciones financieras, a la que dedica la mayor parte de sus esfuerzos, y la que genera la mayor parte de sus beneficios es la actividad crediticia. Sin embargo, ésta afronta una serie de riesgos, entre ellos el riesgo de crédito. Definiéndose como el riesgo de que un cliente o contraparte no pueda o no quiera cumplir con un compromiso que ha celebrado con una institución financiera [2]; su medición contribuye a que el negocio bancario se efectúe bajo ciertos lineamientos que le permitan mantener una cartera de crédito lo más sana posible y así disminuir las posibles pérdidas por incumplimiento. La estabilidad del banco, así como la correcta colocación de los recursos, contribuyen en gran medida a la estabilidad macroeconómica del país, por ello es de gran importancia que las instituciones financieras logren incorporar correctamente las mediciones de los riesgos inherentes a su operación.

Sin embargo, la historia económica del mundo ha demostrado la existencia de crisis

financieras, por lo que ha surgido la necesidad de establecer una serie de mecanismos de monitoreo y control para las operaciones financieras, que son producto de diversas transacciones en el mercado mundial, que permitan dar señales que sean captadas con mayor precisión y de esta manera las entidades puedan anticiparse e incluso sobrevivir a las crisis; manteniendo una regulación interna de sus operaciones y una vigilancia permanente de las mismas, que impida su desbordamiento económico.

Es así, como el riesgo de crédito ha cobrado gran importancia durante las últimas décadas. Y los controles se convierten cada vez en una herramienta indispensable para las instituciones, ya que una inadecuada gestión del riesgo puede conllevar a una situación de insolvencia e insostenibilidad.

El riesgo de crédito es el principal riesgo del sector financiero. Por ello, la manera en que una institución financiera elige y administra su riesgo de crédito es un determinante importante para su rentabilidad y solidez.

De esta manera, como la exposición al riesgo de crédito sigue siendo la fuente más importante de los problemas en los bancos, éstos deberían estar conscientes de la necesidad de identificar, medir, monitorear y controlar el riesgo de crédito y de determinar si su capital es adecuado frente a estos riesgos y si tienen amplia compensación para los riesgos ocurridos.

En este sentido, la identificación del riesgo de crédito es el primer componente de una administración de riesgo eficiente. Las instituciones financieras deben contar con un sistema de gestión del riesgo de crédito que produzca una calificación correcta y oportuna. Para ello, la identificación de los factores de riesgo a los cuales está expuesta la institución constituye el pilar sobre el cual se edifica la gestión del riesgo.

El objetivo de la administración del riesgo de crédito es maximizar la tasa de rendimiento ajustada por el riesgo del banco, manteniendo la exposición al riesgo de crédito dentro de límites aceptables. Una adecuada administración de riesgos debe ayudar a la identificación de aquellas oportunidades de negocio, donde el equilibrio entre riesgo y beneficio sea atractivo, fortaleciendo la solidez y rentabilidad de las

instituciones, ya que es esencial para el éxito a largo plazo de cualquier organización bancaria.

Por tal motivo, las entidades buscan desarrollar eficientemente una administración de riesgos más precisa y a la vez, que se ajuste y cumpla con los cambios regulatorios locales e internacionales que demandan reportes más estandarizados y transparentes. Así como también, hacer innovaciones internas y externas encaminadas a buscar la excelencia operacional en todos los niveles de la entidad.

Hay que tomar en cuenta que el análisis de créditos se considera un arte ya que no hay esquemas rígidos para realizarlo, por el contrario es dinámico y exige creatividad por parte del oficial de crédito. Sin embargo, es importante dominar las diferentes técnicas de análisis de créditos y complementarlas con experiencia y buen criterio, asimismo es necesario contar con la información suficiente que permita minimizar el número de incógnitas para poder tomar la decisión correcta.

La evolución de la gestión del riesgo de crédito y el incremento de la competencia ha obligado a que las entidades financieras, miren más allá de las prácticas tradicionales, buscando nuevas tácticas y herramientas que los analistas identifiquen como las más idóneas. Es decir, desarrollar e implementar procesos internos para encontrar el balance entre el riesgo asumido, la optimización de recursos y ganancias, y la creación de sus propias reservas.

En la actualidad, la banca ha logrado desarrollar diversas y sofisticadas herramientas para la medición del riesgo, este hecho ayuda en gran medida a colocar sus activos de forma más eficiente y enfocarse en los productos más rentables. Con estas acciones, se logra disminuir el riesgo de pérdida y se prevén las condiciones económicas.

Asimismo las exigencias internacionales como el Nuevo Acuerdo de Capital de Basilea, presenta principios, medidas y acciones, con el objetivo de mejorar la adecuación del requerimiento de capital regulatorio para llevarlo a los niveles de riesgos asociados a las operaciones financieras con el propósito de prevenir pérdidas potenciales en las que podría incurrir. Haciendo mayor énfasis en los modelos internos de medición de riesgo de crédito de cada banco.

Así, se hace necesario analizar las técnicas modernas, debido a que éstas pueden ser un complemento del análisis de crédito tradicional. Ya que, a pesar de la creencia tradicional basada en no asumir riesgos o minimizarlos y rechazar aquellas operaciones que no ofrecían plenas garantías, la gestión moderna del riesgo de crédito establece como objetivo gestionar el riesgo para obtener una rentabilidad acorde con el nivel de riesgo asumido, manteniendo al mismo tiempo un capital adecuado y cumpliendo con la normativa [3]. Esto significa que una operación crediticia con una mayor probabilidad de impago, no tiene porque ser mal negocio, si se obtiene una rentabilidad que compensa suficientemente dicho riesgo.

La diversidad de modelos es tan amplia como bancos hay en el mundo, los intentos de formalización giran alrededor de dos ideas, a saber: los que tratan de captar la “intuición” de los expertos y sistematizarla aprovechando la tecnología, y los que se basan en técnicas econométricas o de análisis discriminante para llegar a un “número indicativo” de la calidad de crédito [4].

En este sentido, existen múltiples modelos tradicionales usados por las instituciones financieras a lo largo de la historia. Sin embargo, se debe tener en cuenta que muchas de las mejores ideas de los modelos tradicionales son usadas en los nuevos modelos. En general, se puede mencionar los sistemas expertos, las redes neuronales, los sistemas rating incluyendo los sistemas de rating internos de los bancos, y los sistemas de credit scoring como los más usados, debido a sus múltiples ventajas.

## **2.1. MODELOS DE RIESGO DE CRÉDITO TRADICIONALES**

### **2.1.1. CREDIT SCORING**

Los modelos tradicionales, o de inteligencia artificial, se basan principalmente en el análisis y experiencia. El credit scoring es el más utilizado hoy en día, porque es un sistema de evaluación automático, rápido, seguro y consistente para determinar el

otorgamiento de créditos, que, en función de toda la información disponible, es capaz de predecir la probabilidad de no pago, asociada a una operación crediticia. Ayuda en el proceso de evaluación, comportamiento y cobranza, siendo capaz de analizar en un tiempo mínimo, gran cantidad de información, en forma homogénea.

Los primeros sistemas de “scoring” se desarrollaron en la década de los cincuenta mediante la implementación de “scores” internos de comportamiento por parte de bancos pioneros en Estados Unidos, los cuales eran utilizados para la gestión de las cuentas de dichos bancos en base a la propia información que manejaban. Posteriormente surgieron los “scorings” de aceptación conforme a los cuales se ranqueaba a los solicitantes de crédito según la propia calificación de buenos o malos que efectuaba una determinada institución bancaria [5].

Así, el scoring se usa como una herramienta numérica que clasifica los clientes con el fin de discriminar entre ellos y garantizar que las decisiones sean objetivas y coherentes. Es decir, usa modelos estadísticos (por ejemplo modelos logit, probit) para transformar datos relevantes en medidas numéricas que guían las decisiones de crédito. De esta manera ayuda a la toma de decisiones más rápidas, menos costosas y más prudentes en términos estadísticos, reemplazando a los procesos de toma de decisiones manuales y subjetivos.

La construcción del modelo debe de estar orientada por un objetivo concreto. Así, los elementos básicos para la realización de un credit scoring son el conjunto de variables, la elección del modelo estadístico, los puntos asociados a un atributo particular en lo que respecta a un solicitante y el umbral o valor de corte.

Dependiendo de la aplicación es necesario distintos tipos de variables, por ejemplo para el caso de préstamos a individuos y pequeñas empresas se usan variables socioeconómicas o datos básicos del emprendimiento productivo, mientras que en el caso de grandes empresas se utilizan variables extraídas de los estados contables, información cualitativa acerca de la dirección, el sector económico, proyecciones del flujo de fondos, etc.

Cuando se trata de personas físicas se utilizan variables socio-económicas: edad, estado civil, número de hijos, nivel de ingresos, categoría laboral, historial de pagos. En este caso, su principal aplicación se centra en créditos de consumo o hipotecas.

De esta manera la selección de variables se convierte en uno de los pilares fundamentales para la construcción de un credit scoring y, reducir el número de variables a las más relevantes es un reto. Aunque intuitivamente más variables traen más información esto no es del todo cierto. En realidad un número mayor de variables, aumenta el “volumen” del espacio muestral. Entonces el tamaño de la muestra como proporción del tamaño del espacio muestral se reduce a medida que aumenta el número de variables. Esto reduce la representatividad de la muestra, lo que impide una buena adaptabilidad del modelo.

Por lo mencionado anteriormente, seleccionar un conjunto de variables reducidas, que contengan la mayor cantidad de información relevante, es un paso crucial para obtener un buen modelo. Entonces, para seleccionar dichas variables se utiliza una combinación de técnicas estadísticas y modelos gráficos es decir *redes bayesianas*.

### 2.1.2. SISTEMAS EXPERTOS

Por otro lado, los sistemas expertos tratan de simular el proceso realizado por el analista financiero o de crédito para adoptar decisiones de crédito. El sistema más conocido es las 5 C's, donde además de la experiencia del analista, mediante la observación de distintos estándares consistentes y subjetivos, se analiza el carácter del potencial prestatario, su capital actual, la capacidad de pago o generación de recursos para el cumplimiento de la obligación pretendida, el colateral o la garantía ofrecida como fuente de repago del crédito y las condiciones del entorno o ciclo económico en el cual desarrolla sus actividades [6].

Sin embargo la principal desventaja de este sistema es la subjetividad, debido a que involucra el criterio subjetivo de cada analista, por ejemplo lo que para un analista es

un factor importante para otro no lo es, mismo que se basa en valoraciones de acuerdo al juicio y experiencia adquirida en la asignación de créditos.

### **2.1.3. REDES NEURONALES**

Otro de los métodos para la cuantificación del riesgo de crédito son las redes neuronales que simulan el proceso de aprendizaje humano; y a través de técnicas econométricas crean un modelo de decisión que emula una red de neuronas, capaz de asumir decisiones en torno al riesgo de referencia. Es una herramienta muy útil de clasificación o predicción ya que pueden incorporar información subjetiva, no cuantificable en las decisiones de aprobación de crédito.

La aplicación de redes neuronales ayuda a la obtención de resultados con mayor precisión que el credit scoring [7].

Además de poseer una ventaja especial sobre los sistemas expertos, cuando los datos son ruidosos o incompletos. Y se han vuelto bastante comunes en el modelado de fraude, cuando interpretabilidad no es tan importante.

Sin embargo, una desventaja importante de las redes neuronales es su falta de transparencia. La estructura interna de la red está oculta y no puede ser fácil de duplicar, incluso utilizando los mismos datos como entradas. Es decir, la red neuronal no revela nada acerca de los pasos intermedios que conducen al resultado final, por lo tanto los pasos intermedios del sistema no se pueden comprobar. Además, las técnicas de regresión tienen una ventaja sobre las redes neuronales, porque hacen que sea más fácil de interpretar y explicar el efecto de las características en la variable de resultado, puesto que es fácil determinar el efecto incremental de un cambio determinado para cada característica.

## 2.2. MODELOS PROPUESTOS POR EL ACUERDO DE BASILEA

De esta manera el escenario actual ha obligado a establecer normas internacionales para la supervisión del riesgo de crédito; es así como el surge el primer Acuerdo de Basilea, en el que se determinó los primeros parámetros regulatorios sobre la gestión del riesgo crediticio de los bancos comerciales. El Comité definió que la metodología sería la cobertura del riesgo a través de un capital mínimo a mantener [8].

Sin embargo, el acuerdo de Basilea I tuvo como principal limitación la definición de exposiciones de crédito, en la que ignoraba la calidad crediticia y, por lo tanto, la diversa probabilidad de incumplimiento de los distintos prestatarios.

Entonces para superar las deficiencias de dicha regulación se presenta el nuevo Acuerdo de Basilea sobre capital bancario, mismo que hace que los requerimientos de capital sean más sensibles a la exposición del riesgo, especialmente a la del riesgo de crédito. Planteando para ello tres modelos: el modelo estándar, el basado en modelos internos, y el método avanzado.

El método estándar se basa en el peso del riesgo de crédito que se determina mediante las calificaciones externas asignadas por agencias independientes de evaluación financiera. Es una versión modificada del Acuerdo de Basilea I, en donde se da una mayor granularidad al portafolio de crédito con el fin de asignar un ponderador de riesgo más adecuado al perfil de riesgo que tenga un portafolio en específico [9].

En este método el banco asigna una ponderación de riesgos a cada uno de sus activos y operaciones fuera de balance y genera un total de activos ponderados por riesgo. Actualmente, las ponderaciones individuales dependen del tipo, de prestatario. En el nuevo Acuerdo para las ponderaciones de riesgo se tienen en cuenta la calificación suministrada por la institución externa de evaluación de crédito (como una agencia de calificación) [10].



Los métodos basados en modelos internos exigen a los bancos internacionales formular sus propios modelos de calificación con el fin de clasificar el riesgo de crédito de sus actividades. Así el banco estima la calidad crediticia de sus prestatarios y los resultados se traducen en estimaciones del importe de las pérdidas futuras potenciales, importe que constituye la base de los requerimientos mínimos de capital [11].

La clasificación de las exposiciones a los riesgos dependerá de las propias estimaciones internas de los riesgos que realice el banco.

El método avanzado requiere que los bancos estimen la probabilidad de incumplimiento, la exposición al incumplimiento para cada transacción. Tomando en cuenta para ello las estimaciones de la pérdida dado incumplimiento y la madurez [12].

Sin embargo este método involucra la estimación de parámetros que requieren muchos datos históricos que no están disponibles para la mayoría de los bancos.

## **2.3. MODELOS CON UN ENFOQUE MODERNO**

Como se mencionó anteriormente, en la actualidad han surgido varios modelos modernos para la medición del riesgo de crédito, mismos que son más sofisticados e incluyen un mayor número de variables para su cálculo. Entre ellos están los modelos KMV, los modelos de forma reducida, CreditMetrics entre otros.

### **2.3.1. MODELO KMV**

Uno de los modelos más utilizados es KMV, el cual fue desarrollado a inicios de la década de los noventa por la calificadora Moodys [13]. Es probablemente la metodología que trata de llevar a la práctica el modelo de Merton de la forma más directa, que toma en cuenta el comportamiento crediticio de los deudores.

Este es un modelo de diversificación basado en las correlaciones del mercado de acciones que permite estimar la probabilidad de incumplimiento entre activos y pasivos

[14]. Se define una “distancia al impago”, que es la diferencia entre el valor de los activos de las empresas y cierto umbral de pasivos, de manera que si esta cantidad es negativa, la compañía estaría en bancarrota y no podría cumplir con sus obligaciones [15].

Este modelo tiene un enfoque microeconómico, en el sentido de que sólo se utilizan los precios de las acciones de la empresa, sin tener en cuenta de forma explícita el ciclo económico [17].

Este modelo intenta encontrar la relación entre la distancia al incumplimiento y la probabilidad de que se produzca, por lo cual se elabora una tabla que relaciona la probabilidad de incumplimiento con los niveles de distancias de default [18].

Se debe tener presente que la principal desventaja de los modelos estructurales, como lo es el KMV, es su dependencia de los supuestos de distribución es decir, la normalidad, lo cual hace que los modelos no se asemejen a la realidad.

Los modelos de forma reducida presentan varias ventajas, respecto a los modelos estructurales como KMV y Moody’s como: su cómputo, y su mejor ajuste a los datos observados de crédito [19].

### **2.3.2. CREDITMETRICS**

Respecto a la metodología más usada para la medición del VAR es el modelo Creditmetrics. Que es una herramienta propuesta por JP Morgan en 1997 para medir el riesgo de un portafolio debido a los cambios en el valor de la deuda originados por los cambios en la calidad crediticia del deudor. No solo incluye cambios en el valor originados en posibles eventos de default sino además las mejoras o desmejoras en la calidad crediticia del emisor [20].

La idea principal de esta metodología es la estimación de la distribución esperada a plazo de los cambios en el valor de una cartera de bonos o préstamos, en un determinado horizonte [21]. CreditMetrics calcula no solo las pérdidas esperadas sino también el valor en riesgo (la volatilidad del valor).

El incentivo para el desarrollo de modelos internos VAR se dio en 1996, cuando el Acuerdo de Basilea modificó su propuesta riesgo de mercado y permitió que algunos bancos utilicen sus propios modelos internos, en vez del modelo estándar propuesto por los reguladores, para calcular su riesgo [22].

La metodología KMV difiere significativamente de CreditMetrics en que se utiliza las “frecuencias esperadas de impago” como probabilidades de incumplimiento, en lugar de depender de las probabilidades de transición asociadas a algún esquema de calificaciones, sean éstas las internas de cada banco o las producidas por alguna de las agencias calificadoras [23].

Una vez que se tiene una visión general de los modelos de medición de riesgo de crédito se puede concluir que los modelos tradicionales presentan subjetividad y falta de integración a la gestión de riesgo. Además los sistemas tradicionales se quedan en la etapa de calificación, ya que no hay un vínculo teórico identificable con la probabilidad de impago y la severidad de la pérdida. En el mejor de los casos se puede establecer una correspondencia entre calificaciones y probabilidades de quiebra mediante estudios empíricos *ex post*, cuando la información lo permite.

Los modelos modernos por su parte, no siempre la información del mercado representa el valor de las empresas y es necesario usar técnicas estadísticas avanzadas. Sin embargo, la principal ventaja de los modelos modernos es que permiten determinar la probabilidad individual de incumplimiento.

Finalmente, la implantación de modelos para la medición y gestión del riesgo es sin duda un aspecto fundamental que debe ser tomado en cuenta y al cual las instituciones financieras deben dirigir sus esfuerzos, para que éstos sean coherentes con la realidad de cada entidad y de esta manera sirva de apoyo para que las decisiones de riesgo sean más objetivas, más eficientes y más ágiles.

## Capítulo 3

# REDES BAYESIANAS: ANTECEDENTES Y DESARROLLO

### 3.1. INTRODUCCIÓN

En las últimas décadas se han desarrollado numerosas técnicas de análisis y modelización en distintas áreas de estadística y la inteligencia artificial. La Minería de Datos (*Data Mining*) es un área moderna interdisciplinar que engloba aquellas técnicas que operan de manera automática (requieren de la mínima intervención humana) y además son eficientes para trabajar con una gran cantidad de información disponible en las bases de datos de diferentes problemas prácticos, sean estos de diagnóstico, clasificación o predicción.

Tradicionalmente, este tipo de problemas se resolvían de forma manual aplicando técnicas estadísticas clásicas, pero el incremento del volumen de los datos ha motivado el estudio de técnicas de análisis automáticas que usan herramientas más complejas. Por lo tanto, la *Data Mining* identifica tendencias en los datos que van más allá de un análisis simple. Técnicas modernas de Minería de datos (reglas de asociación, árboles de decisión, algoritmos de regresión, redes neuronales, máquinas de vectores soporte,

redes bayesianas, etc.) se utilizan en ámbitos muy diferentes para resolver problemas de asociación, clasificación, segmentación y predicción.

Entre los diferentes algoritmos de *Data Mining*, los modelos gráficos probabilísticos (en particular las *redes bayesianas*) constituyen una metodología potente basada en la probabilidad y la estadística que permite construir modelos de probabilidad conjunta manejables que representan las dependencias relevantes entre un conjunto de variables (cientos de variables en aplicaciones prácticas). Así, los modelos resultantes permiten realizar *inferencia probabilística* de una manera eficiente.

Las redes bayesianas deben su nombre a la obra de Thomas Bayes (1702 -1761), quien consideró la cuestión de cómo hacer *inferencias inductivas* de una población a partir de datos muestrales observados. Si antes el interés se centraba en deducir consecuencias de hipótesis específicas, a Bayes le preocupó el problema inverso establecer afirmaciones a partir de las observaciones de consecuencias de esa hipótesis. El teorema que lleva su nombre permite calcular las probabilidades de las causas tomando en cuenta los “efectos observados”.

Una *red bayesiana*, en su definición más simple es un *grafo acíclico dirigido y direccionado* que permiten la representación de una distribución conjunta de un conjunto de variables aleatorias. Cada vértice del grafo <sup>1</sup> representa una variable aleatoria y las aristas <sup>2</sup> representan, las dependencias directas entre variables. En una *red bayesiana*, las premisas de *independencia condicional* son exploradas para reducir los parámetros necesarios para caracterizar una distribución de probabilidades y para calcular de forma eficiente las probabilidades a posteriori dadas las evidencias.

Por otro lado, la estructura de una *red bayesiana*, puede ser definida manualmente mediante las relaciones entre variables dadas por especialistas o expertos, o puede ser definida a partir de bases de datos utilizando *algoritmos de aprendizaje de estructura* <sup>3</sup>.

---

<sup>1</sup>Es un par  $G = (N/A)$ , donde  $N$  es el conjunto de nodos y  $A$  es un conjunto de arcos definidos sobre los nodos.

<sup>2</sup>También conocidos como arcos, una arista es un par ordenado  $(X,Y)$ . En la representación gráfica, un arco  $(X,Y)$ .

<sup>3</sup>Se puede ampliar la revisión de aprendizaje en redes bayesianas en Neapolitan (2004), Buntine (1996), Heckerman (1995).

Las redes bayesianas aplicadas a problemas de clasificación son llamadas *Clasificadores Bayesianos*. Estos modelos tienen como objetivo describir y distinguir las clases y clasificar elementos que no fueron clasificados. En nuestra investigación se utilizarán los clasificadores bayesianos *Tree Augmented Naive Bayes*.

Dependiendo del tipo de aplicación, el uso práctico de una *red bayesiana* puede ser considerada en diferentes modelos como los sistema experto neuronales, árboles de decisión, en modelos de análisis de datos (regresión lineal), árboles de clasificación etc. No obstante sin tener en cuenta alguna consideración teórica específica, los siguientes aspectos y características de las redes bayesianas hacen que sean preferibles a otros modelos:

1. *Adquisición de conocimiento*: Capacidad para reunir y fusionar el conocimiento de diversas formas en un mismo modelo (datos históricos, empíricos expresados en reglas lógicas, ecuaciones etc.)
2. *Representación del conocimiento*: La representación gráfica de una *red bayesiana* es sencilla, intuitiva y comprensible, lo que facilita tanto la validación del modelo, su evolución y en especial su uso potencial. Típicamente, un tomador de decisiones siempre va a estar mucho más inclinado a confiar en su conocimiento a un modelo que desconoce completamente.
3. *Utilización del conocimiento*: Una *red bayesiana* es versátil, puede utilizar el mismo modelo para evaluar, predecir, diagnosticar, o para optimizar las decisiones.
4. *Calidad de oferta en el software*: En la actualidad existe software para la captura y procesamiento de las *redes bayesianas*. Estas herramientas tienen características que han ido evolucionado y mejorando continuamente: el aprendizaje de probabilidades, el aprendizaje de la estructura de la red, capacidad bayesiana para integrar las variables continuas y variables de decisión, etc.

## 3.2. APLICACIONES E IMPORTANCIA DE LAS REDES BAYESIANAS EN LA TOMA DE DECISIONES

### 3.2.1. APLICACIONES

Las *redes bayesianas*, organizan un caso problema mediante un conjunto de variables y las relaciones de dependencia que se manifiestan entre ellas. Dado este modelo se puede hacer inferencia bayesiana, es decir, estimar las variables o valores desconocidos a partir de variables o valores ya conocidos. Así los diferentes tipos de modelos bayesianos poseen diferentes aplicaciones para diagnóstico, clasificación, y decisión que brinde información importante en cuanto a cómo se relacionan las variables que pueden ser interpretadas como relaciones de causa y efecto.[25]-

Ante las ventajas mencionadas, las aplicaciones de *redes bayesianas* a lo largo de la historia han sido múltiples [24], la primera aplicación se dio durante el periodo 1988-1995, centrándose en problemas de diagnóstico, de manera específica en medicina, un proyecto financiado para diagnosticar enfermedades neuromusculares que posteriormente se convertiría en una herramienta muy útil para los médicos. Las *redes bayesianas* han sido aplicadas en esta área porque ofrecen la posibilidad de integrar fuentes heterogéneas de conocimiento (conocimiento humano y datos estadísticos).

Por otro lado las *redes bayesianas*, también han sido utilizadas en otros campos como la depuración de programas, procesos de producción, todos estos enfocados en la solución de problemas e identificación de probabilidades que disminuyan el riesgo, como es el uso primordial que se le da en la rama financiera. Básicamente el uso que se les da a las *redes bayesianas*, en el estudio y tratamiento de datos consiste en mejorar esquemas de riesgo, con la finalidad de que se convierta en un sistema de ayuda a la toma de diferentes decisiones.

En Marketing, por ejemplo las *redes bayesianas* se han aplicado para realizar el análisis de satisfacción de clientes, imagen de marca, segmentación de clientes y productos, etc. Las características de las *redes bayesianas* que han permitido el desarrollo

en este campo han sido principalmente la gestión de incertidumbre porque, obviamente, todas las acciones de marketing se toman en un contexto de incertidumbre, y lo que se busca principalmente es aumentar la probabilidad de éxito, así como también la capacidad de incorporar datos incompletos durante el aprendizaje, porque los datos utilizados en la minería de datos aplicada a Marketing a menudo proceden de fuentes de información incompletos<sup>4</sup>.

Si bien en el campo de la banca y las finanzas no existe una literatura extensa de aplicaciones, esta técnica de minería de datos es de gran potencial en diferentes ámbitos. En los últimos años se han convertido en una herramienta de gestión de riesgo de suma importancia para el sector financiero, debido al gran aporte en la definición de condiciones probabilísticas de inferencia, con un amplio manejo de variables. En este ámbito han sido utilizadas para representar el conocimiento en métodos de razonamiento, explotación de información, análisis de los deudores del sistema financiero y predicción de ventas; basados en la teoría de la probabilidad, tal que se pueda predecir el valor de las variables no observadas y explicar los resultados arrojados de las variables observadas.

