UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS Fundada en 1551
FACULTAD DE INGENIERÍA INDUSTRIAL
UNIDAD DE POST GRADO

Reconocimiento de patrones de morosidad para un producto crediticio usando la técnica de árbol de clasificación CART

TESIS para optar el Grado Académico de: MAGISTER EN INGENIERÍA INDUSTRIAL
AUTOR
JESÚS WALTER SALINAS FLORES
ASESOR Ing. NÉSTOR SANTOS
LIMA – PERÚ 2005

. . 1

RESUMEN . 3

INTRODUCCIÓN . 5

I. EL PROBLEMA DE INVESTIGACIÓN . . 7

1.1 FUNDAMENTACIÓN DEL PROBLEMA . 7

1.2 FORMULACIÓN DEL PROBLEMA . 8

1.3 OBJETIVO . 8

1.4 JUSTIFICACIÓN DE LA INVESTIGACIÓN . 8

II MARCO TEORICO . 11

2.1 CARACTERIZACIÓN DEL SISTEMA FINANCIERO NACIONAL . . 11

2.2 RIESGO DE CRÉDITO . . 12

2.3 BASES TEÓRICAS SOBRE EL RECONOCIMIENTO DE PATRONES . 14

2.3.1 RECONOCIMIENTO DE PATRONES . 14

2.3.2 ÁRBOLES DE CLASIFICACIÓN . . 16

2.4 METODOLOGÍA DEL ARBOL DE CLASIFICACIÓN CART . . 18

2.4.1 ESQUEMA GENERAL . . 18

2.4.2 CONSTRUCCIÓN DEL ÁRBOL DE CLASIFICACIÓN . 20

2.4.3 SELECCIÓN DE LAS PARTICIONES . . 20

2.4.4 ESTIMADORES DE ERROR . 25

2.4.5 LA ESTRATEGIA DE PODA . 26

2.4.6 SELECCIÓN DEL MEJOR ÁRBOL PODADO . 27

2.4.7 REGLA DE ASIGNACIÓN DE CLASES . 28

2.4.8 RESUMEN DE GANANCIAS . . 28

2.4.9 IMPORTANCIA DE LAS VARIABLES . 29

2.5 IMPORTANCIA DE LA UTILIZACIÓN DEL ARBOL DE CLASIFICACIÓN CART . . 29

2.6 DEFINICIÓN DE TÉRMINOS BÁSICOS . 30

III METODOLOGÍA DE LA INVESTIGACIÓN . . 31

3.1 TIPO DE INVESTIGACIÓN . 31

3.2 FORMULACIÓN DE LA HIPÓTESIS . . 31

3.3 IDENTIFICACION DE LAS VARIABLES . . 32

3.4 DESCRIPCIÓN DEL PRODUCTO CREDITICIO APLICADO . 32

3.5 DISTRIBUCIÓN DE DATOS EN EL PRODUCTO CREDITICIO APLICADO . 34

3.6 METODOLOGÍA APLICADA . 35

3.7 POBLACIÓN Y MUESTRA . 35

IV PROCEDIMIENTO DE LA TÉCNICA DEL ÁRBOL DE CLASIFICACIÓN CART . 37

4.1 DESCRIPCIÓN DEL SOFTWARE UTILIZADO . . 37

4.2 ANÁLISIS DESCRIPTIVO UNIVARIADO Y BIVARIADO . 37

4.2.1 ANÁLISIS DESCRIPTIVO UNIVARIADO . 38

4.2.2 ANÁLISIS DESCRIPTIVO BIVARIADO . 40

4.3. ANÁLISIS DEL ÁRBOL GENERADO POR EL ALGORITMO CART . 43

4.4 ANÁLISIS PARA LOS MOROSOS Y NO MOROSOS . 46

4.4.1 ANÁLISIS PARA LOS MOROSOS . 46

4.4.2 ANÁLISIS PARA LOS NO MOROSOS . . 49

4.5 IMPORTANCIA DE LAS VARIABLES PARA EL PRODUCTO CREDITICIO . . 51

V. RESULTADOS . 53

5.1 REGLAS DE CLASIFICACIÓN PARA LOS MOROSOS . . 53

5.2 REGLAS DE CLASIFICACIÓN PARA LOS NO MOROSOS . 54

5.3 VALIDACIÓN DE LOS RESULTADOS . 55

5.4 DISCUSIÓN DE RESULTADOS . 56

5.5 MEDIDAS CORRECTIVAS Y CONTROL DE MEJORAS . 57

CONCLUSIONES . . 59

RECOMENDACIONES . 61

BIBLIOGRAFÍA . 63

ANEXO . . 65

Anexo Nº 1 . 65

Anexo Nº 2 . 71

Anexo Nº 3 . 75

TEXTO COMPLETO EPUB . . 85

A mi esposa Violeta, mi compañera por siempre A mis hijos: Lucía y Alejandro, por quienes cada día trato de ser mejor Al Ing. Néstor Santos por su asesoríainvalorable

"Programa Cybertesis PERÚ - Derechos son del Autor"

RESUMEN

En los sistemas bancarios o financieros es importante analizar el riesgo crediticio cuando se le concede un crédito a un cliente para un producto determinado. El presente trabajo de investigación tiene como objetivo encontrar un patrón de comportamiento de la morosidad a partir de la información obtenida al momento de solicitar un crédito para un producto crediticio y a su vez dar a conocer una nueva técnica estadística muy útil para este campo que es la técnica de los Árboles de Regresión CART la cual se aplica en situaciones donde se tienen un conjunto de datos de individuos en los que se han medido variables predictoras o independientes y una variable de clasificación o de criterio que define el grupo al que cada individuo pertenece; y se quiere encontrar un conjunto de reglas de decisión que permitan explicar la clasificación existente y utilizar estas reglas para poder clasificar a un nuevo individuo, esta técnica tiene la ventaja de una alta potencia en sus estimadores.

En el presente trabajo se hace una descripción del CART como un algoritmo de desarrollo de árboles y se presenta una aplicación de dicho algoritmo al reconocimiento de patrones de morosidad y de no morosidad para un producto crediticio de una entidad del sistema financiero nacional que otorga préstamos a personas naturales.