Un ejemplo claro de lo mencionado se da en la medición y mitigación de riesgo operacional [26], en las instituciones financieras a partir de Basilea II, ha dado lugar a la aplicación de esta técnica en este campo ya que la gestión del riesgo operativo, es un tipo de riesgo relativamente nuevo comparado con otros tipos de riesgo mucho más “populares”, como lo son el de crédito o el de mercado. Por este motivo, las bases de datos de pérdidas operativas históricas que poseen las entidades financieras son estadísticamente insuficientes para ser modeladas con un enfoque meramente frecuentistas, razón por la cual ha sido de vital importancia incluir el criterio de los expertos mediante la aplicación de *redes bayesianas* en la modelación de pérdidas operativas, ya que son precisamente los responsables de los procesos internos quienes poseen el conocimiento suficiente para emitir criterios acertados sobre sus potenciales pérdidas.

Finalmente en la última década se ha dado también un desarrollo importante de las

---

<sup>4</sup>En el campo de Marketing Baesens et al (2004) utilizarán las redes para clasificar clientes en función de su potencial de gasto y ofrecer nuevos productos.



redes bayesianas referente a riesgo de crédito, mediante aplicaciones a los “scorings” de crédito. Estos modelos se han propuesto en este campo ya que se conocen en las relaciones probabilísticas entre las variables en el dominio, a diferencia de las redes neuronales. Las técnicas de *redes bayesianas* utilizadas para resolver los problemas de clasificación los datos se denominan *clasificadores bayesianos*, cuyo objetivo es de proporcionar o clasificar individuos que no se han clasificados como, por ejemplo, clasificar un cliente como “bueno” o “malo”, de acuerdo a las variables predictoras [27].

Al respecto se han destacado dos estudios de aplicación de *redes bayesianas* a riesgo de crédito que han sido los pioneros y han dado lugar a estudios posteriores en el mismo ámbito: Baesens et al. (2002) y Chang et al. (2000)[28]. El artículo Baesens describe la aplicación de *clasificadores bayesianos* en modelos de credit scoring. Esta estructura de aprendizaje fue empleada utilizando la simulación de Monte Carlo y la selección de las variables de estos clasificadores se hizo con la Manta de Markov de la variable de respuesta (cliente “bueno” o “malo”). Las principales conclusiones fueron que los clasificadores bayesianos mostraron un buen rendimiento en el scoring de crédito y la selección de variables por Markov dio lugar a modelos mucho más potentes que los tradicionales.

El interes en métodos de credit scoring ha crecido en la medida que los créditos de consumo y tarjetas de crédito han evolucionado en el mercado financiero, por lo que se ha buscado nuevos métodos estadísticos - matemáticos para obtener mejores resultados en los “scorings” de las instituciones financieras. Si bien existen métodos como regresión lineal, análisis discriminante, programación matemática, algoritmos genéticos, en esta investigación se usará la técnica de redes bayesianas ya que el principal objetivo es la construcción de un modelo gráfico que permita determinar las variables más importantes (predictoras) que expliquen a la variable objetivo (clientes buenos y malos)[29].

Además, los modelos gráficos (redes bayesianas), difieren de los otros métodos mencionados ya que también permiten determinar la dependencia entre las variables

predictoras. Así se puede resumir que las redes bayesianas han sido aplicadas a modelos de scoring por las siguientes razones [30]:

- Las *redes bayesianas*, son efectivas para entender las relaciones entre variables predictoras, ya que el número de variables puede ser extenso, dependiendo de la información con la que cuente la institución financiera, y éstas ayudan a eliminar las variables que no influyen en el score de un cliente.
- Segundo, algunos nodos de decisión pueden ser añadidos a una *red bayesiana* en secuencia y optimizar la maximización de utilidad, dependiendo de la política de la institución financiera.
- Tercero, las *redes bayesianas* permiten interacción de conocimiento de expertos.

Así, las *redes bayesianas* permiten crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real de “sentido común” para establecer la probabilidad de las instancias utilizando atributos aparentemente no vinculados. Por ejemplo supongamos que una entidad financiera está preocupada por el posible impago de sus créditos, se pueden utilizar los datos de los créditos anteriores para pronosticar los clientes potenciales que tendrán problemas para pagar sus créditos, a estos clientes de alto riesgo se les puede negar el crédito o se les puede ofrecer otros productos.

En conclusión, el número de aplicaciones de *redes bayesianas* ha ido aumentando año con año y continuarán siendo explotadas en la medida que aparezca nueva tecnología y nuevas herramientas. En la presente investigación se utilizará *redes bayesianas* en la selección de variables en el modelo ya que como veremos posteriormente el número de variables a utilizar en el modelo será grande en función de la información de los solicitantes; los *clasificadores bayesianos* permitirán eliminar algunas variables que no influyan en la puntuación de crédito de los clientes.

### 3.2.2. IMPORTANCIA EN LA TOMA DE DECISIONES [31]

Uno de los mejores usos que se le ha podido dar al manejo de la incertidumbre es la probabilidad y dentro de la inteligencia artificial es notable el uso del método bayesiano como herramienta principal para la definición de hipótesis resultantes del manejo de inferencias dentro del marco del manejo de información probabilística.

La extensión de las *redes bayesianas* es de gran utilidad porque utiliza nodos de decisión y nodos de utilidad, ya que éstos permiten resolver los diferentes problemas de toma de decisiones. La toma de decisiones es tratada como un proceso de optimización en donde se recomienda la mejor alternativa que optimice la utilidad esperada, dadas las observaciones de los factores externos y preferencias que haga la toma de decisiones. En el proceso de construcción de una *red bayesiana* para la toma de decisiones, es importante definir el problema teniendo claras todas las características del caso de estudio, tal que, posteriormente sea posible obtener los datos que serán las variables fundamentales en el desarrollo del problema y se plantee un modelo del cual sea viable obtener la mejor solución que se acople a las necesidades de la solución del problema.

Las *redes bayesianas* representan la incertidumbre con probabilidades, ofreciendo una idea de qué tan fuertemente puede ocurrir un evento, usando la comparación como medio de información sobre posibles consecuencias de una situación. La capacidad del proceso de creación de las redes bayesianas apunta su aplicación a tácticas de toma de decisiones estratégicas. Utilizando la estructura bayesiana para la optimización de problemas, los expertos trabajan como un grupo de toma de decisiones ofreciendo diferentes opiniones sobre la misma situación; esto permite tener diferentes soluciones para que permita identificar variables y las relaciones entre ellas en el modelo y la cuantificación de los gráficos con probabilidades numéricas.

Debido a las limitaciones de probabilidad para la inferencia de una red bayesiana, usualmente no es posible evitar contradicciones entre las diferentes soluciones de los expertos en la toma de decisiones en grupo; ya que se necesita identificar un conjunto de alternativas en el que grupo de expertos tienen opiniones que consideran el problema según el orden de preferencias.

## Capítulo 4

# FUNDAMENTACIÓN TEÓRICA

### 4.1. INTRODUCCIÓN

En los últimos años, la industria bancaria ha cambiado profundamente por los procesos de liberalización, innovación financiera y tecnológica. Esto, ha generado en las entidades financieras una evolución en el ámbito de la modelización de procesos para la medición y gestión del riesgo<sup>1</sup>, al pasar de procesos empíricos a metodologías apoyadas en técnicas estadísticas.

Un buen entendimiento de los conceptos de riesgo, permite a los administradores planear adecuadamente la forma de anticiparse a posibles resultados adversos y sus consecuencias y, de este modo, prepararse para enfrentar la incertidumbre sobre las variables que puedan afectar dichos resultados.

De esta manera, la necesidad de medir el riesgo y promover que las instituciones financieras hagan una correcta evaluación de ellos ha sido un esfuerzo de todos los bancos a nivel mundial. La importancia de tener un modelo de cálculo de probabilidad

---

<sup>1</sup>Una muestra de que la gestión de riesgo ha tomado vital importancia se deja ver en el estudio realizado por Deloitte & Touche (Roisenvit y Zárate, 2006), el cual revela que cuatro de cada cinco instituciones globales de servicios financieros han creado el puesto de director de riesgos.

de incumplimiento confiable con una alta capacidad de discriminación radica en que ésta impacta considerablemente el cálculo de provisiones, afectando directamente el balance y las utilidades que podría llegar a tener la entidad.

En términos generales, el resultado de un crédito otorgado puede manifestarse de dos formas:

1. El cliente cancela el crédito otorgado.
2. El cliente se declara insolvente y no paga el monto del crédito.

La institución bancaria no sufre ninguna pérdida con la primera alternativa, mientras que sufre una pérdida con la segunda. Aunque la entidad no puede saber de antemano el resultado, el análisis de los aspectos apropiados del acreditado resultan ser un buen indicador de cuál podría ser la resolución del crédito. Con ello, el banco puede asignar una probabilidad al evento de que el acreditado liquide el monto dentro del plazo pactado.

En el análisis de crédito tradicional, se hace el supuesto de que las pérdidas individuales de cada crédito son independientes unas de otras. Es decir, el que un acreditado incumpla no implica que otro también lo hará.

Uno de los métodos más utilizado para obtener la distribución de pérdida de una cartera consiste en aproximar ésta por medio de una distribución de probabilidad normal<sup>2</sup>. Esta aproximación requiere que los créditos que componen la cartera tengan una probabilidad de incumplimiento común.

Primero, el uso de la distribución normal supone que la distribución de pérdidas es simétrica, cuando empíricamente ha mostrado ser sesgada hacia las pérdidas. En otras palabras, más de la mitad de las pérdidas que ocurren son menores a la pérdida esperada. Esto por la existencia de créditos con montos superiores al promedio que provocan eventuales pérdidas superiores a la esperada. Segundo, la estimación de correlaciones

---

<sup>2</sup>Utilizada especialmente porque es de sencilla comprensión y produce estimaciones fácilmente interpretables.

entre acreditados requiere de supuestos adicionales, así como de información detallada [32].

Es así, que durante muchos años el área de modelación estuvo restringida a la de los modelos normales clásicos, los cuales eran utilizados para describir la mayoría de los fenómenos aleatorios, aún en los casos en que era bastante razonable suponer distribución no normal para el comportamiento de la variable en estudio. En algunos de estos casos, y aún en la actualidad, se utilizan transformaciones para intentar lograr normalidad. Las transformaciones deben garantizar también simultáneamente los otros supuestos asumidos en un modelo clásico, como constancia de la varianza y linealidad (aditividad) en los parámetros, lo cual ocurre raramente [33].

Así, el uso de datos transformados como base del análisis estadístico es adecuado solo cuando la escala que hace que los supuestos requeridos se cumplan, tiene significado en el área de estudio, dado que las conclusiones se aplican a las poblaciones transformadas (Mead et al, 1993). Nótese que mientras el supuesto de aditividad concierne a la construcción del modelo, los otros dos, normalidad y homogeneidad de varianzas se relacionan con la variación aleatoria y no hay garantía de que ambos requisitos resulten de la misma transformación (Mead et al, 1993; Box y Cox, 1982).

Una propuesta interesante, basada en un concepto innovador de modelación, surgió en 1972, ya que podría suceder que la variable objetivo solo tomara valores en un intervalo o bien que no sea continua, o ni siquiera cuantitativa. De esta manera los errores no pueden ser normales. Esta generalización da lugar a los modelos lineales generalizados (GLM).

Otra posible generalización consiste en reducir restricciones sobre los errores manteniendo sin embargo la propiedad de normalidad, es decir, contemplando errores no independientes o heterogéneos. Esta generalización, que permite dotar de estructura a la variabilidad de los errores del modelo, da lugar a los modelos mixtos. Y, los modelos lineales generalizados junto con los modelos mixtos pueden ser fusionados dando lugar a los modelos lineales generalizados mixtos (GLMM) adaptando las propiedades de ambas propuestas de modelización [34].

#### 4.1.1. MODELO LINEAL MIXTO

Los modelos lineales mixtos propuestos inicialmente por Nelder y Wedderburn en 1972, son una extensión o generalización del modelo lineal estándar en los que se eliminan algunas limitaciones. Entre otras cosas, ofrece una mayor variedad de distribuciones para la variable respuesta, es decir, posibilita el ajuste de un modelo con diferentes distribuciones sin necesidad de realizar modificaciones en los datos. Además, permite establecer una relación entre la variable dependiente y la combinación lineal de factores de predicción mediante una función adecuada del valor esperado de la misma [35].

Concretamente, el modelo lineal mixto permite correlación y heterocedasticidad entre los datos dotándose así de una mayor versatilidad, aunque todavía asume normalidad en la distribución. Robinson (1991) muestra, mediante ejemplos, como el empleo de modelos lineales mixtos es de gran utilidad en campos aparentemente tan diversos como son la geología, la biología o la economía, entre otros. Una interesante aplicación de los modelos lineales mixtos en el ámbito actuarial es la presentada por Frees et al. (1999, 2001). Los autores demuestran, con datos longitudinales, la relación existente entre los modelos de credibilidad y los modelos lineales mixtos [36].

Estos modelos son una herramienta de interés para la modelización, ya que permite el ajuste de funciones lineales sobre observaciones que no verifiquen el supuesto de independencia. El modelo lineal mixto incluye en su formulación efectos fijos, comunes a toda la población, y efectos aleatorios, específicos de cada unidad de muestreo (toma un valor diferente en cada cluster). El modelo de “efectos mixtos” incorpora un nuevo parámetro fijo: la varianza del efecto aleatorio. La varianza del efecto aleatorio recoge la variabilidad entre los diferentes individuos, mientras que la varianza del error recoge la variabilidad dentro de cada individuo no explicada por el modelo. Si la varianza del efecto aleatorio fuera nula, el modelo coincidiría con el modelo de efectos fijos o de regresión lineal.

En notación matricial, los modelos lineales mixtos se puede representar como [37]:

$$y = X\beta + Z\gamma + \varepsilon$$

donde:

$y$  es el vector de observaciones ( $n \times 1$ )

$\beta$  es el vector de efectos fijos ( $p \times 1$ )

$\gamma$  es el vector de efectos aleatorios ( $q \times 1$ )

$\varepsilon$  es el vector de términos de error aleatorio ( $n \times 1$ )

$X$  es la matriz ( $n \times p$ ) asociada a los efectos fijos en relación a las observaciones  $y$  y  $\beta$

$Z$  es la matriz ( $n \times q$ ) asociada a los efectos aleatorios en relación a las observaciones  $y$  y  $\gamma$

Se supone que  $\gamma$  y  $\varepsilon$  son variables aleatorias correlacionadas con media cero y matrices de covarianza  $G$  y  $R$ , respectivamente.

$$E [\gamma] = 0, \quad Var [\gamma] = G$$

$$E [\varepsilon] = 0, \quad Var [\varepsilon] = R$$

$$cov (\varepsilon, \gamma) = 0$$

Así, la esperanza y la varianza ( $V$ ) del vector de observación están dados por:

$$E [y] = X\beta$$

$$Var [y] = V = ZGZ^T + R$$

Es decir, los niveles observados de un efecto aleatorio son una muestra aleatoria de la población de niveles y la esperanza incondicional es la media de  $y$  sobre toda esa población. La matriz  $V$  es un componente importante para trabajar con modelos



mixtos, ya que contiene fuentes de variación aleatoria y define cómo estos modelos difieren de los calculados con Mínimos Cuadrados Ordinarios (MCO).

Si sólo dispone de un modelo de efectos aleatorios la matriz  $G$  es el principal objetivo; por otra parte, para las medidas repetidas o para análisis espacial, la matriz  $R$  es relevante.

Si se asume los efectos aleatorios se distribuyen normalmente, de la siguiente manera:

$$\gamma \sim N(0, G), \quad \varepsilon \sim N(0, R)$$

Entonces, el vector de observación se distribuye normalmente  $y \sim N(X\beta, V)$

Los supuestos clásicos de independencia y homogeneidad de varianzas para los términos aleatorios del modelo lineal general (muestreo ideal) se flexibilizan en el marco del modelo mixto general. La inclusión de efectos aleatorios produce observaciones correlacionadas. Tanto la estructura de correlaciones como la presencia de varianzas heterogéneas pueden ser especificadas a través de la modelación de las matrices de covarianza  $G$  y/o  $R$ . A través de  $G$  y  $R$  es posible modelar correlaciones entre efectos de tratamiento, entre parcelas experimentales ocasionadas por la distribución espacial y/o temporal de las mismas en el campo y/o considerar diferentes precisiones de ensayos cuando se combinan experimentos.

## 4.2. MODELOS LINEALES GENERALIZADOS (GLM)

En los modelos no lineales, existen algunos que se denominan “*linealizables*”, son modelos en los que existe una función de la esperanza que es una función lineal de los parámetros del modelo. Y si, estos modelos linealizables provienen de datos cuya distribución es una familia exponencial (distribución normal, gamma, binomial, Poisson, etc) entonces se tiene un modelo lineal generalizado.

Los modelos lineales generalizados son una extensión de la regresión lineal, ya que permiten que el modelo lineal se relacione con la variable de respuesta a través de una función de enlace y que la magnitud de la varianza de cada medición sea una función de su valor predicho. Es un modelo que vincula las respuestas (variables “dependientes”) con otras variables “independientes” o “explicativas”.

La idea básica de los modelos lineales generalizados consiste en abrir el abanico de posibilidades u opciones para la distribución de la variable respuesta (esto es, “relajar” el supuesto de distribución normal), siempre y cuando pertenezca a una familia más amplia de distribuciones: la familia exponencial, así como permitir que la relación entre el valor esperado (media) de la variable y la combinación lineal de los parámetros (parte sistemática del modelo) no sea siempre la identidad, sino cualquier función monótona [33].

#### 4.2.1. SUPUESTOS BÁSICOS

Los supuestos básicos de un GLM son:

1. Las observaciones son independientemente distribuidas de acuerdo a alguna distribución de la familia exponencial (por ejemplo, normal, binomial o Poisson).
2. El valor de la media de las observaciones está relacionado con la predicción lineal por medio de una función de enlace (por ejemplo, la identidad, probit, logit o log).
3. Los predictores lineales son funciones lineales de los parámetros.

Cuando se utiliza un GLM (con el parámetro de unidad de dispersión), puede dar lugar a los llamados sobredispersión o extra-variación, es decir, la desviación residual es mayor que la esperada, el número de grados de libertad de residuos (Williams, 1982).

### 4.2.2. ESTRUCTURA DEL MODELO

Un modelo lineal generalizado descompone la variabilidad observada en una variable de respuesta en dos componentes (sistemático y aleatorio) vinculados entre sí por una función de enlace.

Por lo tanto se debe considerar tres componentes [71]:

1. Componente aleatoria: se supone que las observaciones  $y_i$  son independientes, con una distribución que sea de la familia exponencial.
2. Componente sistemática: indica la relación entre las variables independientes.
3. Función de enlace: vincula la media (esperanza) de la distribución de las  $y_i$  con la componente sistemática.

#### Distribución de $y$

Normalmente, el vector  $y$  se supone que es independiente de la distribución de densidad de la familia exponencial o similar a la familia exponencial:

$$y_i \sim indep.f_{Y_i}(y_i)$$

$$f_{Y_i}(y_i) = \exp\{[y_i\gamma_i - b(\gamma_i)]/\gamma^2 - c(y_i, \gamma)\} \quad (4.1)$$

donde, por conveniencia, se escribe la distribución de forma canónica. De manera general las distribuciones se pueden escribir en la forma de la ecuación 4.1

#### Función de Enlace

En general, esta función relaciona los parámetros de la distribución de los diferentes predictores. Ello se obtiene mediante el modelado de la transformación de la media,  $\mu_i$ , que es una función de  $\gamma_i$ , como un modelo lineal de los predictores:

$$E[y_i] = \mu_i$$

$$g(\mu_i) = x_i' \beta \quad (4.2)$$

donde  $g(\cdot)$  es una función conocida, llamada *función de enlace* (ya que asocia la media de  $y_i$  y forma lineal de los predictores)  $x_i'$  es la  $i$ -ésima fila de la matriz del modelo, y  $\beta$  es el vector de parámetros en el predictor lineal.

### Predictores

En la práctica, es necesario tomar decisiones en cuanto a que los predictores que se van a incluir en el lado derecho de 4.1 y en qué forma se van a incluir. Un punto clave en el uso de GLM es que muchas de las consideraciones en el modelado son las mismas que para LMMs<sup>3</sup> ya que los lados de la derecha del modelo para la media son los mismos.

### Los modelos lineales

Esta clase generalizada de modelos incluye un modelo lineal como un caso especial. La distribución normal se puede escribir en forma 4.1 mediante la definición de:

$$\gamma_i = \mu_i$$

$$b(\gamma_i) = \frac{1}{2} \mu_i^2$$

$$\gamma^2 = \sigma^2 \quad (4.3)$$

$$c(y_i, \gamma) = \frac{1}{2} \log 2\pi\sigma^2 + \frac{1}{2} y_i^2 / \sigma^2$$

---

<sup>3</sup>Modelos Lineales Mixtos (LMMs) por sus siglas en inglés Linear Mixed Models

Con  $g(\mu_i) = \mu_i$  y  $\mu_i = x_i'\beta$  se genera el modelo lineal.

### 4.3. MODELOS LINEALES GENERALIZADOS MIXTOS (GLMMs)<sup>4</sup>

#### 4.3.1. INTRODUCCIÓN

A menudo, los investigadores se enfrentan con datos no normales por lo que tratan de realizar ciertos ajustes como la transformación de datos para lograr la normalidad y homogeneidad de varianza, utilizando pruebas no paramétricas o depender en la solidez de los clásicos análisis de varianza de no normalidad para los diseños de equilibrio. Además se puede ignorar los efectos aleatorios o tratarlos como factores fijos. Sin embargo, estos procesos pueden fallar (por ejemplo, datos que cuentan con muchos valores cero no pueden hacerse normales por la transformación). Incluso cuando éstos tienen éxito, podrían violar las hipótesis estadísticas (aunque no paramétricas, por ejemplo, de homogeneidad de varianza en todos los grupos) o limitar el alcance de la inferencia (no se puede extrapolar las estimaciones de efectos fijos para los nuevos grupos)[38].

Entonces, para superar estas dificultades se plantea los modelos lineales generalizados mixtos que combinan las propiedades de dos modelos estadísticos, los modelos lineales mixtos (que incorporan efectos aleatorios) y los modelos lineales generalizados (que usan datos no normales mediante el uso de funciones de enlace de la familia exponencial). Los efectos aleatorios no sólo determinan la estructura de correlación entre las observaciones sobre el mismo individuo, también tiene en cuenta la heterogeneidad de los individuos, debido a las características observadas [39].

Los modelos lineales generalizados mixtos dan una mayor flexibilidad en el desarrollo de un modelo apropiado para los datos. Son muy populares cuando hay una respuesta discreta en agrupaciones, por ejemplo, en los estudios longitudinales o en

---

<sup>4</sup>Por sus siglas en inglés

las estructuras de datos jerárquicos. Es normal pensar que los efectos aleatorios tienen una distribución normal. Recientemente, se ha examinado si erróneamente suponiendo una distribución normal para los efectos aleatorios es importante para la estimación de los parámetros de efectos fijos. Si bien se ha demostrado que la mala especificación de la distribución de los efectos aleatorios tiene un efecto menor en el contexto de modelos lineales mixtos, la conclusión generalizada de los modelos mixtos es menos clara. Algunos estudios indican un impacto menor, mientras que otros informan de que el supuesto de normalidad que realmente importa sobre todo cuando la varianza de los efectos aleatorios es relativamente alto. Como no está claro si la hipótesis de la normalidad es realmente satisfecha en la práctica, es importante que los modelos mixtos generalizados estén disponibles a relajar el supuesto de normalidad. La sustitución de la distribución normal con una mezcla de distribuciones gaussianas se especifica en una cuadrícula mediante la cual sólo los pesos de los componentes mixtos se calculan utilizando un enfoque penalizado para asegurar una distribución homogénea de los efectos aleatorios [40].

Los modelos lineales generalizados tienen amplias aplicaciones en el análisis de datos con diferentes respuestas, tales como: datos binarios y los datos ordinales, que también contienen muchas distribuciones continuas. Versiones extendidas de estos modelos se utilizan cuando estas respuestas son dependientes. Lo que causa la variación de la matriz de covarianza de las observaciones [41].

Los modelos lineales generalizados mixtos toman una posición intermedia entre los modelos marginales y los modelos con efectos no lineales mixtos. En los años recientes, han ganado importancia estos modelos.

En general, el GLMM se presenta como una distribución de la familia exponencial. Así, los casos especiales de GLMM como las regresiones logit o probit para datos binarios o regresión de Poisson para datos continuos se deducen como formas específicas de la función enlace [42].

Es decir, si el modelo que se considera está formulado en términos de su esperanza condicional (se conoce la distribución de las observaciones de los efectos aleatorios y la

distribución de los efectos aleatorios), y además el modelo es linealizable (existe una función de enlace que aplicada a la media condicional está linealmente relacionada a los parámetros) se tiene un modelo lineal generalizado mixto [43].

La idea detrás de los modelos lineales generalizados mixtos (GLMMs) es conceptualmente sencilla: incorporar efectos aleatorios en la parte de predicción lineal de un modelo lineal generalizado. Este simple cambio permite dar cabida a la correlación en el contexto de una amplia clase de modelos de datos no distribuidos normalmente. Visto de otra manera, una manera práctica para construir la distribución multivariante para datos que no siguen una distribución normal que puede albergar una cierta flexibilidad en la estructura de la asociación, así como un amplio conjunto de variables de predicción [47].

En el análisis de riesgo de crédito, es importante el comportamiento de una cartera en su conjunto con la finalidad de estudiar el riesgo al que está sujeta la institución prestadora de créditos. Si bien, existe el interés de conocer el número de incumplimientos en un periodo; los incumplimientos ocurren de tal manera que no es posible pronosticar el número exacto de sucesos, ni el momento exacto de su acontecimiento. Esto se debe a que las probabilidades de incumplimiento no son constantes periodo a periodo, porque están sujetas a factores de riesgo como la situación económica del país.

Aunque existen factores que inciden en las probabilidades de incumplimiento, es muy difícil determinar una relación que indique cómo el incumplimiento de un crédito predispone el incumplimiento de otro. Así, cuando la situación económica del país es una recesión probablemente provocará un aumento en las probabilidades de incumplimiento. Por ello, es necesario tomar en cuenta la dependencia existente entre las probabilidades de incumplimiento mediante un factor estocástico sujeto a distintos factores de riesgo.