Se presenta el marco teórico empezando con una caracterización del sistema financiero nacional. Luego se dan algunas referencias sobre el riesgo de crédito y se dan las bases teóricas sobre el reconocimiento de patrones. Finalmente se presenta la metodología del árbol de clasificación CART que es objetivo del presente trabajo y una definición conceptual de términos.

La metodología de la investigación esta referida a la descripción del producto crediticio y la estrategia para la prueba de hipótesis. Para aplicar el algoritmo de árbol de clasificación CART se ha trabajado con un producto crediticio del sistema financiero nacional que otorga préstamos en dólares a personas naturales. En este caso particular, el fin que cumple el producto crediticio consiste en otorgar líneas de crédito a todas aquellas personas naturales dependientes que cumplan con los requisitos establecidos por la entidad bancaria.

El producto crediticio se ofreció a personas dependientes de empresas o instituciones privadas y gubernamentales que se encontraron en situación de activos (empleados, obreros, etc.) o jubilados cuyos ingresos mensuales sean mayores a US $ 700.00 o su equivalente en moneda nacional.

Para el procesamiento de la información se trabajó con la información de todos los clientes que obtuvieron el producto crediticio. El trabajo se realizó en dos etapas:

Etapa I: Descripción univariada y bivariada de las variables descriptivas y de la variable criterio, para lo cual se utilizó el software estadístico SPSS versión 10.0.

Etapa II: Aplicación del algoritmo de árbol de clasificación CART, para lo cual se utilizó el software CART 4.0 for Windows de la compañía Salford Systems [7]; para generar los resultados de la clasificación usando como medida de impureza el índice de Gini con costos de mala clasificación iguales.

Finalmente, se presenta el procedimiento de la técnica del árbol de clasificación CART para el producto crediticio empleado, con la presentación, análisis e interpretación de los datos usando un análisis descriptivo univariado y bivariado. Luego se presenta el análisis del árbol generado por el algoritmo CART.

En los anexos se presenta un manual del software CART y las reglas de clasificación para los morosos y los no morosos.

"Programa Cybertesis PERÚ - Derechos son del Autor"

INTRODUCCIÓN

La evaluación de riesgos financieros de un banco o de una entidad crediticia es llevada a cabo por la Gerencia de Riesgo usando por lo general técnicas subjetivas o utilizando técnicas descriptivas univariadas/bivariadas o tomando como referencia resultados o modelos obtenidos en otras instituciones crediticias.

Cuando se otorgan préstamos o créditos a las personas naturales, se van a tener situaciones en las que se registran un conjunto de datos de “N” individuos o clientes clasificados en dos o más grupos ya definidos y en los que se han medido “p” variables predictoras o independientes y una variable de clasificación o de criterio que define el grupo al que cada individuo pertenece.

Para poder explicar la clasificación existente, las técnicas multivariadas tratan de encontrar una regla discriminante que clasifique correctamente a los individuos ya formados y posteriormente, aplicar esta regla a nuevos individuos.

Entre las técnicas multivariadas disponibles se tienen al análisis de regresión múltiple, el análisis de regresión logística, el análisis discriminante lineal y cuadrático, etc.

Sin embargo, existe otro enfoque para poder explicar la clasificación y predecir a los individuos que lo constituyen los sistemas de clasificación, dentro de los cuales se tiene al algoritmo CART (Classification and Regression Tree).

"Programa Cybertesis PERÚ - Derechos son del Autor"

I. EL PROBLEMA DE INVESTIGACIÓN

1.1 FUNDAMENTACIÓN DEL PROBLEMA

En el Perú hay cada vez más entidades crediticias, por lo que el mercado es cada vez más competitivo. Por esta razón para no perder mercado y a su vez, evitar degradar la calidad/riesgo de su cartera de colocaciones, una entidad crediticia debe aprovechar la experiencia y el conocimiento de sus propios clientes.

La cartera de crédito al consumo implica el manejo de un gran número de clientes. Las entidades financieras requieren procesar un gran número de solicitudes de crédito, por tanto es importante que la administración de riesgo de los bancos ejerza un control efectivo sobre el proceso de evaluación de un cliente a fin de que se le otorgue o se le niegue un crédito.

La evaluación de riesgos financieros de un banco o de una entidad crediticia es llevada a cabo por la Gerencia de Riesgo usando por lo general técnicas subjetivas o utilizando técnicas descriptivas univariadas/bivariadas o tomando como referencia resultados o modelos obtenidos en otras instituciones crediticias.

Sin embargo, existen técnicas que permiten encontrar patrones de comportamiento basado en un conjunto de variables independientes, que son factibles de aplicar en el mercado peruano como son los denominados árboles de clasificación, los cuales son

gráficos que ilustran reglas de decisión y que tienen como finalidad explicar la clasificación y predecir a los clientes que provienen de estos sistemas de clasificación, los cuales corresponden a una serie de reglas de decisión que predicen o clasifican observaciones futuras.

1.2 FORMULACIÓN DEL PROBLEMA

¿Es posible encontrar un patrón de comportamiento de la morosidad a partir de la información obtenida al momento de solicitar un crédito para un producto crediticio?

1.3 OBJETIVO

Determinar las características o el reconocimiento de patrones de morosidad del producto crediticio de una entidad del sistema financiero nacional, como un método que discrimine mejor el proceso de otorgación de un crédito, utilizando el algoritmo de árbol de clasificación CART.

1.4 JUSTIFICACIÓN DE LA INVESTIGACIÓN

En estos días los banqueros están buscando microempresas y profesionales independientes por una simple razón: en los países en vías de desarrollo éstos representan un gran mercado urgido de financiamiento, y estos empresarios y empresarias han demostrado ser riesgos crediticios sólidos.

Investigaciones realizadas han descubierto que las microempresas conformaban el segmento mayoritario, el cual representa a más del 70% de la fuerza laboral quienes son atendidos como clientes no convencionales en el Banco de Crédito del Perú.

La concesión de préstamos a profesionales independientes con ingresos muchas veces impredecibles representa una gran cantidad de desafíos para bancos como Solución, que otorga préstamos de un promedio de US $1.300, y para el Banco del Trabajo del Perú, con 100.000 clientes microempresarios.