En este contexto, los modelos lineales generalizados mixtos son una herramienta muy conocida en las estadísticas que parten del hecho de que los eventos de impago por parte de los deudores son eventos dependientes, es decir se rompe con el supuesto

del modelo logit tradicional de independencia entre los individuos analizados. Así, la probabilidad de incumplimiento de una empresa depende de un grupo de factores económicos, dando como resultado que el incumplimiento entre las empresas sea condicionalmente independiente entre ellas<sup>5</sup>.

Estos modelos son útiles cuando el interés del analista se encuentra en los perfiles de respuesta individual en lugar de la media marginal  $E[Y_{ij}]$ . La inclusión de efectos aleatorios en el predictor lineal refleja la idea de que existe una gran heterogeneidad natural a través de los individuos en sus coeficientes de regresión [39]. La capacidad de los GLMMs para captar la dependencia tanto dentro del año, así como entre los años de manera predeterminada los hacen interesantes en lo que se refiere a riesgo de crédito.

Actualmente es ampliamente aceptado que los eventos de incumplimiento (la incapacidad de las contrapartes para cumplir con sus obligaciones financieras) muestren dependencia. Una primera observación para apoyar este punto de vista es que la probabilidad de incumplimiento parece variar con el tiempo de acuerdo a los ciclos económicos. Una segunda razón para comprender la dependencia entre incumplimientos es que los modelos, incluyendo la dependencia de incumplimientos, reproducen mejor comportamientos empíricos de incumplimientos. Las estimaciones de los modelos con incumplimientos independientes simplemente no son consistentes con los datos observados. Captar esta dependencia es, por supuesto, de inmediato interés para instituciones financieras al momento de prestar dinero o sostener inversiones crediticias riesgosas, desde un desproporcional largo número de incumplimientos sobre un horizonte de tiempo ajustado puede tener serias consecuencias. Esto hace que el concepto de la dependencia de incumplimiento sea uno de los elementos más importantes en la modelización del riesgo de la cartera de crédito [48].

En los GLMM las probabilidades de incumplimiento de los deudores son vistas como resultado de dos partes diferentes, efectos fijos y aleatorios. Los efectos fijos son

---

<sup>5</sup>Los incumplimientos entre las empresas son independientes pero están condicionados a los valores que puedan tomar los factores económicos.



variables explicativas o covariables que se cree que tienen un impacto en el incumplimiento del deudor y se registran en una base regular, como la tasa de interés libre de riesgo o ratios clave en el balance de un deudor. Las covariables son cuantitativas (métricas o continuas) como en el caso de una tasa de interés, como cuando las observaciones son indexadas por uno o más factores de clasificación, por ejemplo, la calificación crediticia de acuerdo a una agencia de calificación. Además, las covariables pueden ser compartidas, regionales, relacionadas con la industria o con un deudor específico, como en el caso de los ratios en el balance [48].

Los efectos aleatorios constituyen una parte estocástica de la probabilidad de incumplimiento. Un efecto aleatorio compartido o variable mixta puede intuitivamente ser considerado como un estado anual de la economía. Si esta situación es idéntica en toda la cartera de los deudores el riesgo de impago aumenta cuando el efecto aleatorio es inusualmente alto y esta es la clave para introducir dentro del año la dependencia por incumplimiento. Por lo general se hace la suposición de que condicionada a efectos aleatorios, la probabilidad de incumplimiento (condicional) es independiente. Pero ya que los efectos aleatorios son estocásticos su influencia en la distribución de probabilidad conjunta debe estar integrada, lo que crea dentro del año la dependencia entre las respuestas. Con el fin de obtener la dependencia entre años, se añade la correlación serial de las realizaciones del estado anual de la economía [48].

Los GLMMs ofrecen ventajas ante el modelo de regresión logística [49]:

- Maneja una variable de respuesta multinomial.
- Maneja datos de desequilibrio.
- Da más información sobre el tamaño y la dirección de los efectos.
- Tiene una estructura del modelo explícito, post hoc adaptable para diferentes análisis (en lugar de requerir diferentes diseños experimentales).
- Puede hacer un solo análisis combinado con todos los efectos aleatorios a la vez.

### 4.3.2. ESTRUCTURA DEL MODELO

La forma general del modelo lineal generalizado mixto (en notación matricial) es [50]:

$$y = X\beta + Z\gamma + \varepsilon$$

donde:

$y$  : vector ( $N \times 1$ ) de la variable de respuesta

$X$  : matriz ( $N \times p$ ) de las  $p$  variables de predicción

$\beta$  : vector ( $p \times 1$ ) de los coeficientes de efectos fijos

$Z$  : matriz ( $N \times q$ ) de los  $q$  efectos aleatorios

$\gamma$  : vector ( $q \times 1$ ) de efectos aleatorios  $\gamma_i \sim N(0, \sigma_a^2)$

$\varepsilon$  : vector ( $N \times 1$ ) de errores que no se explica por el modelo (asume distribución normal)

El efecto aleatorio  $\gamma_i$  sigue una distribución normal con media 0 y varianza  $\sigma_a^2$ . Si la varianza es pequeña, la contribución de  $\gamma_i$  es también bastante pequeña y todos los grupos tendrán una curva logística similar. Sin embargo si  $\sigma_a^2$  es relativamente grande, cada grupo tendrá un intercepto diferente.

Para nuestro caso, el uso de un índice adelantado de la economía ecuatoriana tiene una ventaja importante ya que la probabilidad de incumplimiento de un cliente se correlaciona con la de otros clientes del mismo grupo.

Los tres elementos básicos de este modelo son [16]:

- El vector de efectos aleatorios, que contiene el factor sistémico para cada periodo de tiempo.
- Distribución de la familia exponencial para la distribución condicional de la variable de respuesta dado los efectos aleatorios. La variable de respuesta se supone que es condicionalmente independiente dados los efectos aleatorios.
- Función de Enlace que relaciona la media de la variable de respuesta dado los efectos aleatorios, llamado *predictor lineal*.

### Distribución Condicional de $y$

La distribución condicional de  $y$  dado  $u$ , al igual que en el modelo lineal generalizado (4.1 y 4.1) la variable de respuesta  $y$  suele ser, pero no necesariamente supone que contiene elementos condicionalmente independientes, cada uno con una distribución de densidad de la familia exponencial o similar a la familia exponencial:

$$y_i | u \sim indep. f_{Y_i|u} (y_i | u)$$

$$f_{Y_i|u} (y_i | u) = \exp\{[y_i \gamma_i - b(\gamma_i)] / \gamma^2 - c(y_i, \gamma)\} \quad (4.4)$$

La media condicional de  $y_i$  se relaciona con  $\gamma_i$  mediante  $\mu_i = \partial b(\gamma_i) / \partial \gamma_i$ . Se trata de una transformación de la media que se desea modelar como un modelo lineal tanto con efectos fijos como aleatorios.

$$E[y_i | u] = \mu_i$$

$$g(\mu_i) = x_i' \beta + z_i' u \quad (4.5)$$

De la misma manera que en el GLM,  $g(\cdot)$  es una función conocida, llamada la *función de enlace* (ya que une la media condicional de  $Y_i$  y la forma lineal de los predictores),  $x_i'$  es la fila  $i$ -ésima de la matriz del modelo de efectos fijos, y  $\beta$  es el vector de parámetros de efectos fijos. Además, como se añadió  $z_i'$ , que es la  $i$ -ésima fila de la matriz del modelo de efectos aleatorios, y  $u$  el vector de efectos aleatorios; se debe tener en cuenta que se usa  $\mu_i$  para denotar la media condicional de  $y_i$  dado  $u$ , no la media no condicional. De esta manera, se asigna una distribución de efectos aleatorios:

$$u \sim f_U(u) \quad (4.6)$$

Es así como, la distribución condicional de  $y$  dado  $u$  es una extensión del modelo lineal generalizado; es decir,  $\mu_i$  representa la media condicional en lugar de la media marginal o incondicional de lo contrario, todo es lo mismo). En consecuencia, se denota la varianza condicional de  $y_i$  dado  $u$  como  $\gamma^2 v(\mu_i)$  con el fin de mostrar dependencia de la media condicional  $\mu_i$ .

#### 4.3.3. MODELOS DE MIXTURA DE BERNOULLI [43]

Sea  $\Psi = \Psi_1, \dots, \Psi_p$  un vector aleatorio  $p$  – dimensional ( $p < m$ ). El vector aleatorio de  $Y = Y_1, \dots, Y_m$  sigue un proceso de mixtura de Bernoulli con factores  $\Psi$  si existen  $m$  funciones,  $p_i : R^p \rightarrow [0, 1]$ ,  $1 \leq i \leq m$ , tales que  $P(Y_i = 1 \mid \Psi = \psi) = p_i(\psi)$

Se tiene un vector de créditos  $y = (y_1, \dots, y_m)$  en  $\{0, 1\}^m$ . Entonces, la distribución conjunta condicional de los incumplimientos del portafolio es:

$$P(Y = y \mid \Psi = \psi) = \prod_{i=1}^m p_i(\psi)^{y_i} (1 - p_i(\psi))^{1-y_i}$$

En otras palabras  $Y_i \sim B(P_i)$  donde  $B(P_i)$  denota una distribución de Bernoulli con parámetro  $P_i$ . En este caso  $P_i$  es una variable aleatoria tal que  $P_i = p_i(\Psi)$  y las variables  $Y_i \mid \Psi$  son independientes.

Intuitivamente, la distribución con la que se realiza cada pérdida se escoge de forma aleatoria con base en la realización de  $\Psi$ .

La distribución incondicional de los incumplimientos de  $Y$  se la obtiene integrando la distribución conjunta condicional de los incumplimientos del portafolio sobre la distribución del vector de factores. Así, la probabilidad de incumplimiento de un crédito  $i$  está dada por  $\bar{p}_i = P(Y_i = 1) = E(p_i(\Psi))$ .

Dado que el incumplimiento es un evento poco frecuente se analiza la opción de aproximar una variable aleatoria de Bernoulli a partir de variables aleatorias de Poisson dando origen a los modelos de mixtura de Poisson, en este modelo es permitido que un

crédito incumpla más de una vez en un horizonte de tiempo pero que la probabilidad que ocurra dicho evento sea baja; la notación para determinar este evento es  $\tilde{y}_i \in \{0, 1, 2, \dots\}$  y representa el número de incumplimientos que puede presentar un crédito en un horizonte de tiempo fijo.

#### 4.3.4. MODELOS DE MIXTURA DE POISSON

Sea  $\Psi = \Psi_1, \dots, \Psi_p$  un vector aleatorio  $p$  – dimensional ( $p < m$ ). El vector aleatorio de  $\tilde{Y} = \tilde{Y}_1, \dots, \tilde{Y}_m$  sigue un proceso de mixtura de Poisson con factores  $\Psi$  si existen  $m$  funciones,  $\lambda_i : R^p \rightarrow (0, \infty)$ ,  $1 \leq i \leq m$  tales que el vector  $\tilde{Y}$  es un vector de distribuciones independientes de Poisson con parámetros  $\lambda_i(\Psi)$ .

Se define  $Y_i = I_{[\tilde{Y}_i > 0]}$  entonces el vector aleatorio de  $Y = Y_1, \dots, Y_m$  sigue un proceso de mixtura de Bernoulli con factores  $\Psi$  si existen  $m$  funciones,  $p_i : R^p \rightarrow [0, 1]$ ,  $1 \leq i \leq m$ , tales que  $P(Y_i = 1 \mid \Psi = \psi) = p_i(\psi)$  sigue un proceso de mixtura de Bernoulli donde  $p_i(\cdot) = 1 - e^{-\lambda_i(\cdot)}$ . Si  $\lambda_i(\cdot)$  es pequeño entonces  $p_i(\cdot) \cong \lambda_i(\cdot)$ . La probabilidad de incumplimiento  $p_i(\cdot)$ , es aproximadamente la intensidad  $\lambda_i(\cdot)$ .

Sea  $\tilde{M} = \sum_{i=1}^m \tilde{Y}_i$ ,  $\tilde{M}$  es aproximadamente el número de créditos que incumplen. Entonces,

$$P(\tilde{M} = k \mid \Psi = \psi) = \frac{\left( \sum_{i=1}^m \lambda_i(\psi) \right)^k \exp \left( - \sum_{i=1}^m \lambda_i(\psi) \right)}{k!}$$

En este modelo, la forma como se escoge  $\lambda_i(\psi)$  y la distribución de  $\Psi$  son fundamentales para calcular la distribución de  $\tilde{M}$ .

Sin embargo, el objetivo final es calcular la distribución de la pérdida del portafolio agregado. Este cálculo depende de la implementación que se haga del modelo. Específicamente, existen dos implementaciones interesantes del modelo de mixtura de Poisson. La primer implementación sigue la misma línea de los modelos de mixtura de Bernoulli

donde estos se interpretan como GLMM. Los modelos de mixtura de Poisson también pueden verse como un caso particular de este tipo de modelos.

## 4.4. REDES BAYESIANAS

### 4.4.1. INTRODUCCIÓN

Los *modelos gráficos probabilísticos* son modelos especialmente diseñados para manejar la incertidumbre, la cual caracteriza a la mayoría de los problemas de la vida real. Esta característica, conjuntamente con su representación gráfica, la cual resulta intuitiva para el experto, facilita la comprensión de la relación entre las variables del modelo. Esto explica porque expertos los han aplicado en diferentes problemas.

Las *redes bayesianas* [47] son grafos acíclicos dirigidos donde los nodos representan variables y los arcos representan las relaciones condicionales de (in)dependencia entre las variables. Las *redes bayesianas* están basadas en *modelos gráficos probabilísticos* donde cada variable aleatoria sigue una función de probabilidad condicional dado un valor específico de sus padres. En el caso de que las variables predictoras sean continuas se emplean las redes Gaussianas [48] y en el caso de que se tengan variables discretas y continuas se aplican las redes condicionales Gaussianas [49].

Los *clasificadores bayesianos* por su parte son modelos particulares de las redes bayesianas que se amplían en la tarea de *clasificación supervisada*. Han sido ampliamente utilizados en el campo de la minería de datos y el aprendizaje automático obteniendo buenos resultados. Existen diferentes tipos de clasificadores dependiendo de las relaciones de dependencia a contemplar entre las variables predictoras.

En esta sección se describirán a partir del Teorema de Bayes, los clasificadores bayesianos ampliamente utilizados en la literatura. Paradigmas como Naive Bayes, y de forma particular Naive Bayes Aumentado en Árbol (TAN).

#### 4.4.2. TEOREMA DE BAYES

El Teorema de Bayes, dentro de la teoría probabilística, proporciona la distribución de probabilidad condicional de un evento “A” dado otro evento “B” (probabilidad posteriori), en función de la distribución de probabilidad condicional del evento “B” dado “A” y de la distribución de probabilidad marginal del evento “A” (probabilidad simple o a priori).

La *probabilidad a priori* es la probabilidad de una variable en ausencia de evidencia. La *probabilidad a posteriori* es la probabilidad de una variable condicionada a la existencia de una determinada evidencia; la probabilidad a posteriori de  $X$  cuando se dispone de la evidencia  $e$  se calcula como  $P(X | e)$ .

Partiendo de la fórmulas de probabilidad condicional que menciona que dadas dos variables  $X$  e  $Y$ , la probabilidad de que ocurra “A” dado que ocurrió “B” es la probabilidad condicional de “A” dado “B” y se denota como  $P(A|B)$ :

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \quad (4.7)$$

Lo mismo ocurre si se cambia de orden:

$$P(B/A) = \frac{P(B \cap A)}{P(A)} \quad (4.8)$$

y a partir de las fórmulas anteriores se obtiene

$$P(A/B) = \frac{P(A)P(B/A)}{P(B)} \quad (4.9)$$

La expresión anterior es lo que se conoce como *Teorema de Bayes*.

**Teorema 1** : Sean  $A_1, A_2, \dots, A_n$  eventos mutuamente excluyentes tales que, cualquier evento “B” en el espacio muestral pertenece a uno y solo a uno de estos eventos. Entonces la probabilidad de que ocurra cualquier evento  $A_k$  dado que ha ocurrido el evento “B” se calculará por la siguiente fórmula:

$$P(A_k/B) = \frac{P(A_k \cap B)}{P(B)} \quad (4.10)$$

Por lo tanto, sustituyendo la fórmula de probabilidad condicional, se obtiene la fórmula general para el *Teorema de Bayes*:

$$P(A_k/B) = \frac{P(A_k)P(B/A_k)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \dots + P(A_n)P(B/A_n)} \quad (4.11)$$

Donde:

- El numerador es la probabilidad conjunta.
- El denominador es la probabilidad marginal que ocurra el evento “B”.
- Como “A” y “B” son eventos estadísticamente dependientes, el *Teorema de Bayes* se puede representar también utilizando el diagrama de árbol.

## PROBABILIDAD CONJUNTA

- **Teorema 2** Dado un conjunto de variables  $\{X, Y, \dots, Z\}$ , la probabilidad conjunta especifica la probabilidad de cada combinación posible de estados de cada variable  $P(x_i, y_j, \dots, z_k) \forall i, j, k, \dots$ , de manera que se cumple que:

$$\sum_{i,j,\dots,k} P(x_i, y_j, \dots, z_k) = 1 \quad (4.12)$$

### 4.4.3. DEFINICIÓN DE RED BAYESIANA

Las *redes bayesianas o probabilísticas*<sup>6</sup> se fundamentan en la teoría de la probabilidad y combinan la potencia del *Teorema de Bayes* con la representación de los grafos

<sup>6</sup>Conocidas también como Bayesian Networks.



dirigidos; las mismas permiten representar un modelo causal por medio de una representación gráfica de las independencias o dependencias entre las variables que forman parte del dominio de aplicación [48].

Al representar una distribución de probabilidad, las *redes bayesianas* tienen una semántica clara, lo que permite procesarlas para hacer diagnóstico, aprendizaje, explicación, e inferencias. Según la interpretación, pueden representar causalidad y se refieren como redes causales [47], pero no necesariamente tienen que representar relaciones de causalidad, sino de correlación.

**Definición 3** *Una red bayesiana es un grafo acíclico dirigido (DAG) en el cual cada nodo representa una variable de un dominio y cada arco una dependencia probabilística, en el que se especifica la probabilidad condicional de cada variable dados sus padres. Así, una red bayesiana está compuesta de los siguientes elementos (Jensen 2001):*

- *Un conjunto de variables y un conjunto de aristas direccionadas entre las variables.*
- *Cada nodo contiene estados finitos y mutuamente excluyentes.*
- *Las variables y aristas direccionadas representan un grafo acíclico dirigido.*
- *Cada variable  $A$ , con padres  $B_1, B_2, \dots, B_n$ , así como también una tabla de probabilidades condicionales asociada,  $P(A/B_1, B_2, \dots, B_n)$ .*

En otras palabras, las *redes bayesianas* son *grafos acíclicos y direccionados* que permiten la representación conjunta de probabilidades de un conjunto de variables. Cada nodo representa una variable y cada arista representa una relación de dependencia entre variables. En una *red bayesiana*, si existe una arista direccionada entre dos nodos  $A$  y  $B$ , se dice que  $A$  es “padre” de  $B$  y que  $B$  es “hijo” de  $A$ .

La estructura de una red bayesiana se puede determinar de la siguiente manera:

- Se asigna un vértice o nodo a cada variable ( $X_i$ ), y se indica de qué otros vértices es una causa directa; a ese conjunto de vértices “causa del nodo  $X_i$ ” se lo denota como conjunto  $\pi_{x_i}$  y se lo llamará “padres de  $X_i$ ”.

- Se une cada padre con sus hijos con flechas que parten de los padres y llegan a los hijos.
- A cada variable  $X_i$  se le asigna una matriz  $P(x_i/\pi_{x_i})$  que estima la probabilidad condicional de un evento  $X_i = x_i$ .

Cuando ya se haya realizado la estructura de la red y se han especificado todas las tablas de probabilidad condicional, se puede conocer la probabilidad de una determinada variable dependiendo del estado de cualquier combinación del resto de variables de la red; para ello se debe calcular las variables a posteriori de cada variable condicionada a la evidencia; estas probabilidades a posteriori se podrán obtener de forma inmediata a partir de la fórmula de probabilidad conjunta.

Cada nodo dentro del grafo dirigido de la *red bayesiana* representa una variable aleatoria que tiene asociada una función de probabilidad condicional. La estructura de la *red bayesiana* provee información sobre las relaciones de dependencia e independencia condicional existentes entre las variables. Estas relaciones simplifican la representación de la función de probabilidad conjunta como el producto de las funciones de probabilidad condicional de cada variable.

Una *red bayesiana* para un conjunto de variables aleatorias  $X = (X_1, \dots, X_n)$ , es un par  $B = (G, P(\Theta))$ , donde  $G$  es un grafo dirigido acíclico, cuyos nodos se encuentran en correspondencia uno a uno con las variables en  $X$ , y  $P$  es un conjunto de funciones de probabilidad local definidas por un conjunto de parámetros  $\Theta$ . Se denota  $P_{ai}$  y  $p_{ai}$  para indicar respectivamente, a los padres y a las configuraciones de los padres del nodo  $X_i$  en  $G$ . Así, la función de probabilidad conjunta representada por la estructura  $G$  está dada por:

$$p(x) = p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i \mid P_{ai}) \quad (4.13)$$

La figura (4.1) muestra un ejemplo concreto de *red bayesiana* que representan un cierto conocimiento sobre medicina. En este caso los nodos representan enfermedades,

síntomas o factores que causan algunas enfermedades. En este ejemplo la variable a la que apunta el arco es la variable dependiente, en este caso fiebre depende de tifoidea y gripe.

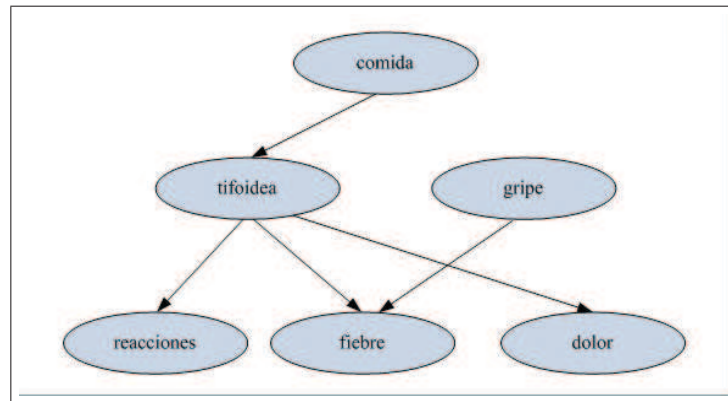


Figura 4.1: Ejemplo de Red Bayesiana

P( C )		P(T/C)		
Ins	Sal		Ins	Sal
0.2	0.8	Si	0.7	0.1
		No	0.3	0.9

Figura 4.2: Probabilidades condicionales ejemplo red bayesiana

Pueden observarse las suposiciones de independencia acertadas por la red, por ejemplo, reacciones es condicionalmente independiente de comida, gripe, fiebre y dolor (nodos no descendientes de reacciones) dado tifoidea (su único nodo padre). Es decir, se observa en la red que:

$$P(R \mid C, T, G, F, D) = P(R \mid T) \quad (4.14)$$

Donde R es reacciones, C es comida, T es tifoidea, G es gripe, F es fiebre y D es dolor.

Por otro lado en la figura 4.2 se muestra la probabilidad asociada a algunos de

los nodos de la red. La tabla  $P(C)$  guarda las probabilidades a priori de la variable comida. La tabla  $P(T/C)$ , la probabilidad de tifoidea dado la variable comida.

#### 4.5. TIPOS DE REDES BAYESIANAS

Existen varios tipos de *redes bayesianas* dependiendo del tipo de variables que contenga el grado (discretas, continuas, o ambas) y del tipo de distribución que se considere para cada variable. Si las variables del problema son todas discretas, el modelo asociado es una *red bayesiana* discreta o red bayesiana multinomial. Si las variables del problema siguen una distribución normal, la red es una red bayesiana normal o red bayesiana Gaussiana y aunque se trabaja en la definición de una red bayesiana para otro tipo de variables aleatorias continuas, todavía no se ha determinado el mecanismo de inferencia cuando la red está formada por otro tipo de variables continuas no Gaussianas. La inclusión de ambos tipos de variables en la red darán lugar a las redes bayesianas mixtas. Además de las anteriores redes las cuales son dependientes del tipo de variables, si el razonamiento probabilístico está condicionado por el tiempo, existirá un cuarto tipo de red denominada red bayesiana dinámica.

Antes de continuar con la explicación más detallada de cada uno de los distintos tipos de redes bayesianas, mencionadas anteriormente, es necesario hacer un inciso sobre los modelos gráficos probabilísticos no dirigidos, denominadas también redes de Markov, como un tipo especial de modelo gráfico que no puede ser considerado como una red bayesiana, pero que posee ciertos aspectos comunes.

En las redes de Markov, la información cualitativa del problema es representada por un grafo no dirigido, donde las relaciones de dependencia entre las variables del problema son relaciones de asociación o correlación, sin establecerse ninguna variable como causa o como efecto, de forma que la información de la que se dispone indica que un conjunto de variables presenta distintos niveles de asociación o correlación.

Posteriormente a la construcción del grafo no dirigido, se buscará la distribución de probabilidad conjunta asociada a las variables del problema como una factorización de

funciones. Los usos más habituales de la redes de Markov están orientados en campos como la física, la robótica y en análisis de imágenes y de textos.

La principal deficiencia de los grafos no dirigidos es su incapacidad para representar relaciones de independencia no transitivas; en estos modelos, dos variables independientes estarán conectadas en el grafo siempre que exista alguna otra variable que dependa de ambas. Por tanto, numerosos modelos de dependencia útiles desde un punto de vista práctico no pueden ser representados por grafos no dirigidos, y es así, donde aparece la necesidad de representar las relaciones causales a través de los grafos dirigidos y, más concretamente, con el uso de las redes bayesianas.

### Redes Bayesianas Discretas

Estas redes se caracterizan porque todas las variables que aparecen en el modelo son discretas, de manera que cada variable sólo puede tomar un conjunto determinado de valores. Cuando las variables del problema son binarias, respondiendo a los procesos de Bernoulli, la red se denomina como red *bayesiana multinomial*.

La figura (4.1), presenta un ejemplo de red bayesiana discreta, en el que además de presentar el diagrama de relaciones se muestran las distribuciones de probabilidad condicionada asociadas a los valores de las variables, donde cada variable tiene determinado sus valores. Los valores posibles de cada variable serán la apreciación de que se produzcan o no, tras su observación, siendo las respuestas sí o no.

### Redes Bayesianas Gaussianas

En las redes bayesianas gaussianas, las variables aleatorias del problema son normales, siendo la distribución conjunta de las variables del problema  $X = \{X_1, \dots, X_n\}$  una normal multivariante  $N(\mu, \Sigma)$ , dada por la función de densidad:

$$f(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp \left\{ -1/2(x - \mu) \Sigma^{-1}(x - \mu) \right\}, \quad (4.15)$$

Donde:

$v$  : vector de medias de dimensión  $n$ .

$\Sigma$  : matriz de covarianzas de  $n \times n$ .

$|\Sigma|$  : determinante de la matriz de covarianzas.

Partiendo del teorema de la factorización de la probabilidad y de la densidad de la distribución normal de las variables se obtiene que:

$$f(x_i) = (pa(X_i)) \sim N(\mu_i + \sum_{j=1}^{i-1} \beta_{ij}(x_j - u_j), v_i) \quad (4.16)$$

Donde:

$\beta_{ij}$  : Es el coeficiente de regresión de  $X_j$  en la regresión de  $X_i$  sobre sus padres.

$v_i$  : es la varianza condicionada de  $X_i$  dados sus padres.

$\beta_{ij}$  : Puede ser 0, si y solo si no hay una arista dirigida de  $X_j$  a  $X_i$ .

### Redes Bayesianas Mixtas

Este tipo de redes se caracterizan por utilizar tanto variables discretas como variables continuas en el modelo gráfico probabilístico dirigido.

Al combinar los dos tipos de variables y poderlas especificar en el modelo, las variables continuas han de ser Gaussianas y las variables discretas deberán preceder a las continuas en el grado.

En estas redes el conjunto de nodos  $V = \{X_1, \dots, X_n\}$ , formada por ambos tipos de variables, se divide por un lado en discretas ( $\Delta$ ) y por el otro en continuas ( $\tau$ ), siendo  $V = \Delta \cup \tau$ . Así, el conjunto de todas las variables aleatorias se denotan como:

$$X = (x)_{\alpha \in V} = (i, \zeta) = \left( (i_\delta)_{\delta \in \Delta}, (\zeta_\gamma)_{\gamma \in \tau} \right) \quad (4.17)$$

Donde:

$i$  : variables discretas.

$\zeta$  : variables continuas.

## 4.6. APRENDIZAJE DE REDES BAYESIANAS

Según Pearl [47] se distinguen dos tipos de aprendizaje bayesiano, cuando se trabajan con redes bayesianas: el *aprendizaje estructural* y *aprendizaje paramétrico*.

### 4.6.1. APRENDIZAJE PARAMÉTRICO

A partir de la estructura de la red, se obtiene las probabilidades a priori de los nodos raíz y las probabilidades condicionales de las demás variables requeridas a través del uso de bases de datos.

El aprendizaje de los parámetros es simple cuando todas las variables son completamente observables en el conjunto de entrenamiento. El método más común es el llamado estimador de máxima verosimilitud, que consiste sencillamente en estimar las probabilidades deseadas a partir de la frecuencia de los valores de los datos de entrenamiento, de forma análoga a como se hace en *Naïve Bayes*.

### 4.6.2. APRENDIZAJE ESTRUCTURAL

Consiste en obtener la estructura de la red, obteniendo las relaciones de dependencias e independencias entre variables existentes. Los algoritmos que aprenden la estructura de la red se encuentran:

- Algoritmos que se basan en un procedimiento que busca la mejor estructura en el espacio de posibles soluciones, midiendo la calidad de cada red candidata mediante funciones de evaluación. Estos son algoritmos que se caracterizan por el tipo de función y por el procedimiento de búsqueda.
- Algoritmos basados en detección de independencias, que toman como entrada el conjunto de relaciones de independencia condicional y generan la red que mejor representan estas relaciones.
- Algoritmos híbridos que se basan en la combinación de ambas metodologías.

El aprendizaje estructural conlleva explorar un espacio de grafos. Esta tarea es muy compleja. En la medida en que se incrementa el número de variables (nodos), el número de posibles grafos a construir con ellas se dispara, por eso en muchas ocasiones se restringe el espacio de búsqueda a grafos con características concretas.

Hay dos aproximaciones básicas al aprendizaje de redes sin restricciones. La primera de ellas reúne métodos que exploran las relaciones de dependencia existentes entre pares, tripletas u otros subconjuntos de variables para elegir la forma en que deben conectarlas. El estudio de esas relaciones requiere establecer un criterio cuantitativo para medir la dependencia entre variables, y es dicho criterio el que guía la construcción de la red. Un ejemplo de algoritmo que se engloba en esta familia de técnicas es el de construcción de TAN (Clasificador Naive Bayes Aumentado en Árbol).

## 4.7. CLASIFICADORES BAYESIANOS

Las redes bayesianas pueden ser aplicadas de forma clara y directa en problemas de clasificación además están diseñadas para encontrar las relaciones de dependencia e independencia entre todas las variables que conforman el dominio de estudio. Esto permite realizar predicciones sobre el comportamiento de cualquiera de las variables desconocidas a partir de los valores de las otras variables conocidas. Permitiendo que cualquier variable de la base de datos pueda comportarse como incógnita o como evidencia según sea el caso.

Siendo la tarea de clasificación como un caso particular de la tarea de predicción, mencionada anteriormente, de las *redes bayesianas*. Un clasificador puede verse como un caso especial de una *red bayesiana* en el cual una función asigna un valor de un atributo, llamado clase, a instancias o conjuntos de características especificados por atributos, que pueden ser tanto continuos como discretos. En el caso de que la base de conocimiento sea una red bayesiana, la función de clasificación estará definida a partir de las probabilidades condicionadas. Siendo la estructura de esta red dependiente del tipo de clasificador que se emplee, como se verá posteriormente.



Los clasificadores bayesianos tienen un uso extendido debido a que presentan ciertas ventajas:

- Por lo general, son fáciles de construir y de entender.
- El proceso de inducción suele ser considerablemente rápido, requiriendo solamente uno o varios pasos.
- Son muy robustos cuando existen atributos irrelevantes.
- Toman evidencias de muchos atributos para realizar la predicción.

A continuación se explican los clasificadores bayesianos que usaremos en nuestra investigación para la selección de variables.

#### 4.7.1. NAIVE BAYES

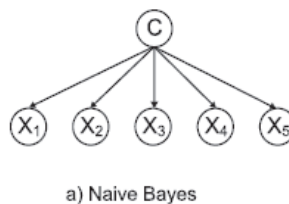


Figura 4.3: Estructura Naive Bayes

El clasificador Naive Bayes [Minsky, 1961] es el modelo de clasificación construido bajo el supuesto de que todas las variables predictoras son condicionalmente independientes dada la variable clase  $C$ . Sin embargo en el dominio de problemas reales esta asunción no se cumple en numerosas ocasiones. A pesar de ello el paradigma Naive Bayes en algunas ocasiones ha mostrado un buen funcionamiento en problemas de este tipo [Domingos y Pazzani, 1997, Hand y Yu, 2001].

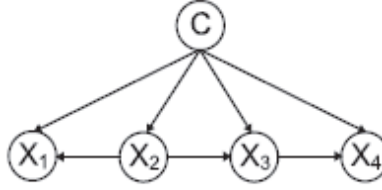


Figura 4.4: Estructura de Tree Augmented Naive Bayes

El clasificador *Naive Bayes* [Minsky, 1961] aplica el Teorema de Bayes para predecir la instancia  $x$ , la clase  $c \in C$  con la máxima probabilidad a *posteriori*:

$$p(c|x)\alpha p(c, x) = p(c) \prod_{j=1}^n p(x_j|c) \quad (4.18)$$

donde  $p(x_i|c)$  representa la probabilidad condicional de  $X_i = x_i$  dado que  $C = c$ . Como resultado el clasificador naive Bayes predice en base a:

$$c^* = \arg . \text{máx} . p(c)p(x_i|c) \quad (4.19)$$

En este paradigma de clasificación todas las variables predictoras están incluidas en el modelo. Los clasificadores bayesianos se caracterizan por representar la distribución de probabilidad conjunta  $p(x_1, \dots, x_n, c)$  entre todas las variables. La estructura completa de los clasificadores bayesianos puede ser definida por la factorización  $p(c, x) = p(c)p(c|x)$ .

#### 4.7.2. TREE ARGUMENTED NAIVE BAYES (TAN)

Naive Bayes aumentado a árbol (TAN) [Friedman y col., 1997] [51] es otro tipo de clasificador Bayesiano que permite dependencias entre las variables predictoras. La principal restricción entre las dependencias de las variables es que cada variable predictora únicamente puede tener un máximo de dos padres: la variable clase  $C$  y alguna otra variable predictora  $X_1, \dots, X_n$ . Para construir una estructura del clasificador naive Bayes aumentado a árbol se necesita previamente aprender las dependencias entre las

diferentes variables predictoras  $X_1, \dots, X_n$ . En [51] se presenta un algoritmo en el cual se tiene una cantidad de información mutua condicionada a la variable clase, en lugar de la cantidad de información mutua en la que se basa el algoritmo [Chow y Liu, 1968]. La cantidad de información mutua entre las variables  $(X_i, X_j)$ , condicionada a la variable clase se define como:

$$I(X_i, X_j \setminus C) = \sum_c \sum_{x_i} \sum_{x_j} p(x_i, x_j, c) \log \frac{p(x_i, x_j \setminus c)}{p(x_i \setminus c)p(x_j \setminus c)} \quad (4.20)$$

Como puede observarse el algoritmo para la construcción del modelo TAN requiere de varios pasos. En primer lugar se calculan las cantidades de información mutua para cada par de variables condicionadas a la variable clase  $I(X_i, X_j \setminus C)$ . Una vez calculados estos valores, la estructura de la red bayesiana del clasificador TAN, se construye en dos fases. La primera se inicia con la construcción de un grafo no dirigido con  $n$  nodos, uno por cada variable predictora  $X_1, \dots, X_n$  en el cual el peso de cada arista viene dado por la cantidad de información mutua entre las dos variables unidas por la arista condicionada a la variable clase  $C$ . El algoritmo asigna por tanto un peso  $I(X_i, X_j \setminus C)$  a las aristas que las conectan  $X_i$  y  $X_j$ . El árbol es construido con las aristas con mayor información mutua condicionada.

Este proceso es repetido anexando la estructura de la red con las aristas con mayor información mutua a no ser que se formen ciclos; caso contrario las aristas son descartadas y se procede a analizar la siguiente arista con mayor información.

El procedimiento termina cuando  $n - 1$  aristas han sido seleccionadas. Finalmente el grafo no dirigido se convierte en un dirigido, escogiendo una variable predictora de forma aleatoria como padre.

En la segunda fase la estructura del clasificador es aumentada a una estructura de clasificador Naive Bayes, agregando el nodo de la variable clase  $C$ , y conectando con un arco el nodo clase  $C$  a cada variable predictora  $X_i (i = 1, \dots, n)$ . La figura muestra un ejemplo de una estructura de un clasificador aumentado en árbol.

### 4.7.3. VALIDACIÓN DE CLASIFICADORES

Existen diversos criterios para la evaluación de los algoritmos de clasificación. Su elección dependerá del dominio del problema. Por ejemplo, *el porcentaje de casos bien clasificados*, mide la bondad (precisión) del clasificador. La bondad de un clasificador es una estimación de la probabilidad de la clasificación correcta de una instancia elegida al azar [Kohavi, 1995b].

En clasificación es aconsejable inducir el modelo a partir de un conjunto de datos llamado conjunto de prueba, el cual se aplica en la fase de clasificación. Con ello se evitan resultados denominados “optimistas”. Es importante que para estimar la precisión de un clasificador se utilice un método con poca varianza.

#### Matriz de Confusión

La matriz de confusión detalla el resultado de la clasificación. En la diagonal principal se reportan los casos correctamente clasificados y en la opuesta se detallan los errores de la predicción. Las columnas representan las clases presentes en los datos y las filas las clases en las que son predichas las instancias. En la tarea de clasificación con dos valores, dado un clasificador y una instancia se producen cuatro valores de salida como son: *verdadero positivo*, si la instancia es clasificada correctamente y su clase pertenece a la positiva; *verdadero negativo* se genera cuando la instancia es correctamente clasificada con la clase de valor negativo; *falso positivo* cuando la instancia es de la clase negativa y es clasificada como clase positiva y por último falso negativo se presenta cuando el clasificador clasifica erróneamente una instancia de la clase positiva como un caso de clase negativa. En base a estas cuatro salidas se puede construir una tabla de contingencia, o matriz de confusión, representando al conjunto de datos de prueba. A partir de la matriz de confusión se pueden extraer algunas medidas para comprender la distribución y naturaleza de los errores cometidos por el clasificador.

Matriz de Confusión		
Subgrupo	Verdadero	Falso
Verdadero	Verdaderos Positivos	falsos positivos
Falso	falsos negativos	verdaderos negativos

Matriz de Confusión

*Sensibilidad de un clasificador:* representa la fracción de verdaderos positivos y se calcula:

$$sensibilidad = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos negativos}} \quad (4.21)$$

*Especificidad de un clasificador:* representa la proporción de verdaderos negativos y se calcula:

$$especificidad = \frac{\text{verdaderos negativos}}{\text{verdaderos negativos} + \text{falsos positivos}} \quad (4.22)$$

*Valor predictivo de la clase positiva:* representa la precisión del clasificador en términos del porcentaje de casos positivos correctamente clasificados, se calcula de la siguiente:

$$valor\ predictivo = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos positivos}} \quad (4.23)$$

*Valor predictivo de la clase negativa:* se calcula de la siguiente forma:

$$valor\ predictivo = \frac{\text{verdaderos negativos}}{\text{verdaderos negativos} + \text{falsos negativos}} \quad (4.24)$$

## MÉTODOS DE VALIDACIÓN DE CLASIFICADORES

Existen diferentes métodos de validación como son el método H (Holdout), remuestreo (random subsampling) y validación cruzada (cross-validation), entre otros. A continuación se explican brevemente los métodos de validación mencionados.

**Holdout** El método holdout (método H) divide aleatoriamente el conjunto de datos en dos subconjuntos. Uno de ellos con  $2/3$  partes para el grupo de entrenamiento, el cual es utilizado para inducir el clasificador, y el segundo grupo con el  $1/3$  de datos restante lo forma el grupo de prueba.

**Remuestreo** El método de remuestreo (random subsampling) es una variante del método H, y consiste en aplicar el método H múltiples veces sobre diferentes subconjuntos del grupo de entrenamiento y grupo de prueba.

**Validación Cruzada** En el método de k-rodajas validación-cruzada (k-fold cross-validation) [Stone, 1974] [52], la base de datos es particionada aleatoriamente en k subconjuntos, aproximadamente del mismo tamaño, donde  $k-1$  subconjuntos constituyen el conjunto de entrenamiento y el restante el conjunto de prueba. El clasificador es inducido a partir del conjunto de entrenamiento. Este proceso se repite k veces y cada ocasión, el clasificador prueba con un conjunto de prueba diferente. La estimación de la precisión del clasificador se calcula promediando la exactitud de cada uno de los k procesos de clasificación y la desviación estándar es calculada de la misma manera. La desviación estándar deberá ser aproximadamente la misma, independiente del número de subconjuntos [Stone, 1974] [52] propuso un método para mejorar el resultado de éxito del clasificador obtenido con k-rodajas validación - cruzada llamado validación cruzada estratificada. Los subconjuntos son estratificados de tal forma que contienen la misma proporción de casos respecto a las clases que el conjunto de datos.

Existe un caso especial del método de k-rodajas validación-cruzada llamado dejar-uno-fuera (*leave-one-out*). En este método de validación la base de datos es particionada k veces, siendo k el número de casos originales n. Se forma un conjunto de entrenamiento con los n-1 casos dejando uno caso fuera, el cual será utilizado en la clasificación. [Efron, 1983] [54] aclara que este método es el menos sesgado pero presenta una varianza alta, la cual en algunos casos presenta una estimación de la precisión

del clasificador no deseada. Este tipo de validación, generalmente es aplicado a bases de datos con relativamente pocos casos.

**Bootstraping** El método bootstraping introducido por [Efron y Tibshirani, 1993] [57] se utiliza para la validación de clasificadores y está basado en la selección de una muestra de datos con reemplazamiento. El muestreo con reemplazamiento consiste en extraer elementos de una población de forma que, tras cada extracción, el elemento extraído se vuelve a introducir y puede volver a ser seleccionado.

El *método bootstraping* introducido por [Efron y Tibshirani, 1993] se utiliza para la validación de clasificadores y está basado en la selección de una muestra de datos con reemplazamiento. El muestreo con reemplazamiento consiste en extraer elementos de una población de forma que, tras cada extracción, el elemento extraído se vuelve a introducir y puede volver a ser seleccionado.

**Área bajo la curva ROC** La curva ROC permite la visualización del comportamiento del clasificador a través de un sistema de coordenadas en el que se representa la sensibilidad en el eje y versus 1-especificidad en el eje x. El gráfico se construye a partir de distintos valores de corte en la probabilidad de la clase del clasificador, los cuales son representados por los diferentes puntos de la curva (1-especificidad, sensibilidad). La curva ROC se construye a partir de variar el valor del umbral  $\tau \in [0; 1]$  cuyo valor sirve como punto de corte para la clasificación. Normalmente se escoge un valor del umbral que minimice la función de coste del error.

El cálculo del área bajo la curva AUC (area under the ROC curve) es un índice ampliamente utilizado [Hanley y McNeil, 1982] [53]. Dicho índice resume el comportamiento y precisión del clasificador. Existen diferentes métodos para calcular el área bajo la curva ROC como son: el método no - paramétrico, paramétrico y por reemplazamiento. El método de la regla trapezoidal propuesto por [Bamber, 1975] es un

método no - paramétrico, que suele utilizarse para el cálculo del área bajo la curva ROC y es equivalente a la prueba estadística de Mann-Whitney.

El valor del área bajo la curva ROC está entre 0.5 (asociado a la diagonal del cuadrado de la gráfica) y 1 (el cual corresponde al punto superior izquierdo) este último asociado a un buen comportamiento del clasificador. La precisión del clasificador aumenta a medida que la curva se desplaza desde la diagonal hacia el vértice superior izquierdo.

En los últimos años se ha incrementado el uso del análisis del área bajo la curva ROC en la comunidad de minería de datos, especialmente en problemas con bases de datos no balanceadas y funciones de error no simétricas, generando el área de investigación en el aprendizaje supervisado denominado sensible al coste.

#### 4.7.4. SELECCIÓN DE VARIABLES

En el dominio financiero es común tener un volumen importante de información, el cual es manejado desde el campo de la minería de datos y el aprendizaje automático por medio de la selección de un conjunto de variables denominado FSS (*feature subset selection*). [Kohavi y John, 1997]. El problema de la selección de un subconjunto de variables, puede plantearse en términos de analizar las variables del dominio del problema y determinar si aportan información que disminuya la incertidumbre sobre las demás variables, ya que no necesariamente todas las variables implicadas son *relevantes*. Por medio del proceso de la selección de variables se reduce el número de variables del dominio del problema al descartar las variables con información *irrelevante o redundante*.

Se considera una variable “*irrelevante*” cuando el conocimiento del valor de la misma no aporta información a la variable clase (objetivo), mientras que se considera una variable “*redundante*” cuando su valor puede ser calculado o determinado a partir de otras variables predictoras. Estos dos tipos de variables pueden disminuir la precisión del clasificador.



Los modelos de clasificación supervisada pueden dividirse en dos categorías dependiendo de la técnica de selección de variables que se aplique: enfoque directo *wrapper* y enfoque indirecto o *filter*.

### Enfoque Directo o Wrapper

El proceso de la selección de variables no debe restringirse únicamente al aprendizaje de los datos sino que también se deben considerar las características del clasificador. En el enfoque wrapper, el algoritmo de clasificación es usado para evaluar la mejora en la precisión del modelo inducido con el subconjunto de variables seleccionadas.

En la aproximación directa o wrapper [Kohavi y John, 1997] cada subconjunto de variables es determinado por un algoritmo de búsqueda, el cual está implícito en la construcción del clasificador. El algoritmo de búsqueda es guiado por una función de mérito o porcentaje de casos bien clasificados obtenida en base a la validación cruzada de  $k$  particiones del conjunto de entrenamiento.

Como se mencionó anteriormente, el objetivo general del proceso de la selección de un conjunto de variables para inducir el modelo es mejorar la precisión del clasificador. Funciones como logaritmo de la verosimilitud (log-likelihood) de los datos y las medidas basadas en la información (como la entropía condicional a la clase o información mutua) han sido ampliamente utilizadas para selección de variables. Estas medidas son conocidas como medidas *filter* y el proceso de la selección de un conjunto de variable basadas en estas medidas es conocido como aproximación indirecta o *filter*. La aproximación indirecta o *filter* [Blum y Langley, 1997, Lewis, 1998] tiene en cuenta las características de los datos y aplica medidas filter para determinar las relaciones existentes entre pares de variables. Inicialmente se lleva a cabo la selección de un subconjunto de variables para después inducir el clasificador.

La información mutua o *mutual information* [Shannon, 1948] es una de las medidas filter más ampliamente utilizadas. En base a la teoría de la información, calcula la relación que existe entre una variable y otra por medio de la verosimilitud de los

datos. Esta medida indica la cantidad de incertidumbre que el conocimiento de una variable predictora es capaz de despejar con respecto al estado en el que se encuentre una segunda variable.

El ranking de las variables es otra medida *filter* y es una de las métricas para establecer una escala de importancia (ranking de las variables). Dichos ranking son realizados de forma univariada, es decir, tan solo se tiene en cuenta la relación existente entre la variable predictora que está siendo analizada y la variable clase. Siempre las primeras variables de la escala han de ser las más importantes dentro del conjunto de variables a seleccionar.

## 4.8. ALGORITMO CHAID

### 4.8.1. INTRODUCCIÓN

Existen varios algoritmos para construir árboles de decisión, las diferencias principales entre los distintos algoritmos radican en las estrategias de poda o terminación y en la regla adoptada para particionar nodos. El algoritmo a utilizarse para la consecución del fin mencionado, será el algoritmo *CHAID*, las siglas *CHAID* corresponden al término inglés Chi- squared Automatic Interaction Detector (detector automático de interacciones mediante chi-cuadrado). Es una técnica estadística desarrollada por Kass (1980) muy eficaz para segmentar o generar árboles.

El método *CHAID*, que utiliza la significación de una prueba estadística como criterio, evalúa todos los valores de una variable predictora potencial. Funde los valores considerados estadísticamente homogéneos respecto a la variable criterio y conserva inalterados todos los valores heterogéneos

A continuación, el algoritmo selecciona la mejor variable predictora para formar la primera rama del árbol de decisión, de forma que cada nodo esté compuesto por un grupo de valores homogéneos de la variable seleccionada. Este proceso se repite hasta que el árbol se ha desarrollado por completo. La prueba estadística utilizada depende

del nivel de medida de la variable criterio. No se trata de un método binario, es decir, puede generar más de dos categorías en cualquier nivel del árbol. Por lo tanto tiende a crear un árbol más ancho que otros métodos. Así también, aprovecha los valores perdidos, tratándolos como una categoría válida individual.

El algoritmo *CHAID* es la herramienta de clasificación utilizada en el estudio de riesgo de crédito por algunas entidades bancarias, *CHAID* se diseñó para identificar las interacciones a incluir en modelos de regresión. Maneja con facilidad las interacciones que tantas dificultades plantean a otras técnicas de modelación. Las interacciones son combinaciones de variables independientes que influyen en el resultado. Además, se la considera como una técnica de explotación de datos (data mining). La explotación de datos consiste en analizar y estudiar grandes masas de datos con el objetivo de descubrir patrones y pautas no triviales desde el punto de vista del aprovechamiento comercial.

#### 4.8.2. METODOLOGÍA CHAID [55]

Para el desarrollo de análisis mediante la metodología *CHAID*, se describen las siguientes tareas:

**Preparación de las Variables:** Esta tarea consiste en seleccionar las variables que definirá el nodo principal del árbol, el cual será de interés para el estudio y elegir un conjunto de posibles pronosticadoras, tanto variables en la medida de lo posible con pocas categorías y variables continuas en la medida de lo posible convertidas en discretas, que permitan realizar una descripción y pronóstico óptimo de la primera variable, es decir del nodo principal.

**Agrupación de las Categorías:** Las variables independientes en el caso de que éstas tengan un perfil similar de la variable inicial o del nodo principal.

**Primera segmentación:** Consiste en la selección de la variable que mejor prediga la variable inicial.

**Segunda segmentación:** Para cada segmento formado en el paso anterior, se busca entre las variables cuyos valores han sido previamente agrupados de la misma forma, por la que tiene mayor poder pronosticador.

**Sucesivas Segmentaciones:** De forma similar se procede según el paso anterior en cada grupo formando la segmentación previa.

#### 4.8.3. DESCRIPCIÓN DE LA TÉCNICA CHAID[56]

*CHAID* funciona con todos los tipos de variables continuas y categóricas. Sin embargo, las variables predictoras continuas se categorizan automáticamente para el análisis. Una de las ventajas de *CHAID*, es que la relación entre la variable dependientes y las variables predictoras se visualiza mediante la imagen de un diagrama de árbol, conocido como árbol de decisión. El árbol de clasificación tiene como base un nodo inicial formado por todos los datos. Después se tiene un primer criterio que divide a este nodo en dos o más grupos llamados nodos hijos, figura 4.5.

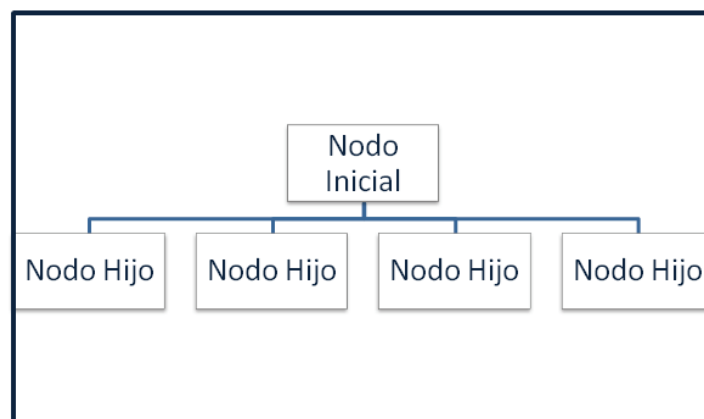


Figura 4.5: Criterio Inicial - CHAID

Para cada nodo hijo se busca otro criterio para dividir los datos nuevamente, por lo cual, los nodos hijos se vuelven nodos padres y de éstos salen nuevos nodos hijos creando así la estructura de un árbol, ver figura 4.6.