Actualmente, en la mayoría de las instituciones financieras los ejecutivos de cuenta evalúan solicitudes de crédito utilizando su propia experiencia y guías publicadas y divulgadas por la institución. Este método tiene las siguientes dificultades:

- *Pobre desempeño* puesto que, la experiencia de la mayoría de los ejecutivos de cuenta es limitada, lo cual significa generar decisiones imprecisas.
- *Inflexibilidad*, porque se requiere experiencia de varios años antes que los ejecutivos

"Programa Cybertesis PERÚ - Derechos son del Autor"

se desempeñen bien. Por lo tanto, este método reacciona lento a los cambios en la población de solicitantes de créditos.

- *Inconsistencia*, puesto que, diferentes ejecutivos poseen estándares diferentes.
- *Baja eficiencia* al evaluar solicitudes consume tiempo e involucra también a personal experimentado de jerarquía superior.

Los bancos se están centrando en desarrollar métodos de evaluación de préstamos para reducir el riesgo y acortar el largo proceso de evaluación de créditos de clientes sin empleo formal e ingresos irregulares que se ven obligados a ofrecer garantías convencionales.

En el sector bancario internacional, la utilización de estas técnicas de árboles de clasificación posee un amplio desarrollo. La posibilidad de obtener características o patrones de comportamiento que permita clasificar a los clientes en solventes o insolventes potenciales es de gran utilidad.

En el sector bancario nacional, el uso de esta técnica de árbol de clasificación permitirá obtener las características de morosidad y de no morosidad usando la información histórica de los clientes al momento de solicitar un crédito.

"Programa Cybertesis PERÚ - Derechos son del Autor"

II MARCO TEORICO

2.1 CARACTERIZACIÓN DEL SISTEMA FINANCIERO NACIONAL

El sistema financiero está formado por todas las entidades y empresas que manejan y utilizan el dinero emitido por el Banco Central de Reserva del Perú, y los títulos, valores y documentos de crédito en general que lo sustituyen, como son los cheques, giros, letras, pagarés, órdenes de pago, cartas fianzas, cartas de crédito, warrants, pólizas de seguro, acciones, bonos, cédulas, etc. La cabeza del sistema financiero es el Banco Central de Reserva del Perú y por lo tanto es también la máxima autoridad en materia monetaria y manejo de los medios de pago en general.

Dentro de este sistema, se pueden diferenciar dos grandes grupos de empresas:

a) Las instituciones financieras y

b) Otras instituciones del sistema financiero

Dentro del primer grupo se encuentran todas aquellas que están organizadas y sobre todo autorizas oficialmente por la Superintendencia de Banca y Seguros para recibir dinero, o los documentos que lo sustituyen, en forma de depósitos del público; solicitar préstamos de otras entidades financieras y captar recursos mediante la emisión de

obligaciones (bonos y cédulas); y utilizarlos junto con sus propios capitales para colocarlos o invertirlos en diferentes formas de operaciones financieras que estimulen la producción y circulación de bienes y servicios en sus diferentes actividades: industria, comercio, agricultura, minería, transportes, pesquería, etc. y en esa forma impulsar el desarrollo sostenido del país.

Sin embargo, no todos los recursos ajenos obtenidos bajo las tres principales modalidades descritas se pueden utilizar para colocarlos e invertirlos, por cuanto es obligatorio conservar en efectivo o en forma de depósitos en el Banco Central de Reserva del Perú, un determinado porcentaje de aquellos, con el nombre de “encaje legal”, no solamente para tener fondos disponibles para hacer frente a los posibles retiros de los depósitos del público, o pago de las obligaciones asumidas, sino también como una forma de regulación del crédito para que éste cumpla con sus fines en la forma que más convenga a los intereses del país. Es por ello que las normas y porcentajes de encaje legal obligatorio que deben tener las instituciones financieras, están dirigidas y controladas por el Banco Central de Reserva del Perú.

Además de los recursos obtenidos de terceros, o capitales ajenos; toda institución financiera deberá tener sus propios capitales o patrimonio aportado por sus accionistas o asociados en su condición de propietarios de la empresa. Sin embargo, en estas instituciones, su patrimonio o capitales propios puede ser solamente 15% del total de los capitales ajenos, como mínimo porcentaje; que en una empresa mercantil sería inaceptable desde el punto de vista de solvencia o recursos propios de la misma.

Dentro del segundo grupo, que se denomina “otras instituciones del sistema financiero”, por cuanto también operan básicamente con dinero y documentos que lo sustituyen, pero que no pueden recibir depósitos del público para colocarlos o invertirlos por ser éstas las dos condiciones básicas para ser institución financiera, se encuentran otras, como las compañías de seguros, los almacenes generales de depósito, el fondo nacional de propiedad social y las empresas de arrendamiento financiero.

2.2 RIESGO DE CRÉDITO

Soler Ramos, José y Otros [8]; señalan que:

“Se entiende por riesgo la posibilidad de sufrir un daño. En este libro se tratará exclusivamente de riesgos económicos-financieros, en los que el daño consiste en una pérdida de valor económico. Para caracterizar completamente el riesgo sería necesario considerar todos los posibles escenarios futuros, asignarles una probabilidad y determinar los resultados económicos derivados de los mismos.”

“La caracterización exhaustiva del riesgo se hace impracticable, ya que se deberían establecer todas las combinaciones posibles de las variables que influyen sobre el valor económico de la cartera o negocio (tipos de interés a diversos plazos, tipos de cambio, precios de acciones, márgenes crediticios sobre los tipos libres de riesgo, etc.); además, su utilidad para la toma de decisiones sería limitada sin una adecuada sistematización.”

"Programa Cybertesis PERÚ - Derechos son del Autor"

"Uno de los problemas principales para la gestión adecuada de los riesgos es la medición de los mismos a través de indicadores que sinteticen adecuadamente el nivel de riesgo y sean sensibles a los factores del entorno que lo producen. Se han desarrollado dos grupos de metodologías:

- Análisis de escenarios
- Técnicas de probabilidad"

"El **análisis de escenarios** consiste en seleccionar unas pocas situaciones consideradas desfavorables y estimar las pérdidas asociadas, en general sin tener en cuenta las probabilidades de ocurrencia. Este planteamiento fue el primero que se desarrolló y todavía resulta útil e incluso insustituible para contemplar situaciones de crisis muy improbables pero no imposibles; sin embargo, es sólo una medida complementaria a otras caracterizaciones del riesgo, ya que presenta serias deficiencias:

- Los escenarios se eligen de manera subjetiva, sin garantizar que cuando un analista define un caso desfavorable lo hace con los mismos criterios que otro, lo cual impide alcanzar la homogeneidad necesaria para comparar los niveles de riesgo en distintos instantes, y entre diferentes actividades y negocios.
- No se llegaría a conocer la probabilidad de sufrir un cierto nivel de pérdidas, incluso si se determinara la probabilidad de cada uno de los escenarios analizados, ya que para ello sería necesario contemplar la totalidad de los escenarios que ocasionarían pérdidas similares".