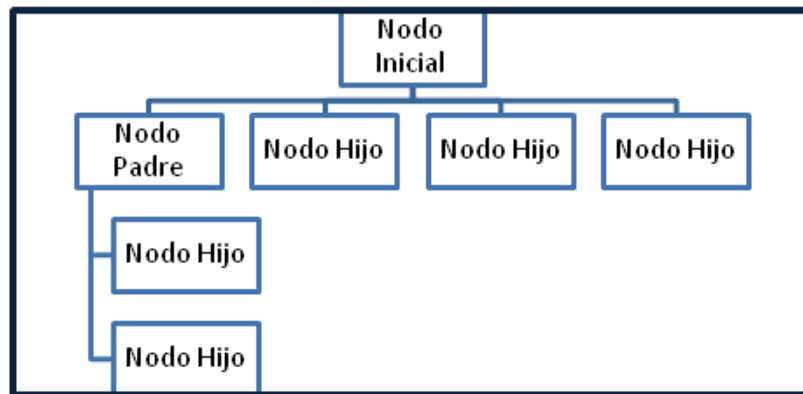


Figura 4.6: Criterio 2 - CHAID

Un árbol de decisión no utiliza un modelo estadístico formal, utiliza un algoritmo para clasificar los datos mediante los valores de las variables.

Algunas de las opciones mencionadas a continuación se pueden definir mediante las opciones avanzadas de CHAID. Estas opciones incluyen la posibilidad de seleccionar la prueba de Chi-cuadrado de Pearson o la de la razón de verosimilitud, el nivel de  $\alpha_{merge}$  y  $\alpha_{split}$ , la profundidad el árbol o el número de individuos que debe tener como mínimo cada nodo.

A continuación se define de forma resumida el algoritmo *CHAID*.

1. Para cada variable predictora  $X$ , se busca el par de categorías de  $X$  cuya diferencia presenta la menor significación (es decir, la que presente el mayor nivel crítico) respecto a la variable criterio  $Y$ . El método usado para calcular el nivel crítico depende del nivel de medida de  $Y$ .

- Si  $Y$  es continua, utilizar la prueba  $F^7$ .

<sup>7</sup>La prueba  $F$  tiene como supuesto:

- Para cada categoría  $X_i$  de la distribución de la variable  $Y$  es normal.
- La desviación estándar de la distribución  $Y$ , es la misma para cada grupo.
- La población de cada categoría  $X_i$  es independiente.

- Si  $Y$  es nominal, se forma una tabla de contingencia con las categorías de  $X$  como filas y las categorías de  $Y$  como columnas. Utilizar la prueba de chi-cuadrado de Pearson o la prueba de la razón de verosimilitud. (Ver prueba de Independencia).
  - Si  $y$  es ordinal, se ajusta un modelo de asociación de  $Y$  (Clogg y Eliaisin, 1987; Goodman, 1979 y Madgidson, 1992). Utilizar la prueba de la razón de verosimilitud  $H^2$ <sup>8</sup>. De la misma manera al obtener el estadístico se realizará la prueba chi-cuadrada, obteniendo así su respectivo valor-p.
2. Para el par de categorías de  $X$  con el mayor nivel crítico, se compara el valor del nivel crítico con el nivel alfa preestablecido  $\alpha_{merge}$ :
- Si el nivel crítico es mayor que  $\alpha_{merge}$ , fundir este par en una sola categoría compuesta, como resultado, se forma un nuevo conjunto de categorías de  $X$  y el proceso se vuelve a iniciar desde el paso 1.
  - Si el nivel crítico es menor que  $\alpha_{merge}$  ir al paso 3.
3. Calcular el nivel crítico corregido para el conjunto de las categorías de  $X$  y las categorías de  $Y$ , mediante la corrección de Bonferroni correspondiente.
4. Seleccionar la variable predictora  $X$  cuyo nivel crítico corregido sea el menor (la que sea más significativa). Comparar su nivel crítico con el nivel alfa preestablecido  $\alpha_{split}$ .
- Si el nivel crítico preestablecido es menor o igual que  $\alpha_{split}$  dividir el nodo conforme al número de categorías de  $X$ .
  - Si el nivel crítico preestablecido es mayor que  $\alpha_{split}$  no dividir el nodo. Este nodo es un nodo terminal.
5. Continuar con el proceso de desarrollo del árbol hasta que se cumpla una de las reglas de parada.

El filtro y las reglas de parada, son criterios que se establecen para que el análisis de segmentación tenga límites, pues sino se puede dar el caso de que se produzca una gran cantidad de nodos de tamaño muy pequeño y difíciles de segmentar.

En primer lugar se puede hablar de un *criterio de segmentación*, que se encarga de no permitir segmentaciones que no sean estadísticamente significativas. Este filtro puede ser aplicado en la agrupación de categorías de una variable y en la selección del mejor pronosticador. La  $\alpha_{merge}$  de 0.05 es el valor utilizado más comúnmente. Si la significación (valor-p) de la prueba estadística de dos categorías de la variable independiente es menor que este valor, se rechaza la hipótesis nula provocando que las 2 susodichas categorías queden separadas y se pueda proseguir con la segmentación. En cambio, si el valor es superior a 0.05 las categorías se funden, si quedan agrupadas todas las categorías de todas las variables, la segmentación se detiene. Los valores extremos permiten comprender con mayor eficacia el efecto del filtro de segmentación. Si se escoge el mayor valor posible de la entonces, la agrupación o reducción de categorías de las variables se convierte en imposible y, siempre que haya significación entre pronosticador y variable dependiente, la segmentación se formará con una determinada variable tantos grupos como categorías se tengan, tendiendo a crear una segmentación más frondosa, más amplia.

Luego se debe tomar en cuenta reglas de parada como el tamaño, nivel y pureza.

- **Tamaño:** Su principal objetivo es evitar que se formen grupos muy pequeños durante el proceso de segmentación. La regla de tamaño puede aplicarse en dos momentos: antes de la segmentación (Na, nodo padre) y después de la segmentación (Nd, nodo hijo). En el caso del nodo padre, la segmentación se detiene si el nodo que se quiere separar tiene un tamaño menor a Na. En el caso del nodo hijo, no se puede formar un grupo si no tiene un número establecido de componentes, es decir, si al crear un nuevo nodo (nodo hijo) su tamaño es menor a Nd no se crea el nodo y se funde con la categoría más similar.
- **Nivel:** Consiste en determinar un nivel (Ns, profundidad) máximo de segmentación. Por nivel se entiende cada una de las franjas horizontales del árbol. La primera

franja horizontal corresponde al nodo principal, la segunda a la primera segmentación, la tercera a la segunda y así sucesivamente. Este filtro evita que se formen múltiples segmentaciones en segmentos grandes de los datos. Asimismo, contribuye a simplificar los resultados en la medida en que reduce directamente el número de variables necesarias para predecir la variable dependiente.

- **Pureza:** Si un nodo es puro, es decir, todos los casos del nodo tienen el mismo valor para la variable dependiente  $Y$ , el nodo no será dividido y se detendrá la segmentación. Mientras que, si en la segmentación de un nodo todos los nodos hijos tienen los mismos valores de  $Y$  y se ha llegado al nivel de profundidad deseado, el nodo no se divide.

La variable criterio a usarse para la construcción de los árboles de decisión es la variable Tipo Cliente, en cada fase del análisis, CHAID realiza una prueba de independencia Chi Cuadrado entre la variable dependiente y la variable predictora escogida. En la actualidad existen diferentes programas para la realización de árboles de decisión, para la presente investigación se utilizará Clementine SPSS 12.

#### 4.8.4. PRUEBA DE INDEPENDENCIA CHI - CUADRADA

Consiste en comprobar si dos características cualitativas están relacionadas entre sí (por ejemplo: ¿el color de ojos está relacionado con el color de los cabellos?). Este tipo de contraste se aplica cuando deseamos comparar una variable en dos situaciones o poblaciones diferentes, i.e ,deseamos estudiar si existen diferencias en las dos poblaciones respecto a la variable de estudio. Supongamos que de  $n$  elementos de una población se han observado dos características  $X$  e  $Y$ , obteniéndose una muestra aleatoria simple bidimensional  $(X_1, Y_1); (X_2, Y_2); \dots; (X_n, Y_n)$ . Sobre la base de dichas observaciones se desea contrastar si las características poblacionales  $X$  e  $Y$  son independientes o no. Para ello se dividirá el conjunto de posibles valores de  $X$  en  $k$  conjuntos disjuntos  $A_1, A_2, \dots, A_k$ ; mientras que el conjunto de posibles valores  $Y$  será descompuesto en  $r$  conjuntos disjuntos:  $B_1, B_2, \dots, B_r$ . Al clasificar los elementos de la



muestra, aparecerá un cierto número de ellos,  $n_{ij}$ , en cada una de las clases  $k \times r$  así constituidas, dando lugar a una tabla de contingencia de la forma:

	A <sub>1</sub>	A <sub>2</sub>	....	A <sub>k</sub>	Total
B <sub>1</sub>	n <sub>11</sub>	n <sub>12</sub>		n <sub>1k</sub>	n <sub>1.</sub>
B <sub>2</sub>	n <sub>21</sub>	n <sub>22</sub>		n <sub>2k</sub>	n <sub>2.</sub>
...					
B <sub>r</sub>	n <sub>r1</sub>	n <sub>r2</sub>		n <sub>rk</sub>	n <sub>r.</sub>
Total	n <sub>,1</sub>	n <sub>,2</sub>		n <sub>,k</sub>	n

El estadístico será:  $X^2 = \sum_{i=1}^r \sum_{j=1}^k \frac{(e_{ij} - n_{ij})^2}{e_{ij}^2}$ , sigue una distribución  $X^2_{con}(k-1)(r-1)$  grados de libertad.

Donde  $e_{ij} = \frac{1}{n} n_j^2$

## Capítulo 5

# METODOLOGÍA PARA LA CONSTRUCCIÓN DE UN SCORING DE APROBACIÓN

### 5.1. INTRODUCCIÓN

En el actual sistema financiero las entidades compiten por encontrar y captar la mayor cantidad de clientes. Las entidades financieras deben ser precisas y eficientes en la identificación y obtención de clientes más apropiados para su negocio, lo esencial en esta etapa es definir cómo y qué clientes atraer; una de las herramientas para conseguir este objetivo es construir un *scoring de aprobación*. Los *scoring de aprobación* son modelos de puntuación predictivos con capacidad para determinar el riesgo de incumplimiento futuro asociado a una operación crediticia, ordenando la población y permitiendo manejar ofertas comerciales diferenciadas por punto de corte. El modelo se obtiene a partir de variables recabadas en la solicitud de crédito y de la información proporcionada por centrales de riesgo. En la presente investigación se utilizó información interna y externa de los créditos aprobados; la información interna comprende principalmente información socio demográfica y de ingresos del cliente; mientras que, la fuente de información externa corresponde a la otorgada por el buró de crédito, per-

mitiendo conocer el historial crediticio de una persona que aplica, es decir, si presenta o no deuda en los sectores real, financiero y comercial, así como también los saldos que presentan en la diferentes carteras de crédito donde disponga de información; y la calificación de riesgo asignada en cada tipo de crédito.

Lo que se detalla en el presente capítulo es la realización de un scoring de aprobación mediante la utilización de modelos lineales generalizados mixtos, que se usan para incorporar efectos macroeconómicos que más adelante servirá para establecer los *perfiles* de los clientes. Además, se utilizará clasificadores bayesianos que servirán para seleccionar las variables a ser usadas y obtener grupos homogéneos.

### 5.2. METODOLOGÍA

La metodología para la obtención de un modelo score se la puede dividir en la consecución de cinco pasos, los cuales aplicados de manera sistemática permiten alcanzar con éxito lo propuesto, a continuación se detalla dichos pasos:

#### 5.2.1. SELECCIÓN DE LA MUESTRA DE ANÁLISIS

La finalidad del scoring es determinar el mejor universo de clientes, aquellos que sean más rentables y para los cuales se puedan determinar las mejores estrategias, sin incurrir en altos niveles de riesgo, además la selección de la muestra, por otro lado se traduce en determinar adecuadamente el periodo de observación.

Naturalmente, la muestra debe incorporar información suficiente de manera que ésta recoja el tipo de comportamiento de pago, y de esa forma, identificar características que reflejen esas conductas en los nuevos solicitantes de crédito. Por lo tanto, se deben establecer las fechas inicial y final del periodo de observación de tal modo que la muestra evidencie la calidad de pago del cliente y que a su vez sea representativa de la población.

Para cumplir con este propósito se utiliza un indicador de tasa de morosidad de la cartera en cuestión, el cual recoge el porcentaje de clientes catalogados como malos pagadores respecto a la colocación de un mes en particular. Mediante un análisis de cosechas se resuelve que los clientes con pago mayor a 30 días conforman el indicador de tasa de morosidad, de forma específica:

$$\text{Indicador\_de\_mora} = \frac{\text{Número\_con\_Pago} > 30}{\text{Número\_Total\_de\_Clientes\_Aprobados}} \quad (5.1)$$

En base a la serie generada por los valores del índice Tasa de Morosidad se debe resolver en qué periodo la proporción de la cartera es madura<sup>1</sup> y estable<sup>2</sup>, obteniéndose de esa manera la fecha final.

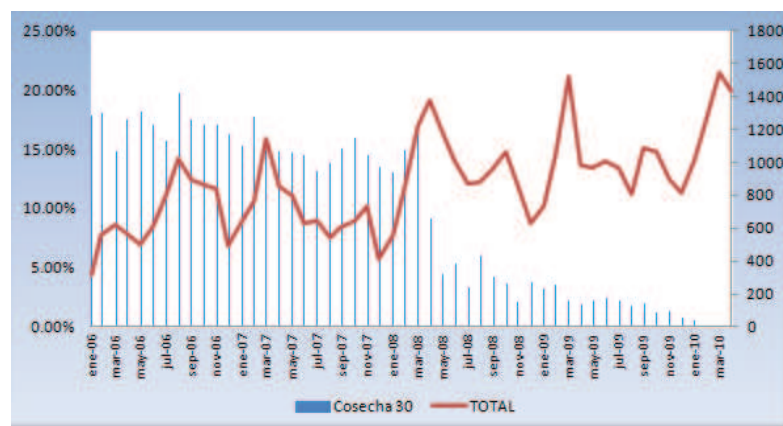


Figura 5.1: Tasa de morosidad y créditos colocados

En la figura anterior se observa que la tasa de morosidad se mueve alrededor de un 15 % desde enero de 2006 hasta mayo de 2008, mostrando a partir del mes siguiente un continuo declive. Por lo tanto, de manera descriptiva, se puede declarar que la cartera

<sup>1</sup>Se dice que la cartera ha madurado si los clientes asociados a ésta ya han presentado su comportamiento real de pago.

<sup>2</sup>Se dice que la cartera es estable si la tasa de morosidad de ésta, en un período determinado, fluctúa alrededor de un valor fijo. Estadísticamente, para verificar que la cartera es estable se debe realizar un análisis de quiebre estructural para determinar el tramo o período donde se pueda considerar que el índice se mantiene estable.

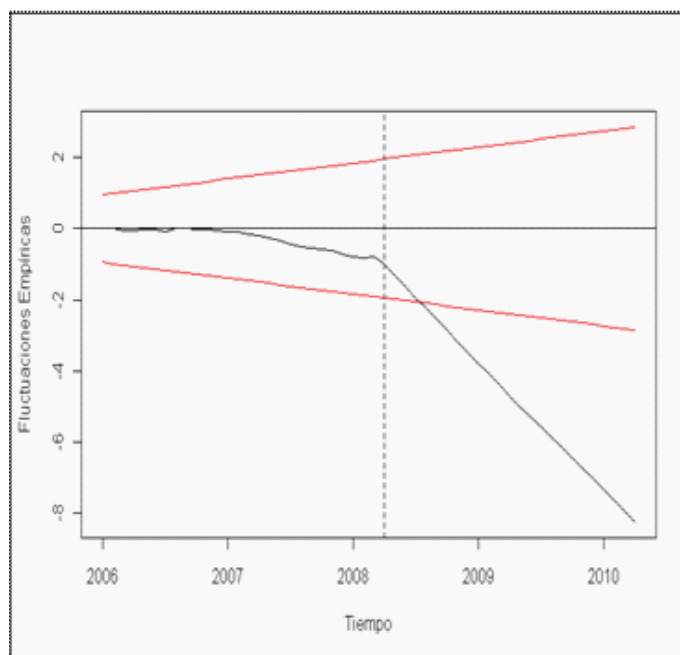


Figura 5.2: Prueba CUSUM recursiva para el índice de tasa de morosidad

es madura en aquel periodo, aseveración que es corroborada estadísticamente con el análisis de estabilidad.

En base a la figura anterior 5.2 se deduce que la serie tasa de morosidad presenta un quiebre estructural en el mes de abril del 2008. El  $p\text{-value} = 1.685e13$  asociado al estadístico de prueba  $S = 2.7433$  para el proceso CUSUM<sup>3</sup> permite rechazar la hipótesis nula de ausencia de quiebre estructural. Además, la banda de confianza permite considerar al mes de mayo como parte del periodo de estabilidad.

Una vez establecido el periodo de la muestra es necesario definir, mediante algún criterio que relacione el comportamiento de pago de los clientes, qué cliente es calificado

<sup>3</sup>CUSUM es una prueba basada en técnicas que son diseñadas para mostrar posibles desviaciones de estabilidad a través de un análisis gráfico. Este método implica el cálculo de la suma acumulativa de residuos recursivos. La idea central de estos procedimientos se puede sintetizar en dos pasos. En el primero, se estiman los coeficientes del modelo de regresión para el periodo histórico y en el segundo, basados en estas estimaciones, se calculan los residuos de las observaciones en el periodo de monitoreo. De tal forma que, si existe un cambio estructural en el periodo de monitoreo, los residuales deberían desviarse sistemáticamente de su media cero. Si se calculan límites de confianza, se obtiene una banda de confianza que debería acotar completamente la evolución de cualquier serie de residuos que obedeciera la hipótesis de estabilidad.

como “malo” o “bueno”. Mediante una definición empírica dada por los ejecutivos de la entidad se decide clasificar a un cliente como malo a aquel que presente un atraso máximo mayor a 30 días y una mora promedio distinta de cero; mientras que aquel que no presente mora alguna será catalogado como bueno. De esta manera se tiene un número de 13390 clientes buenos y 1632 clientes malos.

### CONSISTENCIA DE LA MUESTRA

Las fuentes de información disponibles son la base de datos de la Institución Financiera (información interna) y del Buró de Crédito (información externa). La primera de ellas contiene variables del cliente, de aquí se obtienen principalmente variables socio demográficas. La segunda fuente tiene por objeto el apoyar al sistema financiero e integrar y proporcionar información, previo a la concesión de un crédito, cuyo objetivo principal es registrar el historial crediticio de las personas que hayan obtenido algún tipo de crédito, financiamiento, préstamo o servicio.

Adviértase que el número de clientes buenos (ver Anexo 7.1) es ocho veces el número de malos, así que sea cual fuese el método a utilizar, el modelo carecerá de poder discriminante, por eso cabe la necesidad de extraer una muestra de clientes buenos.

#### 5.2.2. SELECCIÓN DE CLIENTES BUENOS Y MALOS

El segundo paso de la metodología propuesta es obtener un criterio apropiado para establecer si un cliente es bueno o malo; esta definición es determinante en el desarrollo del modelo estadístico a obtenerse debido a que de esta categorización resulta la variable dependiente de cualquier método a aplicarse.

La manera de establecer estos conceptos es a través de una definición empírica dada por ejecutivos de la entidad donde se desarrolla el modelo, analizando cuadros de estadística descriptiva basados en variables auxiliares que relacionen el comportamiento

Rango de Atraso								
Rango	0	1-15	16-30	31-60	61-90	91-120	>120	Total
	13.392							13.992
1-15		941						941
16-30		1.555						1.555
31-60		1.001	1					1.001
61-90		260	11	2				260
91-120		55	36	2				55
>120		10	40	53	30	17	114	10
Total	13.392	3.822	88	57	30	17	114	17.520

Cuadro 5.1: Número de clientes Buenos Malos

de pago de los clientes, tales como mora máxima histórica, mora promedio, contadores de mora (reincidencia), entre otros, que indiquen cómo se clasifica al cliente al interior de una institución financiera.

El planteamiento que se presenta utiliza las variables mora máxima y mora promedio, de tal manera que se categoriza como buenos clientes a aquellos cuyo comportamiento de pago proporciona la mayor rentabilidad y como malos a aquellos que causen mayor impacto en la pérdida de la institución.

De lo anterior se puede definir como buenos a los clientes que se encuentran en moras promedio y máxima igual a cero días, la definición depende del nivel de riesgo que está dispuesta a correr cada institución. Como malos se puede reconocer a los clientes que registren una mora diferente de cero y máxima de 30 días en adelante, que representan un mayor riesgo de crédito. Ver cuadro (5.1).

Según esta definición la cartera de buenos clientes bordea el 76 % de participación del periodo de observación establecido como muestra para el modelo scoring y los malos clientes representan el 9 %; el restante corresponde a cartera “indeterminada”, que por su rendimiento no se la puede clasificar como buena o mala, lo que no altera en si la definición ni la construcción del modelo.

### 5.2.3. SELECCIÓN DE GRUPOS HOMOGÉNEOS

Debido a la estructura del modelo GLMM y la cantidad de créditos que conforman el portafolio, se hace necesario segmentar dicho portafolio en grupos homogéneos, ya que no es posible realizar el modelamiento del incumplimiento de crédito por las siguientes razones:

- Los créditos no son originados al mismo tiempo ni tienen el mismo plazo por lo complica el análisis del portafolio.
- Pueden existir créditos con buen comportamiento crediticio por lo que la estimación del incumplimiento conjunto del portafolio puede ser inconsistente.
- Para cuantificar la dependencia entre incumplimientos se tiene que disponer de la matriz de correlación, por lo que en portafolios con gran cantidad de créditos esta matriz es difícil de construir.

A partir de lo anterior se evidencia la necesidad de identificar grupos homogéneos en el portafolio crediticio para un mejor análisis de la dependencia de créditos; sin embargo, esta agrupación no puede ser realizada a partir de los días de mora del crédito ya que a priori no se conoce el comportamiento de los créditos. En este sentido la técnica utilizada para segmentar la cartera de crédito es un árbol de decisión también conocido como *detección de interacción automática Ji-cuadrada* que permite identificar divisiones óptimas cuando las variables son continuas, nominales e incluso categóricas. Este procedimiento, como habíamos mencionado anteriormente, determina una serie de reglas que permite obtener una partición del portafolio original a partir de alguna variable clave.

No obstante cabe recordar que el número de variables con las que se cuenta información es de 101. Naturalmente que no todas ellas pueden formar parte del árbol de decisión ya que podrían existir variables que no estén relacionadas con la variable objetivo Tipo de Cliente “bueno/malo” y dificultar de esta forma la selección de grupos homogéneos. Lo que concierne ahora, es establecer el conjunto de variables explicativas que se deben considerar como candidatas a ser incluidas en el modelo, para lo



cual se utilizó clasificadores bayesianos, usando toda la población, como se muestra a continuación.

Con la ayuda del software CLEMENTINE, análisis previos (ver figura 5.3) no muestran claramente una selección reducida de variables, para lo cual mediante la eliminación de las variables menos importantes<sup>4</sup> y usando el *Criterio de Pareto*, se realizó varias corridas obteniendo la red final, como se muestra en la figura (5.4).

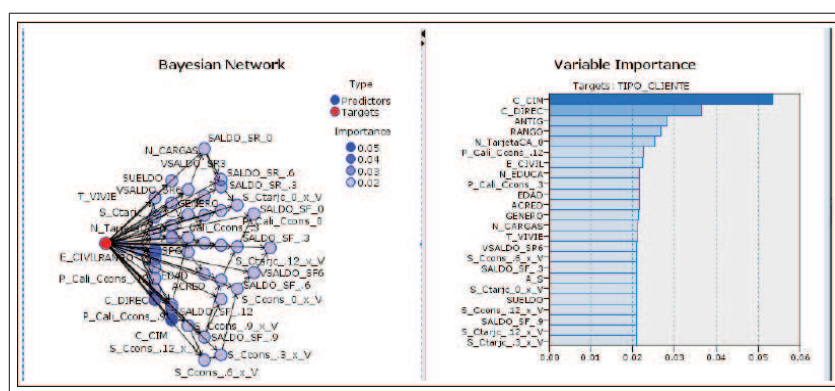


Figura 5.3: Red Inicial para seleccionar variables mediante algoritmo TAN

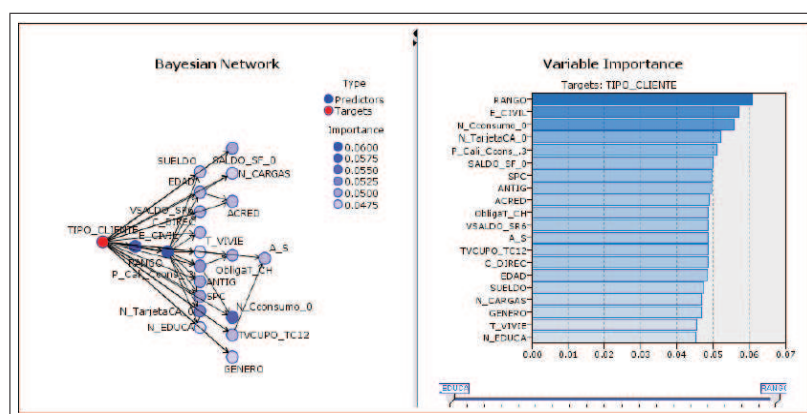


Figura 5.4: Red Final para seleccionar variables mediante algoritmo TAN

<sup>4</sup>El gráfico de importancia de predictores muestra la importancia relativa de cada predictor en la estimación del modelo. La suma de los valores de todos los predictores de la visualización es 1.

La importancia del predictor está relacionada con la importancia de cada variable predictora a la hora de realizar la predicción.