"Las metodologías de reciente difusión, basadas en técnicas de probabilidad, han permitido salvar estos problemas, ayudando a construir tablas en las que se recoge el importe de cada una de las pérdidas posibles junto con la probabilidad de que se alcance dicho nivel."

"Para clasificar los tipos de riesgos suele atenderse a la naturaleza del causante de la pérdida, calificando el riesgo como:

- De mercado, si se debe a variaciones de los precios/tipos negociados en los mercados financieros. A su vez, este riesgo se subdivide en riesgo de interés, de tipo de cambio, de renta variable, de mercancías (commodities), de volatilidad, de correlación, etc.
- De crédito, si se debe al incumplimiento de contratos por insolvencia.
- De liquidez, si se debe a dificultades para financiar con un costo "normal" los negocios y su crecimiento.
- De negocio, si se debe a caídas en el volumen del negocio o de sus márgenes.
- Operativo, si se debe a errores humanos o de los medios de producción o gestión.
- Legal, si se debe a la incapacidad legal para ejercer los derechos que se consideraban como propios, o si se deriva del propio incumplimiento de la ley."

Gestión y control del riesgo de crédito

"Desde un punto de vista general, se puede definir el riesgo de crédito como la

posibilidad de sufrir pérdidas si los clientes y contrapartidas, con las que la entidad tiene contratadas operaciones, incumplen los compromisos contractuales que han adquirido, por falta de solvencia."

"Una entidad financiera asume el riesgo de crédito en los diferentes negocios en los que opera. En la actividad de banca comercial (corporativa, de empresas, de personas, etc.) las entidades asumen riesgo de crédito por los préstamos, líneas de crédito, garantías, avales, etc. que conceden."

2.3 BASES TEÓRICAS SOBRE EL RECONOCIMIENTO DE PATRONES

2.3.1 RECONOCIMIENTO DE PATRONES

Watanabe [11]; señala que:

"Un patrón es una entidad a la que se puede dar un nombre y que está representada por un conjunto de propiedades medidas y las relaciones entre ellas (vector de características). Por ejemplo, un patrón puede ser una señal sonora y su vector de características el conjunto de coeficientes espectrales extraídos de ella (espectrograma). El reconocimiento automático, descripción, clasificación y agrupamiento de patrones son actividades importantes en una gran variedad de disciplinas científicas, como biología, psicología, medicina, visión por computador, inteligencia artificial, teledetección, etc."

El mismo autor menciona que "un sistema de reconocimiento de patrones tiene uno de los siguientes objetivos:

- Identificar el patrón como miembro de una clase ya definida (clasificación supervisada).
- Asignar el patrón a una clase todavía no definida (clasificación no supervisada, agrupamiento o clustering)."

"El diseño de reconocimiento de patrones se lleva a cabo normalmente en tres fases:

- Adquisición y preproceso de datos.
- Extracción de características.
- Toma de decisiones o agrupamiento."

Tau y Gonzalez [10]; señalan que:

"El proceso de reconocimiento de patrones tiene que ver con la actividad de clasificar un objeto o hecho en alguna categoría determinada, en base a particularidades del objeto que son comunes a los miembros de una misma clase."

"Algunos ejemplos de actividades comunes de reconocimiento de patrones

"Programa Cybertesis PERÚ - Derechos son del Autor"

realizados por humanos son:

- Escuchar una conversación.
- Leer un libro.
- Reconocer un amigo entre la multitud.
- Traducir de un lenguaje a otro."

Una definición formal de reconocimiento de patrones que dan los mismos autores es la siguiente:

"Es la categorización de datos de entrada en clases identificadas, por medio de la extracción de características significativas o atributos de los datos extraídos de un medio ambiente que contiene detalles irrelevantes."

Matemáticamente hablando, la clasificación consiste en la partición del espacio n-dimensional definido por las características de un objeto, en varias regiones, donde cada región corresponde a una clase.

Se definen algunos términos básicos:

Patrón - Descripción de un objeto o miembro de alguna categoría.

Clase - Categoría determinada por algunos atributos comunes a sus miembros.

Vector de características - Contiene la medida de las características de un patrón; puede estar formado de números binarios o valores reales."

El diseño de un sistema reconocedor de patrones presenta 4 problemas fundamentales:

1) La representación de los datos de entrada.

2) La extracción de características.

3) La determinación del proceso de decisión óptimo.

4) Problemas de estimación de parámetros.

Un sistema automático de clasificación está formado de las siguientes partes (ver Gráfico Nº 1).

Traductor - Toma información del objeto y la transforma a señales analógicas, que a su vez son convertidas a su forma digital.

Extractor de características - Mide propiedades importantes que representan al objeto. Su propósito es reducir la cantidad de datos obtenidos del traductor sin perder información esencial. El resultado de tales medidas se llama vector de características.

Clasificador - Evalúa las propiedades dadas por el extractor de características y toma una decisión final acerca de la clase del objeto.

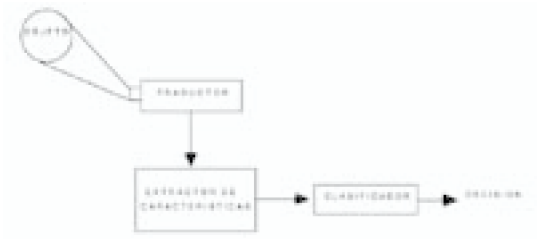


Gráfico Nº 1: Sistema clasificador de Patrones

Los métodos de clasificación pueden ser de varios tipos. El método escogido depende de las características de las clases. Se pueden identificar 3 tipos de metodologías:

- Heurísticas. Se basan en la intuición y la experiencia humana. El diseño es tipo "ad-hoc".
- Matemáticas. Se dividen en 2 tipos:
  - Determinísticas. No requieren de propiedades estadísticas, pero son limitadas.
  - Estadísticas. Emplean propiedades estadísticas de las clases.
- Lingüísticas o sintácticas. Se basan en elementos primitivos (subpatrones) y relaciones entre ellos, definiendo gramáticas formales (variables, primitivas, reglas de producción, etc.)

Existen 2 fases fundamentales en un clasificador:

- Entrenamiento.
- Prueba de reconocimiento.

Tau y Gonzalez [10]; señalan que: "después de seleccionar las características y decidir el tipo de reconocedor a implementar se deben recolectar datos para entrenamiento y pruebas. Durante el entrenamiento se usan datos de entrenamiento, que deben ser representativos del medio ambiente. Para medir la capacidad del clasificador se usan los datos de prueba."

2.3.2 ÁRBOLES DE CLASIFICACIÓN

Bonson □2□; menciona que "la clasificación supervisada es una de las tareas que más frecuentemente son llevadas a cabo por los denominados Sistemas Inteligentes. Por lo tanto, un gran número de paradigmas desarrollados bien por la Estadística (regresión logística, análisis discriminante, K-NN) o bien por la Inteligencia Artificial (redes neuronales, inducción de reglas, árboles de clasificación, redes bayesianas) son capaces de realizar las tareas propias de la clasificación."

La compañía SPSS □9□ menciona que "los árboles de clasificación corresponden a gráficos que ilustran las reglas de decisión. Parten de un nodo raíz que contiene todas las

"Programa Cybertesis PERÚ - Derechos son del Autor"

observaciones de la muestra. A medida que se desarrolla el árbol, los datos se dividen en ramas de subconjuntos de datos exclusivos. El Gráfico Nº 2 muestra un árbol de decisión que describe la política de admisión de un instituto de enseñanza superior, basándose en promedio de notas y puntuaciones de pruebas estándar (MCAT).”



Gráfico Nº 2: Ejemplo de Árbol de Decisión

Acuña, E. [1]; menciona que “el uso de árboles de decisión tuvo su origen en las ciencias sociales.”

“El término árboles es por la gráfica, aunque los árboles son mostrados creciendo hacia la parte final de la página. La raíz del nodo superior, en cada nodo se hace una decisión hasta llegar a un nodo terminal u hoja. Cada nodo no-terminal contiene una pregunta en la cual se basa la división del nodo. Cada hoja contiene el nombre de la clase a la cual pertenece.”

“Las diferencias principales entre los algoritmos para construir árboles se hallan en la estrategia para cortar árboles y la regla para dividir los nodos.”

Hair y otros [4]; sostienen que “los árboles de decisión, tiene aspecto familiar (como el dendograma del análisis cluster), pero se construye e interpreta de forma completamente distinta. Los árboles de decisión son particiones secuenciales del conjunto de datos para maximizar las diferencias de la variable dependiente. Los dos programas más utilizados son CHAID (detector de interacción chi-cuadrado) y CART (árboles de clasificación y regresión). Los árboles de decisión ofrecen una forma concisa de desarrollar grupos que son consistentes en sus atributos pero que varían en términos de la variable dependiente.”

Según Breiman y otros [3]; “un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto de prototipos. Esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto y cada nodo hoja se refiere a una decisión (clasificación).”

“La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezando por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.”

Según Johnson, D. □5□; “se ha sugerido otro enfoque para el análisis discriminante. Puede valer la pena considerar este método si se tiene un gran número de variables

discretas. Este procedimiento, del cual puede disponerse en el software CART, se desarrolla una serie de preguntas a las que se puede responder sí o no. Las respuestas a estas preguntas llevan al investigador a través de las ramas de un árbol, hasta que llega a un lugar en el que se puede hacer una clasificación. En cada nodo del árbol, el procedimiento encuentra la mejor pregunta para hacer en esa etapa en particular, con el fin de ayudar al investigador a tomar la mejor decisión. Al final de cada rama se toma decisión en cuanto a dónde debe clasificarse una observación particular."

El mismo autor menciona que: "CART puede construir un árbol binario de predicción con una estructura que es fácil de entender, interpretar y usar. La construcción de árboles comienza usando un algoritmo que recursivamente busca a lo largo de todas las variables para producir una secuencia óptima de divisiones binarias, de tal manera que tenemos un árbol extremadamente profundo. Un algoritmo de poda es aplicado posteriormente. Dicho algoritmo poda todas las ramas del árbol que perjudican en la asignación de nuevos casos. El árbol puede ser inspeccionado y usado para modificar el algoritmo de organización. El procedimiento es repetido hasta que el mejor conjunto es encontrado. Con cada análisis se puede realizar una cross-validation un número determinado de veces."

La compañía SPSS □9□; menciona que "una de las técnicas más populares en Data Mining es la segmentación basada en árboles. Esta clase de métodos incluye al CHAID, árboles de regresión y clasificación (CART), y al QUEST. Estos métodos empiezan con todos los registros y los dividen en dos o más subgrupos que tienden a ser más homogéneos que la muestra original. Cada uno de estos subgrupos se divide nuevamente en subgrupos más pequeños y así sucesivamente hasta encontrar una buena solución o cuando no hayan más individuos para continuar."

2.4 METODOLOGÍA DEL ARBOL DE CLASIFICACIÓN CART

2.4.1 ESQUEMA GENERAL

Breiman, L. y otros [3], proponen el algoritmo de árbol de clasificación CART. Constituye uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado.

El algoritmo CART se caracteriza por realizar particiones binarias y por utilizar una estrategia de poda basada en el criterio de coste-complejidad.

Los datos con que trabajan corresponden a un conjunto de individuos u objetos N O_1,...,O_S denominados prototipos.

Cuadro Nº 1: Cuadro de Datos para el algoritmo de árbol de clasificación CART

"Programa Cybertesis PERÚ - Derechos son del Autor"

Individuo	x_1	x_2	....	x_j	....	x_d	Clase
O_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1d}	1
O_2	x_{21}	x_{22}	...	x_{2j}	...	x_{2d}	1
.	.	.	.	.	.	.	.
O_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{id}	
.	.	.	.	.	.	.	.
O_S	x_{s1}	x_{s2}	....	x_{Sj}	....	x_{Sd}	J

Fuente: elaboración propia

Una forma sistemática de predecir la pertenencia de una clase es una regla que asigne una clase a cada vector x en X, la regla asigna una de las clases 1,...,J a x.