Variables Internas	Variables Buró de Crédito
Rango	No de créditos de consumo del cliente (N_Consumo_0)
Estado Civil	Número de tarjetas de crédito activas (N_TarjetasCA_0)
Ciudad de Residencia del Cliente	Peor calificación de crédito de consumo (P_Calif_Cons_0)
Edad	Acreditación (ACRED)
Sueldo	Tasa de variación de cupo de tarjetas de crédito respecto a la acreditación (TVCUPO_TC12)
Saldo promedio de la cuenta	
Antigüedad de la cuenta(Antig)	
Número de Cargas (N_Cargas)	
Género	
Tipo de Vivieda	
Nivel de Educación	

Cuadro 5.2: Variables seleccionadas mediante Clasificadores Bayesianos

La figura (5.4) contiene una red con gráficos de nodos que muestra la relación entre la variable objetivo y sus predictores más importantes, así como las relaciones entre variables predictoras. La importancia de cada variable predictora se muestra según la densidad del color, un color más fuerte indica un predictor más importante y viceversa. Así, se puede visualizar que la variable más importante es la variable Rango, seguido de la variable estado civil que ésta a su vez directamente relacionada con las variables edad, ciudad de residencia, tipo de vivienda, nivel de educación, también podemos visualizar una relación directa entre la variable *Tipo de Cliente* con el saldo promedio de la cuenta, número de cargas y género. Finalmente en el cuadro (5.2) podemos observar cuales son las variables que mejor pronostican la variables objetivo<sup>5</sup>.

Por otro lado, con el propósito de evaluar la precisión de los modelos de clasificadores bayesianos en base a la relación de las variables predictoras con la variable Tipo de Cliente, se usó validación cruzada usando datos en muestras independientes para fases de entrenamiento, prueba y validación en la generación del modelo. Se usó una muestra para generar el modelo y otra muestra distinta para comprobarlo y así se obtuvo una buena adecuación del modelo. Además se realizó la comprobación de dos posibles modelos de clasificadores bayesianos obteniendo matrices de confusión (cuadro 5.3), mientras que en el cuadro (5.4) se muestran los resultados en términos

<sup>5</sup>En el Anexo 7.5 se describe las categorías de las variables discretas.

## 5. METODOLOGÍA PARA LA CONSTRUCCIÓN DE UN SCORING DE APROBACIÓN

83

Predicha/Real	Red Inicial		Red Final	
	Buenos	Malos	Buenos	Malos
Buenos	13.049	339	12.339	18
Malos	1263	366	678	982

Cuadro 5.3: Resultados de Modelos de Clasificadores Bayesianos, expresados en Matriz de Confusión.

Modelo	Sensibilidad	Especificidad	Valor Predictivo Positivo	Valor Predictivo Negativo
Red Inicial	93.85 %	93.91 %	99.63 %	46.36 %
Red Final	98.14 %	97.41 %	98.86 %	73.86 %

Cuadro 5.4: Resultados de Modelos de Clasificadores Bayesianos, en términos de la sensibilidad, la especificidad, valor de la predicción de clientes buenos y clientes malos.

de sensibilidad, especificidad, valor de la predicción de los clientes “buenos”, y valor de la predicción de los clientes “malos”.

Nótese en el cuadro 5.4, la red final obtuvo una especificidad de 98.20 %, sensibilidad de 95.16 %, presentando un valor de predicción de buenos clientes de 99.89 %, mientras que el valor de predicción para los clientes definidos como malos es de 59.16 %, con lo cual se concluye que las variables obtenidas en la red final son las que permiten un mejor predicción de la variable Tipo Cliente.

Finalmente la figura 5.5 muestra las curvas ROC obtenidas para los modelos de redes obtenidas, con un intervalo de confianza de 95 % se observa que la red final tiene el área sobre la curva más alta, comprobando de esta forma que la red final con las veinte variables seleccionadas anteriormente predice de mejor manera la variable tipo de cliente.

La anterior selección de variables (cuadro 5.2) facilita de esta forma, el objetivo final que es la obtención de grupos homogéneos del portafolio de manera consistente, que se basa en valores de una de las variables predictoras que para nuestro caso es la variable *Rango*.

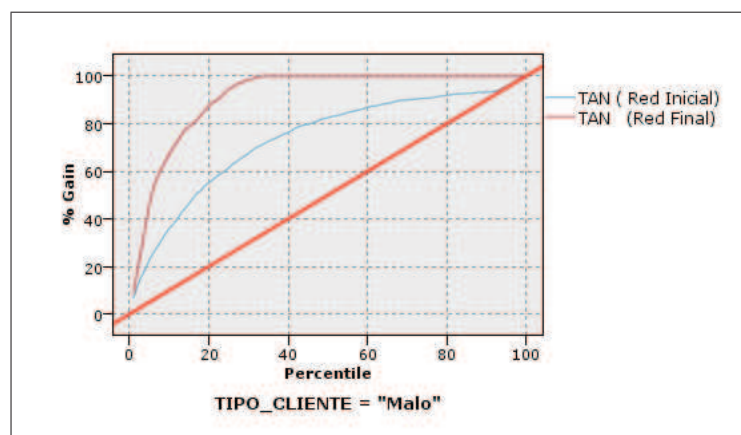


Figura 5.5: Curvas ROC

Nodo por Nodo						
Nodo	Nodo:n	Nodo (%)	Ganancia:n	Ganancia (%)	Resp (%)	Índice (%)
11	546	15.94	387	29.93	70.83	187.75
8	210	6.13	122	9.44	58.10	153.89
12	520	15.18	221	17.09	42.50	112.58
7	464	13.55	191	14.77	41.16	119.04
5	239	6.98	96	7.42	40.17	106.40
10	214	6.25	72	5.57	33.64	89.12
2	152	4.44	29	2.24	19.08	50.54
9	584	17.05	102	7.89	17.47	46.46
6	496	14.48	73	5.65	14.72	38.99

Cuadro 5.5: Cuadro resumen de los tres grupos crediticios homogéneos de una cartera de consumo creados a partir del método CHAID

A partir del método CHAID, y usando la muestra de clientes<sup>6</sup> inicialmente se identifican nueve grupo homogéneos (Cuadro 5.5), que en promedio tiene la misma cantidad de clientes definidos como buenos malos (Ver Anexo 7.6).

Finalmente tras utilizar como variable dependiente, la variable tipo cliente y como variables independientes las seleccionadas por los *clasificadores bayesianos*, el método CHAID arrojó una segmentación de tres grupos, donde el primer grupo (G1) tiene un 16.3% de clientes definidos como malos, el segundo 40.45%, y el tercero 67.09% (5.7). En otras palabras, se puede evidenciar que cada grupo homogéneo tiene un

<sup>6</sup> Como se mencionó anteriormente, el número de clientes buenos es ocho veces el número de clientes malos.

Estadísticas Acumuladas			
Nodo:n	Nodo (%)	Ganancia:n	Ganancia (%)
546	15.94	387	29.93
756	22.07	509	39.37
1276	37.25	730	56.46
1740	50.08	921	71.23
1979	57.78	1017	78.65
2193	64.03	1089	84.22
2345	68.47	1118	86.46
2929	85.52	1220	94.35
3425	100	1293	100

Cuadro 5.6: Cuadro acumulado de los grupos crediticios homogéneos de una cartera de consumo creados a partir del método CHAID

Grupos Homogéneos			
Subgrupo	Nº de créditos	% Malos	%Buenos
G1	1531	16.3 %	83.7 %
G2	1812	40.45 %	59.55 %
G3	963	67.08 %	32.92 %

Cuadro 5.7: Cuadro resumen de los tres grupos crediticios homogéneos de una cartera crediticia de consumo creados a partir del método CHAID.

comportamiento crediticio; así el grupo G1 presenta menor proporción de clientes malos, mientras que el grupo G3 tiene un mayor porcentaje de clientes malos.

#### 5.2.4. SELECCIÓN DEL MODELO

Una vez clasificada la base de datos en grupos homogéneos de clientes, se realiza el modelo lineal generalizado mixto a fin de calcular la probabilidad de incumplimiento. Desde un punto de vista práctico estos modelos representan tal vez la forma más útil de analizar y comparar un portafolio crediticio, en el que se asume independencia condicional de incumplimiento dados factores estocásticos comunes subyacentes; es decir, los incumplimientos individuales son independientes; pero la dependencia entre los incumplimientos se debe a la dependencia de las probabilidades de incumplimiento de cada individuo con un conjunto común de factores económicos, como variables

macroeconómicas.

De esta manera, se ajusta un modelo lineal generalizado mixto con datos mensuales para el periodo enero 2006 – mayo 2008 (fecha de concesión del crédito), grupo al que pertenece cada cliente, número total de créditos concedidos y número de clientes que incumplieron el pago. En el modelo se asume un solo efecto aleatorio que representa “el estado de la economía”; el papel de los efectos aleatorios en el GLMM es capturar patrones de variabilidad en las respuestas que no se pueden explicar solamente por las covariables observadas, sino que se podría explicar por un factor adicional no observado. En nuestro caso, estos factores no observados se encuentran agrupados en el índice adelantado de la economía (ver Anexo 7.2); alternatively nos referimos como riesgo sistemático. Por simplicidad, para el modelo realizado se supone que los efectos aleatorios son independientes e idénticamente distribuidos; es decir siguen una distribución normal.

En primer lugar se realiza un modelo en el que no se considera el índice adelantado de la economía ecuatoriana, dando como resultado, la siguiente probabilidad de incumplimiento:

$$p_{ij} = \frac{1}{1 + e^{-(1,6511*G_1 - 0,4341*G_2 + 0,6587*G_3)}} \quad a_i \sim N(0, 0,32^2) \quad (5.2)$$

Al aplicar el modelo anteriormente descrito (GLMM), la probabilidad de incumplimiento en el caso en que la economía se encuentra en “valle” está dado por:

$$p_{ij} = \frac{1}{1 + e^{-(0,8700*G_1 + 0,3466*G_2 + 1,4395*G_3 - 2,0367*Z)}} \quad a_i \sim N(0, 0,28^2) \quad (5.3)$$

Es decir:

- Si se concedió un crédito en junio 2006 y el cliente pertenece al grupo 1 ( $G_1$ ) la probabilidad de incumplimiento es:

$$p_{ij} = \frac{1}{1 + e^{-(0,8700 - 2,0367 * 0,3407)}} = 0,1730 \quad a_i \sim N(0, 0,28^2)$$

- Si la fecha de conseción fue junio 2006 y el cliente pertenece al grupo 2 ( $G_2$ ) la probabilidad de incumplimiento está dada por:

$$p_{12} = \frac{1}{1 + e^{-(0,3466 - 2,0367 * 0,3407)}} = 0,4140 \quad a_i \sim N(0, 0,28^2) \quad (5.4)$$

- De manera similar, si el cliente pertenece al grupo 3 ( $G_3$ ) y el crédito fue concedido en junio 2006 la probabilidad de incumplimiento es:

$$p_{12} = \frac{1}{1 + e^{-(1,4395 - 2,0367 * 0,3407)}} = 0,6781 \quad a_i \sim N(0, 0,28^2)$$

Como se puede observar, la probabilidad de incumplimiento depende del grupo homogéneo al que pertenece el cliente; además se evidencia que la probabilidad de incumplimiento para el cliente del grupo 1 es más baja, debido a que como se mencionó anteriormente (selección de grupos homogéneos) este grupo está conformado en menor proporción por clientes definidos como malos.

En cambio, cuando la economía se encuentra en “pico” la probabilidad de incumplimiento se obtiene a partir de:

$$p_{ij} = \frac{1}{1 + e^{-(1,6315 * G_1 - 0,4145 * G_2 + 0,6782 * G_3 - 0,0425 * Z)}} \quad a_i \sim N(0, 0,32^2) \quad (5.5)$$

Por ejemplo, para la misma fecha de conseción que en el caso anterior, se tiene:

- Un crédito consedido en el periodo junio 2006, y el cliente segmentado en el grupo 1; la probabilidad se calcula:

$$p_{11} = \frac{1}{1 + e^{-(-1,6315 - 0,0425 * 0,5063)}} = 0,1606 \quad a_i \sim N(0, 0,32^2) \quad (5.6)$$

- Para el mismo periodo de tiempo, pero para el grupo 2 se tiene la siguiente probabilidad:

$$p_{12} = \frac{1}{1 + e^{-(-0,4145 - 0,0425 * 0,5063)}} = 0,3926 \quad a_i \sim N(0, 0,32^2) \quad (5.7)$$

- Si un cliente pidió un crédito en junio 2006 y se clasificó como grupo 3 se obtiene:

$$p_{13} = \frac{1}{1 + e^{-(0,6782 - 0,0425 * 0,5063)}} = 0,6585 \quad a_i \sim N(0, 0,32^2) \quad (5.8)$$

En el ejemplo expuesto para el periodo junio 2006 se tienen diferentes probabilidades de incumplimiento para el mismo cliente, dependiendo del estado en el que se encuentre el ciclo económico; en este caso específicamente, se debe tomar la probabilidad de incumplimiento calculada a partir del modelo de pico debido a que la probabilidad que ocurra pico es 0,50 y la de valle 0,34 (ver Anexo 7.2). Los resultados son coherentes ya que al estar la economía en pico la probabilidad de incumplimiento debe ser más baja pues existe menor factor de riesgo para que se incumpla. De esta manera, cuando se está en “auge” se estima con la probabilidad mensual de un pico; mientras que para la fase de decrecimiento se calcula con la probabilidad de valle.

### 5.2.5. VALIDACIÓN DEL MODELO

Al aplicar un modelo lineal generalizado mixto se está asumiendo implícitamente una serie de supuestos (supuestos sobre los errores), y una vez obtenidos los resultados del modelo, es necesario verificar todos ellos; esto se llama el proceso de validación del modelo. A continuación se verifica la existencia de normalidad en los errores.



El análisis de los errores que a continuación se realizará consiste en constatar que los residuos del modelo lineal generalizado mixto siguen una distribución normal. Esta hipótesis se probará tanto de manera gráfica como de forma analítica.

El objetivo de una prueba de normalidad no es tanto establecer normalidad, sino evaluar el alejamiento de los datos respecto a un modelo normal. En este sentido, el método gráfico es una excelente ayuda para tomar la decisión apropiada, pues el ojo humano detecta más fácilmente los puntos alejados con respecto a una recta que con respecto a otra figura [44].

Para verificar que los residuos se distribuyen normalmente se grafica los errores.

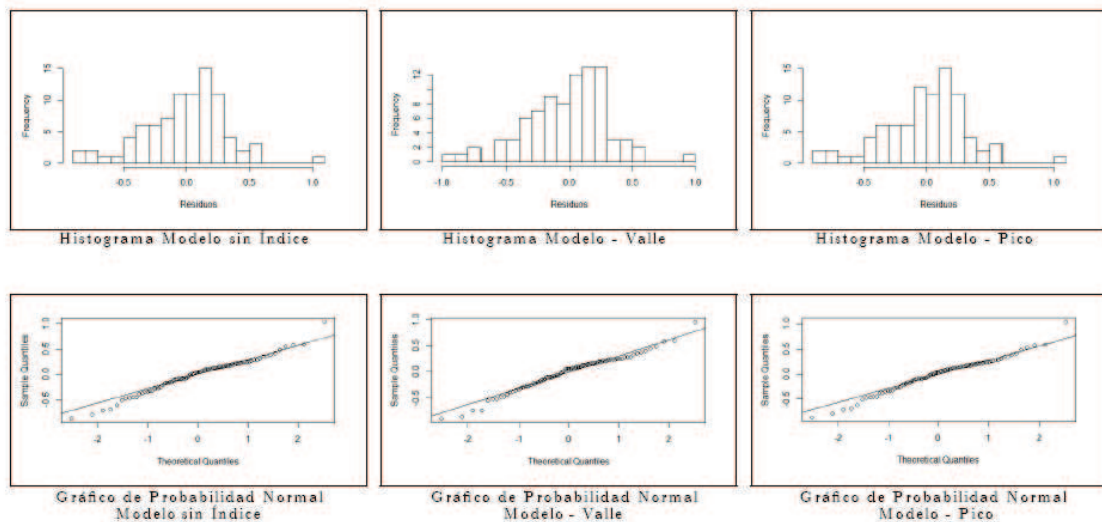


Figura 5.6: Residuos de los modelos desarrollados.

La distribución de los residuos debe ser normal, es decir, los residuos observados y esperados bajo la hipótesis de distribución normal deben ser parecidos. Esta suposición se comprueba tanto con el histograma como con el gráfico de probabilidad normal, que permite comparar gráficamente la función de distribución observada, con la función de distribución normal. En la figura 5.6 correspondiente a los residuos de los tres modelos realizados, se observa para cada uno un buen ajuste a la normal.

Para realizar una prueba de manera formal se plantea la siguiente hipótesis:

$H_o$  : Los residuos se distribuyen normalmente

$H_1$  : Los residuos no se distribuyen normalmente

Existen diversos procedimientos para probar la hipótesis planteada, para este caso se realizará la prueba de Shapiro - Wilks; este test es un contraste de ajuste que se utiliza para comprobar si unos datos determinados han sido extraídos de una población normal, es adecuado para muestras pequeñas ( $n < 50$ ) [45].

El estadístico propuesto por Shapiro-Wilk se basa en medir el ajuste de los datos a una recta probabilística normal. Si el ajuste fuera perfecto los puntos formarían una recta de  $45^\circ$  (frecuencia observada igual a frecuencia esperada). El estadístico de contraste se expresa por medio de la siguiente ecuación:

$$W = \frac{1}{\sum_{j=1}^n (x_j - \mu)^2} \left[ \sum_{j=1}^h a_{j,n} (x_{(n-j+1)} - x_j) \right]^2$$

donde  $n$  es el número de datos,  $x_j$  es el dato en orden ascendente de muestra que ocupa el lugar  $j$ ,  $\mu$  es la media,  $h$  es  $n/2$  si  $n$  es par o  $(n-1)/2$  si  $n$  es impar y  $a_{j,n}$  es un valor tabulado.

Una vez calculado el estadístico  $W$  se contrasta con un valor  $W$  crítico para el nivel de significación elegido. Como este estadístico mide el ajuste a una recta y no la distancia a la distribución normal, la hipótesis nula se acepta cuando el valor  $W$  es superior al valor de contraste tabulado (valor de ajuste muy alto) [46].

A continuación en la siguiente tabla se presenta la prueba de Shapiro - Wilks para los tres modelos:

Modelo sin Índice	Modelo - Valle	Modelo - Pico
0.1296	0.06769	0.1299

Para los tres modelos la significación probabilística de la prueba nos indica una probabilidad mayor de 0.05 (p-value) por lo que se puede concluir que los residuos se distribuyen normalmalmente.

De los resultados del análisis de los residuales se tiene:

- Los gráficos de probabilidad normal evidencian que los errores tienen una distribución aproximadamente normal.
- Se acepta la hipótesis nula de acuerdo al estadístico de Shapiro-Wilk, lo que evidencia que los residuos siguen una distribución normal.

### 5.2.6. PERFIL DE CLIENTES

Se entiende por perfilamiento de clientes a la segmentación de éstos en función del score que hayan alcanzado. Para identificar a cada perfil se utilizará como criterio la morosidad del portafolio a diciembre. La importancia del perfilamiento de clientes radica en el uso que se le puede dar como criterio para generar políticas en el negocio.

Como primer paso se segmentaron los clientes de acuerdo al puntaje (que hace referencia a la probabilidad de ser bueno) en 7 grupos (escalas de 100 para cada grupo) y se calculó la morosidad a diciembre para cada grupo. Finalmente con la serie de morosidad a diciembre para cada uno de los grupos se obtiene tres perfiles de clientes (figura 5.7) con calificaciones: A-, A, A+.

Los clientes con calificación “A-”, corresponden a los clientes con puntajes de 0 hasta 500 puntos, mientras que los clientes cuya calificación asignada fue de “A” los puntajes se encuentran en un rango de entre 500 y 800. Finalmente los clientes con mejor calificación de “A+”.

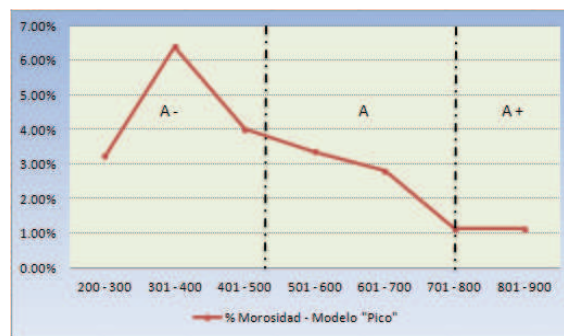


Figura 5.7: Perfilamiento de clientes en función de la tasa de morosidad

### 5.2.7. CONSISTENCIA DE NEGOCIO DEL MODELO

#### Morosidad

Para verificar que el modelo asigna “mejores” puntajes a los clientes se realizó un análisis comparativo de morosidad con el modelo planteado y el modelo de la institución financiera; obteniendo los siguientes resultados:

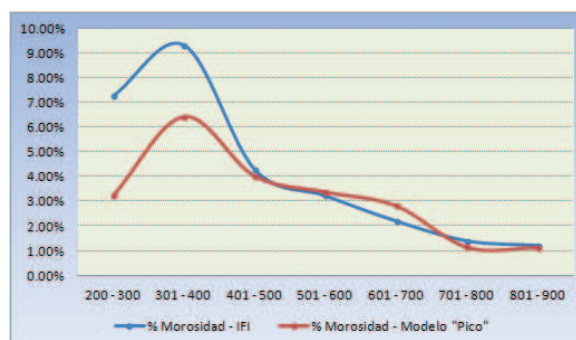


Figura 5.8: Modelo “Pico”: % Morosidad - puntaje score

En la figura 5.8 se observa que la tasa de morosidad tanto de la IFI como del modelo “pico” tienen un comportamiento descendente; un resultado esperado ya que los clientes mayor puntuados tienen baja probabilidad de incumplimiento y por ende una tasa de mora pequeña. Así, por ejemplo para el último grupo (clientes con puntaje entre 801 y 900) la tasa de morosidad alcanza aproximadamente el 1.13 %.

De manera similar, en la figura 5.9 se tiene la tasa de morosidad de la IFI comparada

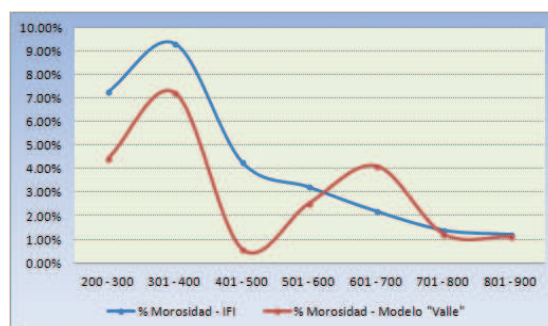


Figura 5.9: Modelo “Valle”: % Morosidad - puntaje score

con la del modelo “valle” en la que se observa menores tasas de morosidad para el caso de los puntajes del este último modelo, con excepción del rango (601 - 700).

Entonces, se visualiza que tanto en el modelo “pico” como en el modelo de “valle” se discrimina mejor a los clientes, otorgando un puntaje más acertado generando como resultado tasas de morosidad más bajas. En el modelo “pico” (ver figura 5.8) únicamente en el rango 601 – 700 la tasa de morosidad es mayor comparada con la de la IFI; lo mismo ocurre para el modelo “valle”, que a pesar de tener tasas de morosidad más altas respecto al modelo anterior, éstas siguen siendo más bajas respecto a las de la institución financiera.

### Análisis temporal con el ciclo de la Economía

Otra forma, para verificar si el modelo es consistente es el análisis temporal de los puntajes asignados por nuestro score en función de la fecha de concesión de cada crédito. En primer lugar se realizó la determinación gráfica de ciclos <sup>7</sup>de la economía ecuatoriana en el periodo de análisis<sup>8</sup>.

<sup>7</sup>Los ciclos son un tipo de fluctuaciones que se presentan en la actividad económica global de las naciones cuyo sistema productivo descansa principalmente en la empresa privada. Un ciclo consta de expansiones, que se producen aproximadamente, al mismo tiempo en muchas ramas de la actividad económica, y que son seguidas de regresiones y recuperaciones, que conducen a la fase de expansión del ciclo siguiente. Burns y Mitchel(1946)

<sup>8</sup>El Centro de Estudios Fiscales, en su Estudio Hechos Estilizados de la Economía Ecuatoriana: El Ciclo económico 1965-2008, determinó que existe una ciclo económico en el periodo comprendido entre IV 2005-2008 II. Ver [https://cef.sri.gob.ec/virtualcef/file.php/43/Estudios\\_Fiscales/Roja\\_NT\\_2010-02.pdf](https://cef.sri.gob.ec/virtualcef/file.php/43/Estudios_Fiscales/Roja_NT_2010-02.pdf)

Como se puede evidenciar en la figura 5.10 se determinaron tres fases del ciclo corto, las fases uno y tres se consideran fases de auge ya que las tasas de crecimiento anuales son superiores a 4 %, mientras que la fase dos se consideró como fase de contracción dado que presenta las tasas de crecimiento más bajas del periodo de análisis.

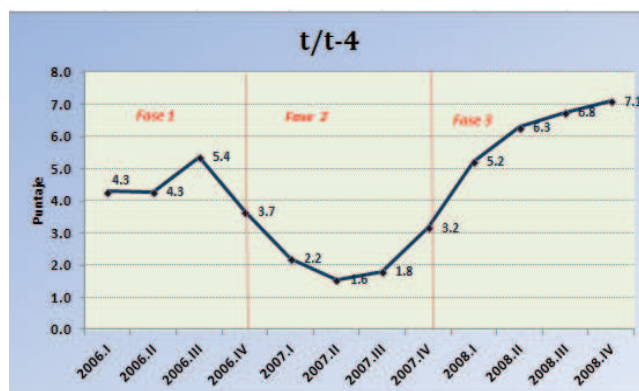


Figura 5.10: Evolución PIB trimestral periodo 2006.I-2008.IV. Fuente: Banco Central del Ecuador

En segundo lugar se realizó una comparación de los puntajes obtenidos por los modelos desarrollados y el modelo interno de la IFI, a nivel de grupos y en base a las fases que se determinaron anteriormente. La figura 5.11, muestra que los puntajes promedios obtenidos por el modelo de pico, en una fase de auge, son mayores que los puntajes asignados por la IFI, tanto para el grupo 1, como para el grupo 2 y menores para el grupo que tres, lo que evidencia que el modelo de pico recoge efectivamente el comportamiento de la economía en dicho periodo.

De manera similar realizando el mismo análisis para el ciclo dos (contracción), se puede observar en la figura 5.12, que los puntajes de los modelos de “pico” y “valle”, asignan puntajes menores a los del modelo interno de la IFI, dado que la probabilidad de incumplimiento aumenta, lo que permite nuevamente decir que estos dos modelos también recogen el comportamiento de la economía en ese periodo.

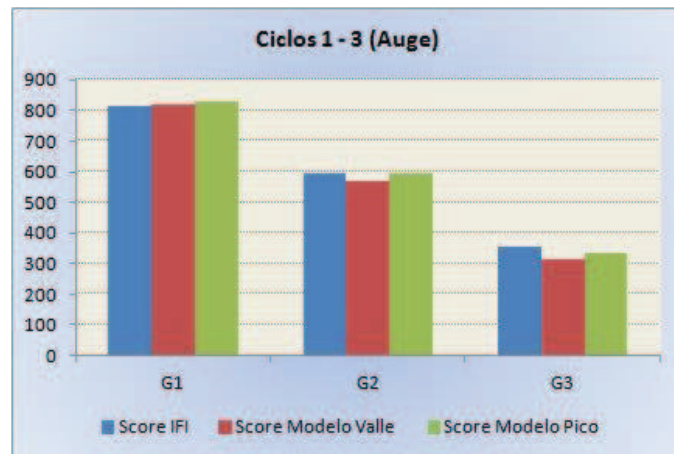


Figura 5.11: Consistencia de Negocio del Modelo - Análisis Fase de Auge

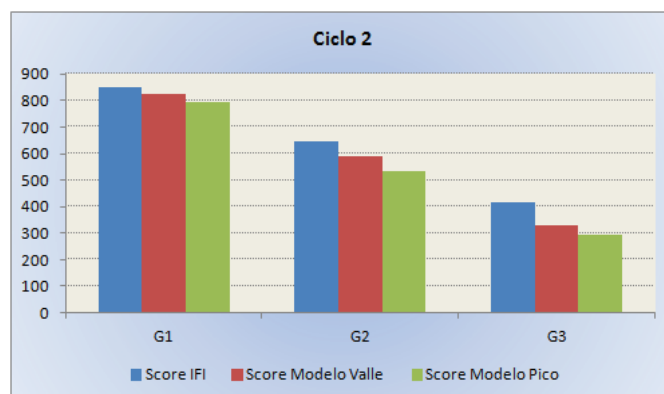


Figura 5.12: Consistencia de Negocio del Modelo -Análisis Fase de Contracción

## Capítulo 6

# CONCLUSIONES Y RECOMENDACIONES

- En los últimos años las diferentes instituciones han buscado nuevas técnicas orientadas a mitigar el riesgo de crédito. Entre ellas, los modelos lineales generalizados mixtos, dichos modelos constituyen una herramienta de gestión de riesgo de crédito, ya que permite calcular de manera adecuada la probabilidad de incumplimiento para monitorear el riesgo implícito en los créditos otorgados; además de capturar la dinámica de la economía, a través de la incorporación de factores de riesgo sistémicos que afectan el comportamiento de pago del conjunto de deudores. Los factores de riesgo se pueden modelar de manera separada del modelo de cálculo de probabilidad individual; empleando para ello, factores sistémicos como variables macroeconómicas aportando al modelo cierta variabilidad que no recogen los modelos tradicionales.
- Dentro de la metodología planteada, para la elaboración del scoring una parte crucial fue definir grupos homogéneos, (para un mejor análisis de la dependencia de los créditos) se usó la metodología CHAID, que permitió identificar divisiones óptimas a partir de las variables seleccionadas mediante clasificadores bayesianos. El uso de clasificadores bayesianos permitió seleccionar un total de 20 variables a partir de 101. Así, la selección de variables a ser usadas en el árbol, mediante



esta técnica, permitió realizar un análisis en conjunto para conocer el grado de información con el que pueden contribuir cada una de las variables en el para discriminar y predecir clientes buenos y malos.

- El uso de esta nueva técnica de clasificadores bayesianos a más de permitir eliminar algunas de las variables que no influyeron en la definición de la variable dependiente (buenos-malo), que como mencionamos en el párrafo anterior es de vital importancia, considerando que el número de variables con las que contaba el portafolio de la institución financiera fue extenso, permite obtener topología gráfica que es particularmente relevante en los modelos de scoring ya que por un lado al describir las relaciones entre las diferentes variables permite también determinar el comportamiento de los prestatarios, mientras que por otro se podría utilizar como un modelo de predicción de riesgo tomando en cuenta que el objetivo en un credit scoring es predecir el riesgo en lugar de explicarlo.
- Tradicionalmente en las muestras se intenta usar un 50 % tanto de clientes buenos como 50 % malos, sin embargo en la práctica la muestra casi siempre no es de proporciones iguales. El aporte de nuestro trabajo es la utilización de la muestra tal cual se seleccionó, es decir, se demostró que los clasificadores bayesianos permiten seleccionar las variables predictoras que mejor expliquen la variable buenos y malos aún cuando la muestra sea desproporcionada.
- Los resultados indican la sensibilidad del riesgo de crédito ante cambios en algunas variables macroeconómicas y sus efectos sobre el comportamiento de pago de los clientes; es decir, las probabilidades de incumplimiento se incrementa o disminuyen dependiendo del estado en el que se encuentre la economía, en el momento de la concesión del crédito.
- Las herramientas utilizadas en la presente investigación, pueden ser una alternativa de modelo interno, considerando las nuevas regulaciones que han venido implementando los organismos reguladores en el sistema financiero ecuatoriano; ya que éstas permiten estimar el riesgo de crédito y obtener un perfil de riesgo adecuado de los clientes, tomando en cuenta no solo características propias del

cliente sino de su entorno, lo que conlleva a la institución a determinar políticas de riesgo propias a las características de cada perfil.

## Capítulo 7

# ANEXOS

### 7.1. SELECCIÓN DE LA MUESTRA DE CLIENTES BUENOS

Como es usual en el campo del muestreo, el tamaño de la muestra es un aspecto importante al momento de extraer las observaciones a partir de la población, puesto que ésta no es más que el número de observaciones necesarias para que los datos obtenidos sean representativos de la población. Es así que el tamaño  $n$  de la muestra se calcula mediante la siguiente expresión:

$$n = \frac{Np(1-p)}{(N-1)\frac{B^2}{z^2\alpha/2} + p(1-p)} \quad (7.1)$$

donde,  $N$  es el tamaño de la población (de buenos),  $B$  es el error,  $\alpha$  el nivel de significancia y  $p$  es la proporción de malos dentro del total de clientes. El tamaño de la muestra calculado con (7.1) es  $n=2913$  que corresponde a la decisión de trabajar con  $B=1\%$  y  $\alpha=5\%$ . Por consiguiente, el número de registros con el que se realizará el modelo discriminante es 4536; 2913 clientes buenos y 1623 clientes malos. La forma en la cual se extraerá la muestra será a través de muestreo sistemático. Este método de selección suele ser más preciso que el aleatorio simple, ya que recorre la población de un modo más uniforme logrando como resultado que la muestra conserve la misma distribución que la población.

## 7.2. ÍNDICE ADELANTADO DE LA ECONOMÍA ECUATORIANA (IACCE)

### 7.2.1. INTRODUCCIÓN

Aunque las instituciones financieras han encontrado dificultades en el transcurso de los años por muchas razones, la causa principal de los problemas bancarios serios sigue directamente relacionada con normas débiles de crédito para prestatarios y contrapartes, una débil administración del riesgo de cartera o una falta de atención a los cambios en las circunstancias económicas u otras que podrían causar el deterioro en el crédito de las contrapartes del banco [58].

Esta experiencia ha llevado a que la gestión de riesgo de crédito, sea objeto de una atención cada vez mayor del regulador y del enfoque estratégico en muchas instituciones financieras. Por ello, recientemente se publicó el nuevo Acuerdo de Basilea III (BIII), cuyos objetivos buscan en primer lugar incrementar la capacidad del sector bancario para absorber choques, en segundo lugar mejorar la gestión de riesgos y el gobierno corporativo y finalmente, introducir elementos macroprudenciales en la regulación que ayuden a contener los riesgos sistémicos sobre la economía [59].

El método cuantitativo más frecuente es el credit scoring, que permite asignar una calificación de riesgo, una puntuación, a un demandante de crédito (cliente real o cliente potencial); y que ayuda a los prestamistas a tomar decisiones vinculadas con el otorgamiento de crédito (principalmente, de consumo). Es decir, determina con una base matemática, quién es sujeto de crédito, cuánto dinero se le otorgará y bajo qué condiciones [60].

Sin embargo, para estar acorde a la normativa internacional es necesario incorporar en esta metodología, condiciones económicas para incrementar el poder explicativo del modelo y a la vez permitir capturar los efectos de la dinámica de la economía que podrían impactar en la capacidad financiera del cliente para cumplir con sus obligaciones.

Esto, de acuerdo a resultados de algunos trabajos en los que sus autores han evaluado la sensibilidad del riesgo de crédito ante cambios en algunas variables macroeconómicas y sus efectos sobre el comportamiento de pago de los clientes.

Es así, que Hardy y Pazarbasioglu (1999), examinaron episodios de crisis bancarias en una muestra amplia de países e identificaron qué variables macroeconómicas y financieras pueden servir como indicadores de alerta temprana. Los resultados, indican que los problemas que enfrentan los bancos están asociados con caídas contemporáneas del crecimiento del PIB, periodos de mayor inflación, expansión del crédito agregado, mayores flujos de capital, el incremento de las tasas de interés reales, caída del tipo de cambio real y choques adversos en los términos de intercambio.

Asimismo, Hoggarth et al. (2005) en su trabajo realizado muestran un nuevo enfoque de las pruebas de estrés en el sistema bancario del Reino Unido, y dan cuenta de la dinámica entre las pérdidas y ganancias de los bancos y las principales variables macroeconómicas. Los resultados muestran que dichos bancos, son sensibles ante una desaceleración de la actividad económica; es decir existe una relación negativa, significativa y duradera entre los cambios de la actividad económica y el indicador de mora (IM) [61].

De manera similar, Alvés (2004) realiza un análisis de cointegración (VEC) entre las variables macroeconómicas y la probabilidad de incumplimiento esperado (EDF) del sector corporativo de la Unión Europea<sup>1</sup>. Los resultados evidencian que las variables macroeconómicas contribuyen a determinar los valores de estado estacionario de la EDF y a definir el perfil de riesgo de los deudores [62].

Otro trabajo, en esta misma línea, es el de Amaya (2005) que estima la relación de largo plazo entre la actividad económica, las tasas de interés, los precios de la vivienda y los indicadores de mora de las carteras hipotecaria y de consumo, para el periodo comprendido entre 1994 y 2005. Los resultados indican que la actividad económica tiene un efecto significativo sobre los niveles de largo plazo de los indicadores

---

<sup>1</sup>La EDF es estimada por medio del modelo KMV de riesgo de crédito de Moody's.

de mora de las dos carteras. Resaltando de esta manera la importancia de avanzar en la regulación del riesgo de crédito y de provisiones anticíclicas [63].

Por su parte, Wong et al. (2005) en su estudio desarrolló un marco de pruebas de stress para el portafolio de créditos de los bancos en Hong Kong. Se realizaron dos modelos de riesgo de crédito, cada uno comprende un modelo de regresión múltiple que explica la probabilidad de incumplimiento y la exposición del riesgo de cada banco al ambiente macroeconómico. El análisis de los resultados muestra una relación significativa entre las tasas de incumplimiento y los factores macroeconómicos, entre los cuales se encuentra el PIB real de Hong Kong, las tasas de interés, el precio de la vivienda y el PIB real de China continental [64].

De esta manera, se concluye que la actividad crediticia tiene un claro comportamiento procíclico. Así, en la fase expansiva del ciclo económico, la cartera tiende a expandirse de forma acelerada contribuyendo a profundizar la propia expansión de la economía. Por otra parte, la contracción de los préstamos que suele observarse durante la fase descendente del ciclo, tiende a acentuar la desaceleración económica haciendo más agudo el ciclo.

Entonces, la probabilidad de pago está estrechamente relacionada con el ciclo económico en general. Es decir, en la fase alta del ciclo las familias y las empresas, sin restricciones de acceso al crédito y con ingresos y ventas crecientes, tienen menores dificultades para hacer frente a sus obligaciones financieras. Por ello, los niveles de morosidad del crédito bancario tienden a reducirse. Por el contrario, en las fases recesivas, caracterizadas por la contracción del crédito y por el retroceso de las ventas empresariales y los ingresos familiares —por la caída del salario y el empleo—, la morosidad suele aumentar [65].

Sin embargo, analizar las variables individualmente puede no ser la mejor forma de estimar el estado general de la economía, debido a la información parcial involucrada en cada una de las series. De esta forma el monitoreo de la actividad económica se debe hacer a partir de indicadores económicos que condensen la mayor información posible para proveer señales acerca de los ciclos económicos. Es así como la construcción

de índices coincidentes y líderes ha jugado un papel importante en la descripción y pronóstico del estado de la actividad macroeconómica [66].

### **7.3. INDICADOR ADELANTADO DEL CICLO ECONÓMICO PARA EL ECUADOR**

Una gran gama de indicadores líderes de actividad económica han sido contruidos con el propósito de que anticipen el comportamiento del ciclo económico.

Una forma utilizada para calcular índices compuestos de variables indicadoras es el uso de modelos de probabilidad explícitos, en los cuales el índice coincidente corresponde a una variable no observable que representa el estado general de la economía y permite identificar aquello que se busca predecir con el índice líder [66].

Los indicadores líderes (o adelantados) poseen una historia dentro de la teoría económica de aproximadamente 60 años, desde que pioneros como Mitchell y Burns (1938), y Shiskin y Moore (1967) empezaron a analizar el comportamiento adelantado de algunas series respecto de la pauta que marcaba el ciclo económico. Desde entonces, varias revisiones se han llevado a cabo y nuevas metodologías han surgido, pero las ideas base siguen siendo las mismas. Las economías de mercado suelen experimentar repetidas pero no periódicas fluctuaciones en su nivel de actividad, y existe un conjunto de variables que reacciona de forma anticipada a este comportamiento general [67].

#### **7.3.1. ASPECTOS METODOLÓGICOS PRELIMINARES**

A continuación, se detalla la metodología utilizada para seleccionar las variables y construir el indicador adelantado del ciclo económico.

#### **VARIABLE DE REFERENCIA**

El primer paso consiste en elegir la variable de referencia. La principal característica que debe cumplir una serie de referencia es que indique adecuadamente el

comportamiento de la variable que se desea explicar. Es importante la elección de la variable de referencia ya que la utilidad del índice que se construya dependerá en alguna medida de la correcta definición de ésta.[68].

### SELECCIÓN DE VARIABLES [67]

Como el objetivo es que el índice recoja el comportamiento de la economía en su conjunto, se deben utilizar series de diversos sectores, algunas de ellas que puedan indicar que el nivel de actividad económica se encuentra próximo a aumentar mientras que otras indiquen una disminución del mismo. Por esta razón, es necesario utilizar un número suficiente de variables, de suerte tal que si alguna de ellas indicase una trayectoria equivocada de la variable de objetivo, no afecte de manera importante al indicador.

Del mismo modo, es necesario tomar en consideración el grado de relación entre las variables escogidas. No pueden incluirse variables que presenten comportamientos muy semejantes; puesto que un movimiento conjunto errado, podría desviar al índice de su trayectoria original.

De esta manera, la selección de las variables antecedentes que formen parte del índice debe ser lo más rigurosa que se pueda, a fin de predecir el comportamiento de la economía con la mayor exactitud posible. En este sentido, la elaboración de los puntos de giro y tendencia ofrecen mayor información respecto de la relación entre las variables adelantadas y la variable objetivo.

Así, el número de puntos de giro (o de inflexión) correctamente predichos y el grado de asociación con la tendencia de la variable objetivo permiten encontrar aquellas series (adelantadas) con un mejor desempeño individual en la predicción de las fluctuaciones, picos y valles de la variable objetivo.

Después de haber determinado la estabilidad y significancia del grado de asociación de las series adelantadas y el signo de correlación de cada una, se eligen aquéllas que



presenten una mejor performance en lo que respecta a la tendencia y los puntos de giro, que proporcionen la mayor cantidad de meses de anticipación, y que no repitan información dentro del indicador.

## METODOLOGÍA

**Filtro de Hodrick y Prescott [69]** Sea  $y_t$  una serie de tiempo para  $t = 1, 2, \dots, T$ . Si  $\tau_t$  es la tendencia de esta serie, entonces la medida de las fluctuaciones cíclicas está dada por:  $c_t = y_t - \tau_t$ .

Hodrick y Prescott proponen que el componente tendencia de una serie es el que minimiza la siguiente ecuación:

$$\sum_{t=1}^T (y_t - \tau_t)^2 + \sum_{t=2}^{T-1} [(\tau_{t+1} + v_t) - (\tau_t - \tau_{t-1})]^2 \quad (7.2)$$

O bien, puede resumirse de la siguiente forma:

$$\min \sum_{t=1}^T C_{t^2} + \lambda \sum_{t=3}^T (\Delta^2 \tau_t)^2 \quad (7.3)$$

donde:

$$\Delta^2 = (1 - L)^2, \text{ con } L \text{ operador de rezagos}$$

El primer término de las ecuaciones (7.2) y (7.3) anteriores es la suma de las desviaciones de la serie respecto a la tendencia al cuadrado, y es una medida del grado de ajuste. El segundo término es la suma de cuadrados de las segundas diferencias de los componentes de tendencia y es una medida del grado de suavidad. Este modelo permite que el componente de tendencia de  $y_t$  cambie suavemente a lo largo del tiempo.

Escribiendo las ecuaciones anteriores en notación matricial se tiene:

$$\text{mín } C' C + \lambda (A \tau)' (A \tau) \quad (7.4)$$

donde el desarrollo matricial de  $A \tau$  es:

$$A \tau = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & . & . & . & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & . & . & . & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & . & . & . & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & . & . & . & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ . \\ \tau_T \end{bmatrix}$$

La condición de primer orden de (7.4), que consiste en igualar a cero la primera derivada y despejar  $\tau$ , brinda como resultado:

$$\tau = (I + \lambda A' A)^{-1} Y \quad (7.5)$$

Luego, de (7.5), el componente cíclico  $C_t$  puede calcularse como:

$$\begin{aligned} C &= Y - \tau = Y - (I + \lambda A' A)^{-1} Y \\ C &= \left[ I - (I + \lambda A' A)^{-1} \right] \end{aligned} \quad (7.6)$$

Es importante mencionar que la serie  $y_t$  se emplea generalmente en logaritmos para que el componente  $C_t$  quede expresado directamente en términos de desviaciones porcentuales de la serie respecto a la tendencia, que se consideran como una estimación del ciclo de la serie.

La cantidad  $\lambda$  es el parámetro de suavidad con el cual se controla la aceleración en el componente de tendencia, es decir, las variaciones en la tasa de crecimiento del componente de tendencia.  $\lambda$  debe ser positiva para que la segunda derivada sea positiva y se garantice así que se obtuvo un mínimo.

**Método de Neftci [70]** El uso de algún indicador adelantado  $I^A$ , puede ser de gran ayuda, puesto que éste presenta sus puntos de giro, picos (máximos relativos) y valles (mínimos relativos) antes que el nivel general de actividad económica.

El objetivo es que el indicador sea capaz de anticipar los puntos de giro en los ciclos de crecimiento económico, tratando de calcular la probabilidad de que se presente un punto de giro. En concreto se calcula las probabilidades secuenciales recursivas mensuales de pico y valle, a partir de la función de densidad estimada con los datos observados hasta el momento  $t$  del indicador adelantado  $I_t^A = (i_1^A, i_2^A, \dots, i_{t-1}^A, i_t^A)$ . Neftci (1982).

El análisis teórico usado por Neftci (1982) consiste en observar un proceso, cuya estructura cambia de forma aleatoria en algún momento del tiempo, y predecir el instante en el que se presenta el cambio súbito. Interesa establecer el cambio de régimen del indicador adelantado para predecir los cambios en la actividad económica real. Así, las observaciones de  $I_t^A$  serán usadas para realizar la inferencia que permita predecir el momento en que la economía cambia de fase. El indicador adelantado se convierte en un predictor de los puntos de giro que la actividad económica tendrá en  $t + h : y_{t+h}^G$ , siendo  $G = P, V$  según el punto de giro se refiera a un pico o a un valle respectivamente. La predicción en  $t$  (asociada al indicador adelantado  $I_t^A$ ), la denotaremos como:  $\hat{y}_{t+h,t}^G$ .

Se supone que la actividad económica real ( $Y$ ) en  $t + h$  tiene dos distribuciones distintas correspondientes a las distribuciones en el régimen de expansión y de contracción que denotamos por  $f_{t+h}^u(Y)$  y  $f_{t+h}^d(Y)$  respectivamente. Dichas distribuciones se estiman a partir de las observaciones pasadas de  $Y_t$ , que denominaremos respectivamente  $f_t^u(Y)$  y  $f_t^d(Y)$ . Un punto de giro en la economía se define como el momento del tiempo en el que cambia la distribución de probabilidad de  $Y_{t+h}$  y el objetivo es predecir los puntos de giro de  $I_t^A$  con la idea de que los hechos que producen un cambio están presentes en esa serie antes que en  $Y_{t+h}$ . Denotaremos por  $Z_t$  al número entero que representa el primer periodo posterior a un punto de giro (pico) del indicador adelantado. Entonces el indicador adelantado  $I_t^A$  tiene dos distribuciones correspondientes a recuperaciones y contracciones que denotamos respectivamente por  $f_t^u(I_t^A)$  y  $f_t^d(I_t^A)$ .

y se definen como sigue:

$$\begin{aligned} f_t^u(I_t^A) \text{ cuando } t < Z_{|} \\ y f_t^d(I_t^A) \text{ cuando } t \geq Z_{|} \end{aligned} \quad (7.7)$$

pudiendo ser estimadas a partir de las observaciones pasadas en el indicador  $(I_{t_1}^A)$ .

Conocida la última información disponible hasta  $t$  del indicador adelantado,  $I_{t_1}^A$ , la probabilidad de observar un pico en  $t$ , correspondiente a un anticipo del pico en la actividad económica futura  $\hat{y}_{t+h,t}^P$ ,  $Z_I \leq t$ , puede ser expresada como:

$$P_t = P(Z_{|} \leq t \mid I_{t-1}^A) = \frac{P(\hat{y}_{t+h,t}^P \mid Z_{|} \leq t) P(Z_{|} \leq t)}{P(I_{t-1}^A)} \quad (7.8)$$

que es la probabilidad a posteriori de un punto de giro dados los datos disponibles del indicador adelantado hasta el momento  $t$ .

Es posible obtener una fórmula recursiva de esta probabilidad a posteriori que para el caso de un pico viene dada por:

$$P_t = \frac{[(P_{t-1} + \Gamma_t^u (1 - P_{t-1}))] f_t^d(\hat{y}_{t+h,t}^P \mid I_{t-1}^A)}{[(P_{t-1} + \Gamma_t^u (1 - P_{t-1}))] f_t^d(\hat{y}_{t+h,t}^P \mid I_{t-1}^A) + (1 - P_{t-1}) f_t^u(\hat{y}_{t+h,t}^P \mid I_{t-1}^A) (1 - \Gamma_t^u)} \quad (7.9)$$

mientras que el cálculo de la probabilidad secuencial recursiva en los valles se llevará a cabo mediante la expresión:

$$V_t = \frac{[(V_{t-1} + \Gamma_t^d (1 - V_{t-1}))] f_t^u(\hat{y}_{t+h,t}^V \mid I_{t-1}^A)}{[(V_{t-1} + \Gamma_t^d (1 - V_{t-1}))] f_t^u(\hat{y}_{t+h,t}^V \mid I_{t-1}^A) + (1 - V_{t-1}) f_t^d(\hat{y}_{t+h,t}^V \mid I_{t-1}^A) (1 - \Gamma_t^d)} \quad (7.10)$$

donde  $\Gamma_t^u = P(Z_{|} = t \mid Z_{|} > t)$  y  $\Gamma_t^d = P(Z_{|} = t \mid Z_{|} \leq t)$  son las probabilidades de que un punto de giro se presente en el momento  $t$  sabiendo que no se presentó

previamente, denominadas probabilidades de transición de un régimen de expansión a uno de contracción en el primer caso y de una contracción a una expansión en el segundo,  $f_t^u(\hat{y}_{t+h,t}^P | I_{t-1}^A)$  y  $f_t^d(\hat{y}_{t+h,t}^V | I_{t-1}^A)$  son las densidades de probabilidad del indicador adelantado según la última observación provenga de una situación de recuperación o de contracción y  $P_{t-1}$  y  $V_{t-1}$  son las probabilidades calculadas para la observación anterior del indicador adelantado.

Cuando nos encontramos en un momento perteneciente a la fase de crecimiento en la economía, se calcula la probabilidad secuencial recursiva de un pico (fin de la expansión)  $P_t$  y cuando supera un valor crítico establecido previamente concluimos que se presenta un pico en el momento  $t = Z|$ , que se puede relacionar con un inminente punto de giro (en  $t + h$ ) en el crecimiento económico:  $y_{t+h}^P$ . Del mismo modo, cuando estemos en un periodo de decrecimiento se calcula la probabilidad secuencial recursiva de un valle (inicio de la expansión).

Cuando la probabilidad estimada es mayor que una probabilidad crítica fijada por el analista, se declara una señal de pico o valle, según la información contenida en la probabilidad secuencial recursiva. Cada vez que la probabilidad estimada supera la probabilidad crítica y no acontece el punto de giro, se produce una falsa señal.

Las probabilidades  $V_t$  o  $P_t$  se reajustan de nuevo a cero cuando se registra un punto de giro en la serie para la cual realizamos la predicción ( $Y$ ), buscando así el punto de giro siguiente.

### 7.3.2. ELABORACIÓN DEL ÍNDICE MENSUAL ADELANTADO DEL CICLO ECONÓMICO PARA EL ECUADOR

El IACCE es un indicador que pretende detectar y pronosticar las inflexiones de la economía de tal manera que se puedan anticipar recesiones o etapas de crecimiento con un trimestre de adelanto, mediante el método de detección de puntos de inflexión de Neftci.

Es un índice probabilístico mensual que resulta de la agregación de series macroeconómicas que han presentado un comportamiento afín y anticipado con relación al ciclo económico del PIB del país, pero que tienen una mayor periodicidad y oportunidad.

La elaboración del indicador IACCE requiere el análisis de una amplia variedad de series. La importancia de la base de datos construida no sólo está asociada a su dimensión sino también al esfuerzo desplegado para cubrir una amplia gama de sectores que abarcan todos los ámbitos de la economía. La principal fuente de la base de datos es el Banco Central del Ecuador (BCE).