Un clasificador o regla de clasificación es una función d(x) definida en X para cada x, d(x) es igual a uno de los números 1, 2,...,J.

La metodología a seguir puede resumirse en dos pasos: que se presentan en el Gráfico 3:

1. Aprendizaje. Consiste en la construcción del árbol a partir de un conjunto de prototipos, S.

2. Clasificación. Consiste en el etiquetado de un patrón, X, independiente del conjunto de aprendizaje. Se trata de responder a las preguntas asociadas a los nodos interiores utilizando los valores de los atributos del patrón X. Este proceso se repite desde el nodo raíz hasta alcanzar una hoja, siguiendo el camino impuesto por el resultado de cada evaluación.

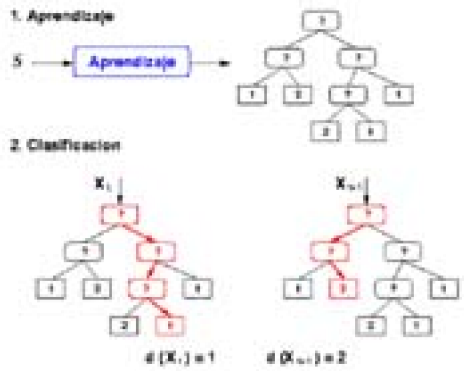


Gráfico Nº 3 Metodología del árbol de clasificación CART

La construcción de un clasificador está basada en una muestra de aprendizaje, la cual consiste de datos $(x_1, j_1),...,(x_s,j_s)$ de S casos donde $x_n \in X$ y $j_n \in \{1,...,J\}$, n=1,...,S. La muestra de aprendizaje es denotada como Λ, donde:

$\Lambda = \{(x_1,j_1),...,(x_s j_s)\}$

Un árbol de clasificación T representa una partición recursiva del espacio de representación, P, realizada sobre un conjunto de prototipos S.

Existe una relación entre nodos de T, regiones en P y conjuntos en S que se resumen en los siguientes puntos:

1. Cada nodo de T tiene asociado un subconjunto de prototipos de S.

2. El nodo raíz tiene asociado el conjunto completo.

3. Cada hoja, t, tiene asociada una región, R_t, en P. Así si $\hat{T}$ es el conjunto de nodos hoja del árbol T:

$$\bigcup_{t \in \hat{T}} R_t = P$$

que se interpreta que los conjuntos de prototipos asignados a los nodos hoja constituyen una partición de P.

4. Cada nodo interior tiene asociado una región en P, que es la unión de las regiones asociadas a los nodos hoja del subárbol cuya raíz es él.

5. La unión de los conjuntos de prototipos asignados a los nodos de un mismo nivel da como resultado el conjunto inicial.

2.4.2 CONSTRUCCIÓN DEL ÁRBOL DE CLASIFICACIÓN

La construcción del árbol de decisión constituye la fase de aprendizaje. Las líneas generales para su construcción pueden resumirse en los siguientes puntos, de acuerdo a un esquema recursivo:

1. El avance está basado en la partición de un nodo de acuerdo a alguna regla, normalmente evaluando una condición sobre el valor de alguna variable. Los prototipos que verifican la condición se asignan a uno de los dos nodos hijo y los restantes, al otro. Cuando un nodo se particiona, pasa a ser un nodo intermedio.

2. El caso base o condición de parada tiene como objetivo detener el proceso de partición de nodos. Cuando se verifica la condición de parada en un nodo, éste es un nodo hoja. Los prototipos asociados a un nodo hoja constituyen un agrupamiento homogéneo, por lo que al nodo se le asigna una etiqueta. En ocasiones, se poda el árbol resultante utilizando alguna regla de poda.

2.4.3 SELECCIÓN DE LAS PARTICIONES

Una partición divide a un conjunto de prototipos en conjuntos disjuntos. Con CART las particiones son binarias, resultado de evaluar una condición que tiene dos únicas respuestas: sí o no.

El objetivo de una partición es incrementar la homogeneidad de los subconjuntos resultantes, o lo que es lo mismo, que éstos sean más puros que el conjunto originario. Cada partición tiene asociada una medida de pureza, que se utiliza para la selección de la mejor partición.

"Programa Cybertesis PERÚ - Derechos son del Autor"

2.4.3.1 FORMULACIÓN DE LA REGLA DE PARTICIÓN

Sea Q el conjunto de preguntas binarias de la forma:

¿X Є A?, A C P

El conjunto Q genera un conjunto de particiones, s, en cada nodo, t y cada nodo t se particiona en t_L (izquierdo) y t_R (derecho) de manera que:

- Los casos de t que verifican la condición ¿X Є A? se asignan a t_L

$$t_L = t \cap A$$

- Los casos de t que no verifican la condición ¿X Є A? se asignan a t_R

$$t_R = t \cap \overline{A}$$

En CART se define el conjunto estándar de preguntas de la siguiente forma:

1. Cada partición depende de un único atributo.

2. Si X_i es un atributo categórico, que toma valores en { c_1, c_2,..., c_L}, Q incluye las preguntas:

¿X_i Є C?

donde C es un conjunto de entre los subconjuntos de { c_1, c_2,..., c_L}.

3. Si X_i es un atributo continuo, Q incluye preguntas del tipo:

¿$X_i \leq v$?

donde v es un valor real, teóricamente cualquiera. En CART, no obstante, v es el punto medio de dos valores consecutivos de X_i. Esto simplifica los cálculos haciendo que los problemas sean tratables.

El conjunto de preguntas que pueden formularse (y que evaluará el CART), aunque puede ser muy amplio, está perfectamente definido y es sistemático.

2.4.3.2 CRITERIOS DE PARTICIÓN

Una vez que se conoce cómo se formulan las preguntas (cada una da lugar a una partición) se trata de conocer cómo calcular la mejor partición entre todas las posibles.

Cada partición tiene asociada una *medida de pureza* y se tratará de incrementar la *homogeneidad* de los subconjuntos resultantes de la partición, esto es, que sean más *puros* que el conjunto originario.