Adicionalmente, con el fin de estabilizar la varianza de las series, se aplicaron logaritmos a las series originales siempre que fuese posible. Otra consideración necesaria concierne a la disponibilidad de las series. Elaborar un indicador líder cuyos componentes no estarán disponibles sino muchos meses después del momento requerido para su construcción resultaría inútil para el establecimiento de cualquier tipo de política. Teniendo esto en cuenta, se escogieron series que tuviesen, a lo sumo, un retraso de dos meses en la presentación de los datos.

El primer paso para la elaboración del índice consiste en elegir la variable de referencia, con la cual se compara el desempeño de las demás variables que potencialmente formarán parte del indicador adelantado. Para este caso, la variable de referencia es el Producto Interno Bruto (PIB).

Una vez elegida la variable de referencia, se recopilan series de tiempo que potencialmente se adelanten a ésta. Sin embargo, como las 96 variables disponibles (de enero 2004 a octubre 2011) se encontraban expresadas en distintas unidades y tenían diferentes tipos de estacionalidad, fueron sometidas a varias transformaciones con el objetivo de homogeneizarlas y depurarlas de cualquier ruido estadístico distorsionante.

Se utiliza la siguiente metodología, la cual en términos generales, consiste en aplicar transformaciones a cada una de las series:

1. Se deflactan las series nominales.

2. En caso de detectarse datos atípicos, se extraen y se interpola el dato faltante con los registros precedente y siguiente.
3. Se aplica el logaritmo natural a todas las series, a excepción de aquellas dadas en porcentajes, o cuyo rango de datos incluye valores negativos.
4. Se extrae el componente cíclico de las series a partir del filtro de Hodrick Prescott<sup>2</sup>, ya que permite ajustar una curva suave al conjunto de datos.

Las series escogidas fueron seleccionadas, debido a que son indicadores que reaccionan rápidamente a un cambio de fase, ya que miden expectativas de los agentes de la economía o porque causan movimientos en la actividad económica.

Adicionalmente, las variables seleccionadas deben poseer las siguientes propiedades:

1. **Mayor correlación adelantada:** Maximice la correlación adelantada con la variable de referencia, procurando que ésta sea lo más anticipada posible.
2. **Capacidad de anticipación de quiebres:** Anticipe los puntos de quiebre del ciclo económico de la variable de referencia, con el menor número de señales falsas.
3. **Disponibilidad:** se prioriza aquellas series mensuales que tengan un pequeño rezago de publicación, se mantengan sobre una base regular y no sufran demasiadas revisiones.

A fin de establecer si los componentes cíclicos de las series adelantan, coinciden o rezagan al ciclo económico del PIB (el cual fue utilizado como ciclo económico de referencia), se calcularon las correlaciones cruzadas entre cada una de las variables analizadas y el ciclo de referencia. Estableciéndose, si el cambio de fase de la variable respecto al ciclo de referencia: adelanta o rezaga

Con base a la información de correlaciones cruzadas y un análisis de componentes principales, del conjunto de variables disponibles (96 series), sólo 11 cumplieron con

---

<sup>2</sup>La descripción de este método fue explicada anteriormente

los requisitos de significancia y consistencia en la correlación y nivel de adelanto. Como condición adicional, se intentó que el comportamiento de cada variable escogida tuviese una interpretación económica.

Las 11 series económicas mensuales dentro de la economía ecuatoriana que presentan las características y cualidades mencionadas anteriormente, para conformar el indicador son:

1. Índice de Precios al Consumidor
2. Deuda Externa Total
3. Ahorro - Captaciones
4. Importaciones de Combustibles y lubricantes (FOB)
5. Exportaciones No Tradicionales: Otros industrializados
6. Crudo Napo - Petroecuador (USD FOB)
7. Precio Napo
8. Exportaciones No Tradicionales: Elaborados de banano
9. Índice de Alojamiento, agua, gas, electricidad y otros
10. Producción Petrolera: Privadas
11. Importaciones de Bienes de Capital Agrícola

Una vez establecidas las series del ciclo económico tanto del PIB real trimestral como de las series seleccionadas, se procedió a determinar los periodos de recesión del ciclo.

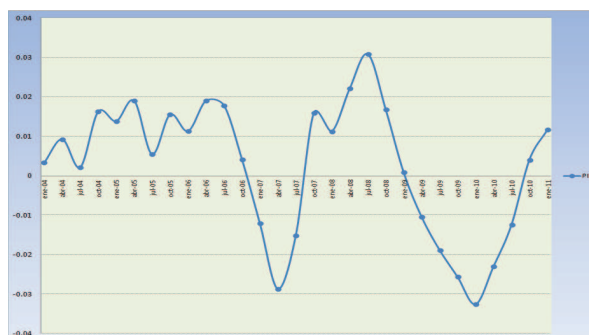


Figura 7.1: Períodos del Ciclo Económico - PIB

Durante el periodo analizado se produjeron 4 contracciones y 4 recuperaciones.



La contracción tuvo una duración de 7 trimestres en promedio y las recuperaciones persistieron en promedio por 6 trimestres.

Los puntos de giro establecidos en el ciclo de crecimiento del producto interno bruto servirán de base para la identificación de los puntos de giro de las series que adelanten al ciclo económico.

Para calcular el índice compuesto adelantado para el ciclo económico en el Ecuador, se utilizó el método de Neftci<sup>3</sup>. Este método calcula la probabilidad de que ocurra un punto de giro a partir de la clasificación de las observaciones del indicador de referencia según pertenezcan a una fase de aceleración o desaceleración; acumulando las probabilidades desde el comienzo de un punto de giro hasta el siguiente.

Una vez disponible toda la información, es posible estimar mensualmente las probabilidades de un punto de giro. Así, cuando se está en expansión se estima la probabilidad mensual de un pico correspondiente al fin de la expansión, mientras que en fases de decrecimiento se calcula la probabilidad de un valle.

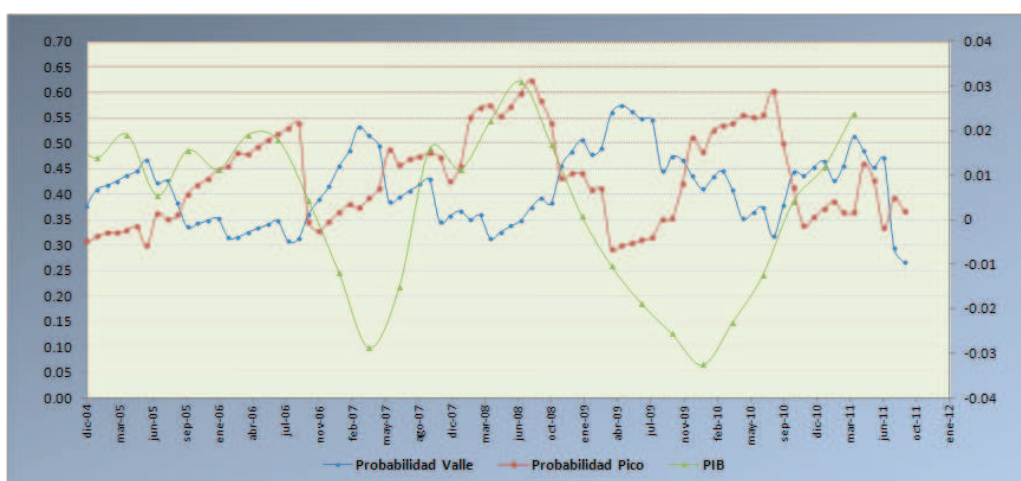


Figura 7.2: IACCE - Pico y Valle

<sup>3</sup>Método explicado anteriormente

#### 7.4. PROBABILIDADES DEL ÍNDICE ADELANTADO DE LA ECONOMÍA ECUATORIANA

Fecha	Probabilidad Valle	Probabilidad Pico
ene-06	0.35	0.45
feb-06	0.32	0.46
mar-06	0.32	0.48
abr-06	0.33	0.48
may-06	0.33	0.49
jun-06	0.34	0.51
jul-06	0.35	0.52
ago-06	0.31	0.53
sep-06	0.31	0.54
oct-06	0.36	0.35
nov-06	0.39	0.33
dic-06	0.41	0.35
ene-07	0.46	0.36
feb-07	0.49	0.38
mar-07	0.53	0.37
abr-07	0.51	0.39
may-07	0.50	0.41
jun-07	0.38	0.49
jul-07	0.39	0.46
ago-07	0.41	0.47
sep-07	0.42	0.47
oct-07	0.43	0.48
nov-07	0.35	0.47
dic-07	0.36	0.43
ene-08	0.37	0.45
feb-08	0.35	0.55
mar-08	0.36	0.57
abr-08	0.31	0.57
may-08	0.33	0.55

## 7.5. DESCRIPCIÓN DE LAS VARIABLES DISCRETAS A UTILIZAR EN EL MODELO

- **Rango:** Esta variable es una variable interna, que clasifica a los clientes por grupos a los que para facilitar la visualización hemos denominado grupos. Las categorías de esta variable son:

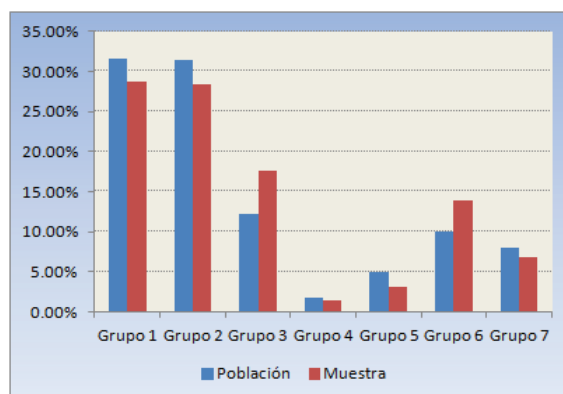


Figura 7.3: Descripción de Variable Rango.

- **Estado Civil:** Esta variable, está compuesta de las siguientes categorías:

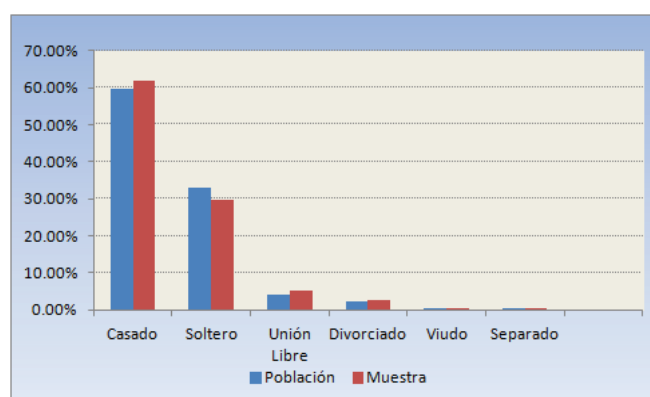


Figura 7.4: Descripción de la Variable: Estado Civil

- **Nivel de Educación:** Esta variable, está compuesta de las siguientes categorías:

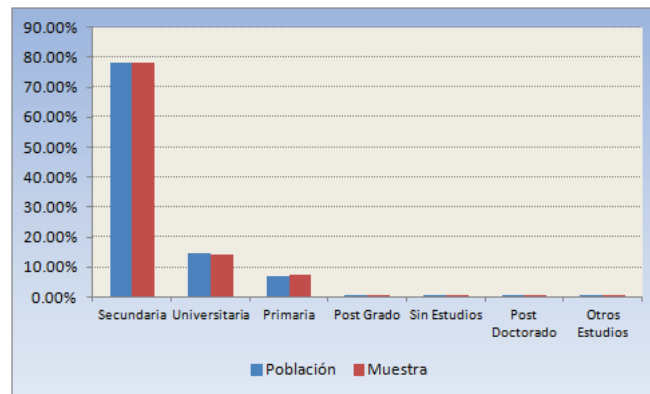


Figura 7.5: Descripción de la variable Nivel de Educación.

- **Tipo de Vivienda:** Esta variable, está compuesta de las siguientes categorías:

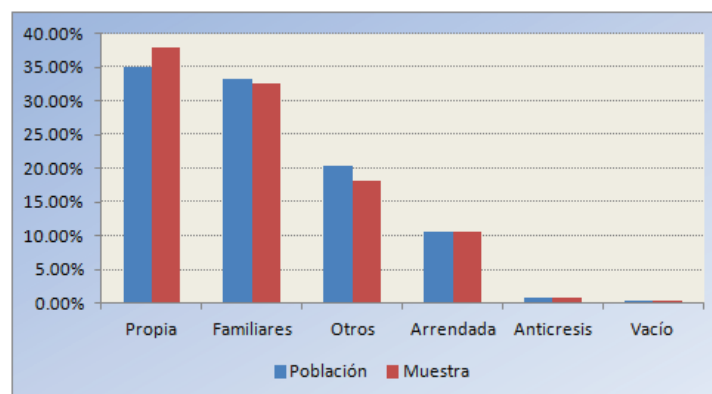


Figura 7.6: Descripción de la Variable Vivienda

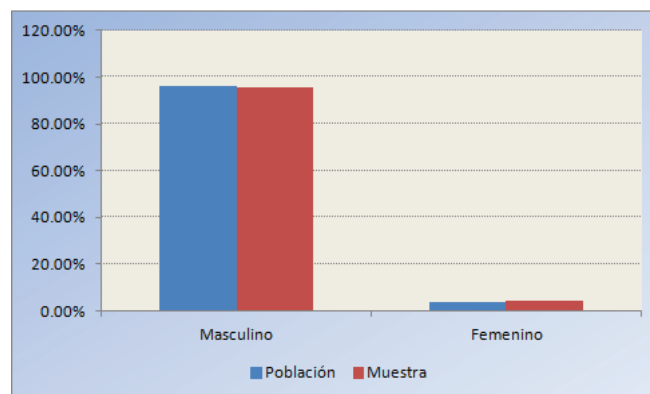


Figura 7.7: Descripción de la variable Género

- **Género:** Esta variable, está compuesta de las siguientes categorías:
- **Ciudad de Residencia del Cliente:**

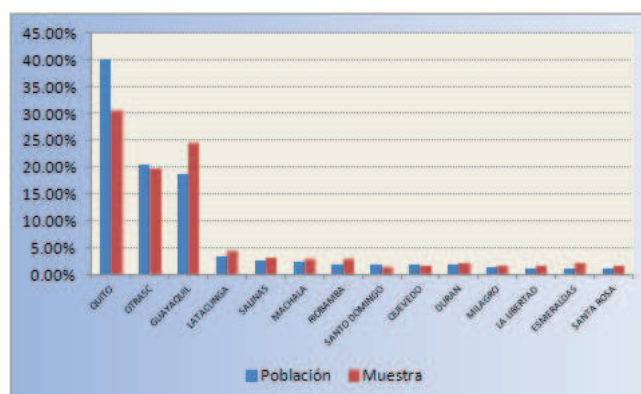


Figura 7.8: Descripción de la Variable Residencia del Cliente.

## 7.6. GRUPOS HOMOGÉNEOS OBTENIDOS MEDIANTE METODOLOGÍA CHAID

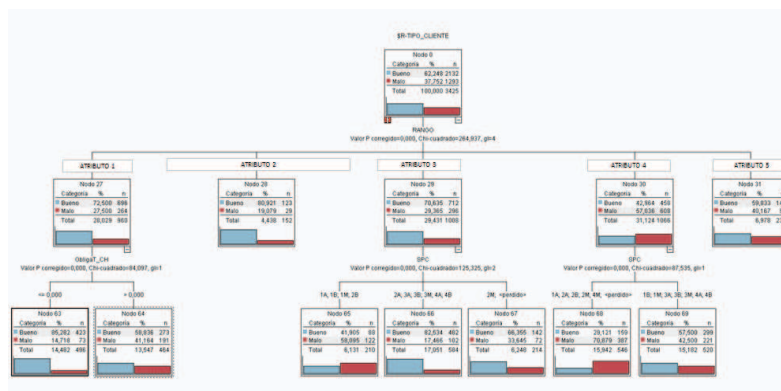


Figura 7.9: Selección de Grupos Homogéneos mediante metodología CHAID

# Bibliografía

- [1] **Blanco H. Gonzalo.** *El Sistema Financiero en México* (1996). Pág. 19.
- [2] **Dirección de Análisis y Medición de Riesgo.** *Ejemplo de Riesgo de Crédito* HSBC México (HBMX).
- [3] **Facultad de Ciencias Económicas y Empresariales,** *La medición del riesgo de crédito y el nuevo acuerdo de capital del Comité de Basilea.* Universidad de Santiago de Compostela.
- [4] **Márquez Diez Javier – Canedo.** *Una nueva visión del riesgo de crédito.*
- [5] **Maximiliano Silva Quiroz.** *Tesis: Métodos de Divergencia y Entropía para mejorar la predicción en Modelos de Clasificación.*
- [6] **Saunders Anthony, Allen Linda.** *Credit risk measurement: new approaches to value at risk and other paradigms.* Pág. 9.
- [7] **Alunan, Marco, y Varcetto (1994).**
- [8] **Arango Uribe Paula Andrea.** *Modelo de Cupos de Inversiones temporales para entidades del sector real a partir del análisis de riesgo de crédito y riesgo de mercado.* Pág. 21.
- [9] **Riesgo de Crédito.** *Gaceta de Basilea II.* HSBC México (HBMX). Dirección de Análisis y Medición de Riesgo. Pág. 3.
- [10] *La medición del riesgo de crédito y el nuevo acuerdo de capital del Comité de Basilea.* Universidad de Santiago de Compostela. Facultad de Ciencias Económicas y Empresariales. Pág. 7.

- 
- [11] *La medición del riesgo de crédito y el nuevo acuerdo de capital del Comité de Basilea*. Universidad de Santiago de Compostela. Facultad de Ciencias Económicas y Empresariales. Pág. 10.
- [12] **Saunders Anthony, Allen Linda**. *Credit risk measurement: new approaches to value at risk and other paradigms*. Pág. 42.
- [13] **Saavedra García María Luisa, Saavedra García Máximo Jorge**. *Modelos para medir el riesgo de crédito de la banca*. Pág. 307.
- [14] **Saavedra García María Luisa, Saavedra García Máximo Jorge**. *Modelos para medir el riesgo de crédito de la banca*. Pág. 307.
- [15] **Elizondo Alan, Altman Edward I**. *Medición integral del riesgo de crédito*. Pág. 184.
- [16] **McNeil Alexander J., Frey Rüdiger, Embrechts** “*Quantitative Risk Management: Concepts, Techniques and Tools*”. Princeton and Oxford.
- [17] **Baixauli J. Samuel**. *Riesgo de crédito*. Pág. 20.
- [18] **Saavedra García María Luisa, Saavedra García Máximo Jorge**. *Modelos para medir el riesgo de crédito de la banca*. Pág. 310.
- [19] **Saunders Anthony, Allen Linda**. *Credit risk measurement: new approaches to value at risk and other paradigms*. Pág. 81.
- [20] **De Lara, A. (2005)**. “*Medición y control de riesgos financieros*” 3. ED. México. Editorial Limusa S.A.
- [21] **Baixauli J. Samuel**. *Riesgo de crédito*. Pág. 18.
- [22] **Saunders Anthony, Allen Linda**. *Credit risk measurement: new approaches to value at risk and other paradigms*. Pág. 84.
- [23] **Elizondo Alan, Altman Edward I**. *Medición integral del riesgo de crédito*. Pág. 184.



- 
- [24] **Norman Fenton, Martin Neil** *Managing Risk in the Modern World*. Pág. 4 - Pág. 10.
- [25] **Miller Rivera Lozano**, *El papel de las redes bayesianas en la toma de decisiones*. Universidad del Rosario Pág,1
- [26] **Felix Doldan Tié**, *Riesgo Redes Bayesianas y Riesgo Operacional*.
- [27] **Samuel Gerssen**, *Bayesian Networks in Credit Rating*.
- [28] **D.J.Hand, K.J McConway**, and **E.Stanguellini**. Graphical Models of applicants for credit.
- [29] **Abbas Heiat**, *Modeling Consumer Credit Scoring Through Bayes Network*.
- [30] **D.J.Hand, W E. Henley**. *Statistical Methods in Consumer Credit Scoring: A review*.
- [31] **Miller Rivera Lozano**. *El Papel de las Redes Bayesianas en la Toma de Decisiones*. Febrero Universidad de Rosario. 2011.
- [32] **Gutiérrez García Javier, Elizondo Flores Jesús Alan**. “*Riesgo de crédito: El enfoque actuarial*”. México.
- [33] **Díaz María del Pilar**. “*Modelos Lineales Generalizados aplicados a las Ciencias Biológicas*”. Universidad Nacional de Córdoba. Noviembre 2011.
- [34] **Badiella Llorens**. “*Modelos lineales generalizados mixtos algunos casos prácticos*”. Universidad Autónoma de Barcelona.
- [35] **Puente Luis Nava, Surendra Prasad Sinha**. “*Ajuste e interpretación de modelos de Regresión logística con variables categóricas y continuas*”.
- [36] **Santolino Prieto Miguel Ángel**. “*Métodos econométricos para la valoración cualitativa y cuantitativa del daño corporal en el seguro del automóvil*”
- [37] **Fikret Isik, PhD**. “*Generalized Linear Mixed Models, An Introduction for Tree Breeders and Pathologists*”. Eugene, Oregon, USA.

- [38] **Bolker Benjamin M., Brooks Mollie E., Clark Connie J. , Geange Shane W. , Poulsen John R. , M. Henry H. Stevens and Jada-Simone S. White.** *“Generalized linear mixed models: a practical guide for ecology and evolution”*
- [39] **Katrien Antonio, Jan Beirlant.** *“Actuarial statistics with generalized linear mixed models”.*
- [40] **Arnost Kom´arek Emmanuel Lesaffre** *“Generalized linear mixed model with a penalized Gaussian mixture as a random effects distribution”*
- [41] **Hoda Rashidi Nejad ,Assieh Rashidi Nejad, Graduated student, Kerman, Iran.** *“Fit generalized linear mixed models (GLMM), using of different likelihoods”*
- [42] **Eugene Demidenko.** *“Mixed models: theory and applications”*
- [43] <http://academic.uprm.edu/rmacchia/agro6998/Modeloslinealesgeneralizados.pdf>
- [44] **Díaz Abel.** *“Diseño estadístico de experimentos”.* Editorial Universidad de Antioquia. Segunda Edición.
- [45] <http://www.xatakaciencia.com/matematicas/contraste-de-shapiro-wilk>
- [46] **Guisande González Cástor,** *“Tratamiento de Datos”,* España
- [47] **Pearl, J. (1988).** *Probabilistic Reasoning in Intelligence Systems.* Morgan Kaufmann, Los Altos CA.
- [48] **Heckerman, D. (1995).** *A tutorial on learning with bayesian networks. Technical Report MSR-TR-95-06,* Microsoft Research, U.S.A.
- [49] **Lauritzen and Spiegelhalter, 1988.** *Local computations with probabilities on graphical structures and their application to expert systems.*
- [50] [http://www.ats.ucla.edu/stat/mult\\_pkg/glmm.htm](http://www.ats.ucla.edu/stat/mult_pkg/glmm.htm)
- [51] **Friedman, N, Geiger, D, y Goldsmidt, M. (1997).** *Bayesian network classifiers. Machine Learning.*

- 
- [52] **Stone M.**, *Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society Series B.*
- [53] **Hanley, J.A. y McNeil, B.J. (1982).** *The meaning and use of the area under a receiver operating characteristic (ROC) curve.*
- [54] **Efron B.** *Estimating the error rate of a prediction rule: Improvement on cross-validation.* Journal of the American Statistical Association.
- [55] **Boada Ramos Alexis Omar,** *Metodología para medir la satisfacción de cliente, aplicación al servicio generado por una telefónica móvil.*
- [56] **Actuarios Trabajando,** *Revista Mexicana de Investigación Aplicada.* Colegio Nacional de Actuarios. 2011 Pág. 82 - Pág. 108
- [57] **Efron, B. y Tibshirani, R.J. (1993).** *An Introduction to the Bootstrap. Monograph on Statistics and Applied Probability.*
- [58] **Comisión de Basilea de Supervisión de Bancos.** “*Principios para la Administración del Riesgo de Crédito*” Julio 1999.
- [59] **Rodríguez Elena.** “*Las nuevas medidas de Basilea III en materia de Capital*”.
- [60] <http://www.rankia.com/blog/berbis-swap/380484-credit-scoring-como-toma-banco-decisiones-crediticias>.
- [61] **Gutiérrez Rueda Javier, Saade Ospina Agustín.** “*Ciclos del Riesgo de Crédito*”. Reporte de Estabilidad Financiera. Septiembre 2009.
- [62] **Alves, I.** “*Sectoral fragility: factors and dynamics*”. Bank for International Settlements. 2004
- [63] **Amaya Carlos Andrés.** “*Evaluación del Riesgo de Crédito en el Sistema Financiero Colombiano*” Banco de la República, Reporte de Estabilidad Financiera, Diciembre 2005.

- [64] **Wong J., Choi, K. and Fong, T.** “*A framework for macro stress testing the credit risk of banks in Hong Kong*”. Hong Kong Monetary Authority Quarterly Bulletin. Diciembre 2005.
- [65] **Díaz Oscar A, Quevedo.** “*Determinantes del ratio de morosidad en el sistema financiero boliviano*”.
- [66] **Melo Luis Fernando, Nieto Fabio, Posada Carlos Esteban, Betancourt Yanneth Rocío y Barón Juan David.** “*Un índice coincidente para la actividad económica colombiana*”. Diciembre 2011.
- [67] **Escobal Javier, Torres Javier.** “*Sistema de Indicadores Adelantados y Coincidentes del Nivel de Actividad para la Economía Peruana*”. Grupo de Análisis para el desarrollo (GRADE). Lima. Agosto 2000.
- [68] **Kikut Ana Cecilia, Muñoz Evelyn.** “*Diseño de un índice sintético adelantado para la inflación: El caso de Costa Rica*”. Banco Central de Costa Rica. Departamento de Investigaciones Económicas. Febrero 1999.
- [69] **Muñoz Evelyn, Kikut Ana Cecilia.** *El filtro de Hodrick y Prescott: “Una técnica para la extracción de la tendencia de una serie”*. Banco Central de Costa Rica. Departamento de Investigaciones Económicas.
- [70] **Moreno Cuartas Blanca, López Menéndez Ana Jesús.** “*Las Opiniones Empresariales como predictores de los puntos de giro del ciclo industrial*”. Departamento de Economía Aplicada. Universidad de Ovied. VO L . 25-1. Enero 2007.
- [71] **Balzarini Mónica, Macchiavelli Raúl E., Casanoves Fernando.** “*Aplicaciones de Modelos Mixtos en Agricultura y Forestería*”.