Función de impureza y medida de impureza

Una *función de impureza* es una función □ definida sobre J-uplas de la forma (c_1, c_2,..., c_J) tales que: a)

$$c_j \geq 0 \text{ para } j = 1, 2, \ldots, J \text{ y b) } \sum_j c_j = 1$$

con las siguientes propiedades:

i) Φ tiene un único máximo en

$$\left(\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J}\right)$$

ii) Φ alcanza su mínimo en (1, 0, 0,..., 0), (0, 1, 0,..., 0),..., (0, 0, 0,..., 1) y el valor del mínimo es 0.

iii) Φ es una función simétrica de c_1, c_2,..., c_J.

Relacionada con la función de impureza está la medida de impureza de un nodo. Dada una función de impureza Φ, se define la *medida de impureza* de cualquier nodo t, y se escribe i(t),como:

$$i(t) = \phi(p(1|t), p(2|t), \ldots, p(j|t))$$

donde p(j| t) es la probabilidad de que un caso del nodo t (un prototipo asociado al nodo t) sea de clase j. Estas probabilidades pueden calcularse empíricamente como la proporción de casos de clase j en el nodo t:

$$p(j/t) = \frac{N_j(t)}{N(t)}$$

La medida de impureza de un nodo es el resultado de evaluar la función de impureza sobre ese nodo tomando las proporciones relativas de cada clase como los c_j. Se debe tener en cuenta, por un lado,

$$A)\ p(j|t) \geq 0$$

$$b)\ \sum_j p(j|t) = \sum_j \frac{N_j(t)}{N(t)} = \frac{1}{N(t)} \sum_j N_j(t) = 1$$

"Programa Cybertesis PERÚ - Derechos son del Autor"

lo que garantiza que los componentes de la J-upla, calculados en términos de proporción relativa son válidos para evaluar la función de impureza. Por otro lado,

i) La máxima impureza (o la mínima pureza) se obtiene cuando *todas* las clases están igualmente representadas en t.

ii) La mínima impureza (o la máxima pureza) se obtiene cuando en t sólo hay casos de *una sola* clase (máxima homogeneidad).

iii) Cualquier permutación de los c_j produce el mismo resultado en el valor de impureza.

Bondad de una partición

La bondad de una partición s en un nodo t debe estar relacionada con la impureza del nodo sobre el que se realiza la partición, t, y con la impureza de los nodos resultantes de la partición, t_L y t_R.

Suponiendo que una partición candidata, s, que divide t en t_L y t_R de forma que una proporción p_L de los casos de t van a t_L y una proporción p_R van a t_R, tal como se presenta en el Gráfico Nº 4.

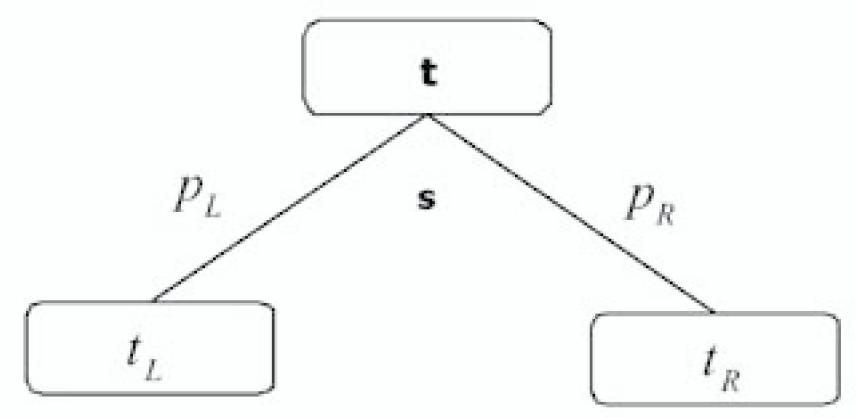


Gráfico Nº 4: La partición s divide t en tL y tR

La *bondad de la partición* s *en un nodo* t, Φ(s, t), se define como el *decrecimiento en impureza* conseguido con ella:

$$\Phi(s, t) = \Delta i(s, t) = i(t) - p_L\, i(t_L) - p_R\, i(t_R)$$

Así, como se conoce cómo calcular i(t), se puede calcular Φ (s, t) para cada partición s y seleccionar la *mejor partición* como la que proporciona la mayor bondad Φ (s, t).

Para establecer el efecto que produce la selección de la mejor partición en cada nodo sobre el árbol final se necesita una medida de la impureza global del árbol.

Impureza de un árbol

Sea un árbol binario T, obtenido mediante una serie de particiones. Sea T el conjunto de nodos terminales del árbol T.

Sea I(t) = i(t)p(t), donde p(t) es la probabilidad de que un caso cualquiera esté en el nodo t. La *impureza del árbol* T, I(T), se define como:

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t)\, p(t)\ )$$

La impureza de un árbol se calcula únicamente en base al conjunto de nodos terminales.

Una conclusión a la que llegaron Breiman y otros [3]; es la siguiente:

"La selección continuada de las particiones que maximizan Δi(s, t) *es equivalente a seleccionar las particiones que minimizan la impureza global* I(T)",

lo que significa que la estrategia de selección de la mejor partición en cada nodo conduce a la solución óptima considerando el árbol final.

Criterios de medida de impureza

Entre los criterios más utilizados se encuentran la medida de entropía y el índice de Gini.

· *Medida de entropía* .

Mide la entropía en un nodo t de acuerdo a la formulación clásica siguiente:

$$i(t) = -\sum_{j=1}^{J} p(j/t) \log p(j/t)$$

Se asume que 0 log 0 = 0

· *Índice de Gini* .

Mide la diversidad de clases en un nodo:

$$i(t) = \sum_{\substack{i,j=1 \\ i \neq j}}^{J} p(i/t)\, p(j/t) = 1 - \sum_{j=1}^{J} p(j/t)^2$$

Criterio de parada

Antes de abordar la solución a este problema se necesita saber cómo asociar una medida de error a un árbol de decisión.

"Programa Cybertesis PERÚ - Derechos son del Autor"

2.4.4 ESTIMADORES DE ERROR

Si $R^*(T)$ es el *error real* asociado al árbol T, el problema que se plantea es el de estimar este valor.

Se define, R(T), el *estimador por resustitución del error global de clasificación* de T, como:

$$R(T) = \sum_{t \in \tilde{T}} R(t) = \sum_{t \in \tilde{T}} r(t)p(t)$$

donde: r(t) = 1 –max {p(j| t)} y p(t) es la probabilidad de que un caso cualquiera esté en el nodo t.

Con esta definición, si T' se construye a partir de T a través de una partición *arbitraria* de un nodo terminal de T, siempre se verifica:

$R(T') \leq R(T)$

o de otra forma, $R(t) \geq R(t_L) + R(t_R)$. R(T) decrece conforme el tamaño del árbol se hace mayor, llevado al extremo se pueden construir árboles en los cuales los nodos terminales tengan un sólo prototipo. Para estos árboles, R(T) = 0.

La utilización de una medida de error, el estimador por resustitución, R(T), induce a lo que se conoce como *sobreaprendizaje* , entendiéndose como que el clasificador está muy ajustado al conjunto de entrenamiento, proporcionando índices de error muy bajos, pero que no son realistas ya que no son extrapolables a otros conjuntos. Así, se requiere establecer un estimador más realista del error real, lo que en esta terminología se conoce como un *estimador honesto* . Estimadores honestos del error asociado a un árbol T, $R^*(T)$, son:

a) $R^{ts}(T)$, *Estimador por conjunto de prueba* .

Se calcula partiendo el conjunto de aprendizaje, S, en dos conjuntos disjuntos, uno destinado al aprendizaje, S^l, y el otro a prueba, S^t.

Este estimador es aplicable si el conjunto de entrenamiento es lo suficientemente grande para ser particionado.

b) $R^{cv}(T)$, *Estimador por validación cruzada* .

Si el conjunto de entrenamiento tiene pocos prototipos no es aconsejable el uso del estimador por conjunto de prueba ya que reduce aún más el tamaño efectivo del conjunto de aprendizaje. Suele utilizarse validación cruzada con V = 10 conjuntos.

En la estimación por validación cruzada, la muestra de aprendizaje Λ es dividida en V subconjuntos, Λ_v, v=1,....,V cada uno conteniendo el mismo número de casos. La v-ésima muestra de aprendizaje es

$\Lambda^{(v)} = \Lambda - \Lambda_v$, v=1,...,V

$\Lambda^{(v)}$ contiene la fracción (V-1)/V del total de casos.

De todo lo expuesto se puede concluir que el procedimiento recomendado para establecer un criterio de parada se basa en dos principios:

1. *Podar en lugar de impedir el crecimiento* **.**

Se tratará de construir un árbol muy grande para posteriormente proceder a podarlo hacia la raíz de manera *adecuada* (se podarán los subárboles que produzcan pequeños beneficios de bondad). El resultado será una secuencia decreciente de árboles.

2. *Utilizar estimadores honestos de* $R^*(T)$.

Estos estimadores se usarán para seleccionar el árbol del tamaño adecuado de entre la secuencia de árboles podados. El estimador que se utilizará dependerá del tamaño del conjunto de aprendizaje.

2.4.5 LA ESTRATEGIA DE PODA

Como se ha mencionado, una de las características distintivas de CART es su estrategia de poda. Breiman y otros [3] inciden en la importancia de la poda frente al de selección de particiones. Su argumento es que resulta más eficiente podar un árbol que detener su crecimiento: la poda permite que un subárbol de un nodo permanezca y la otra desaparezca, mientras que detener el crecimiento poda *ambas* ramas simultáneamente.

El procedimiento general de poda puede resumirse como sigue:

1. Particionar nodos hasta que se cumpla alguna de estas condiciones:

a) que sea totalmente puro, o

b) $N(t) < N_{min}$ (habitualmente $N_{min} = 5$)

O sea, hasta que un nodo sea perfectamente homogéneo o hasta que tenga asociados a pocos prototipos. Resulta evidente que el resultado debe ser un árbol muy grande, al que se denomina T_{max}.

2. Una vez obtenido T_{max} se trata de podar este árbol, obteniendo una secuencia *decreciente* y *anidada* de árboles.

Si T' se obtiene a partir de T por poda, T' es un *subárbol podado* de T y se escribe T' >T. Así, la secuencia decreciente de árboles podados y anidados será la siguiente:

$$T_{\max} \succ T_1 \succ T_2 \succ \cdots \succ \{t_1\}$$

de manera que el árbol $\{t_1\}$ es un árbol que consta de un único nodo.

Uno de estos árboles será el que se seleccione finalmente.

Para realizar esta selección se asocia una medida de error a cada árbol de la secuencia y se escoge aquel que tenga asociado el menor error.

La idea básica subyacente es la de podar aquellos árboles que produzcan pequeños

"Programa Cybertesis PERÚ - Derechos son del Autor"

beneficios de bondad. Se espera que árboles podados (más simples) produzcan mejores resultados que los obtenidos con árboles más grandes (más complejos) al clasificar patrones independientes, esto es, los árboles podados tendrán más capacidad de generalización al no estar tan ajustados al conjunto de aprendizaje (el problema del *sobreaprendizaje*).

A continuación se presenta el procedimiento de poda basado en el criterio de mínimo coste-complejidad.

PODA POR MÍNIMO COSTE-COMPLEJIDAD

La medida de coste-complejidad involucra un proceso en el que un parámetro de penalización es continuamente incrementado, de manera que la poda se realiza en base a diferentes umbrales. Este proceso da lugar a una secuencia anidada de árboles.

La *medida de coste-complejidad* se define como sigue:

- Para cualquier subárbol $T < T_{max}$ se define su *complejidad* como el número de nodos terminales, $| T |$.
- El error de clasificación asociado al árbol T es $R_\alpha(T)$.
- La *medida de coste-complejidad* asociada al árbol T, $R_\alpha(T)$, se calcula

$$R_\alpha(T) = R(T) + \alpha \mid \tilde{T} \mid$$

donde α es un valor real ($\alpha \geq 0$) (*parámetro de complejidad*) que se interpreta como el coste de complejidad por nodo terminal.

Así, $R_\alpha(T)$ es una combinación lineal del coste del árbol y su complejidad, ponderada apropiadamente. De otra manera, se trata de combinar *bondad* y *simplicidad* . Un valor alto de □ produce un incremento en el coste por nodos terminales, lo que se traduce en penalizar un subárbol con un alto número de nodos terminales.

Para cada α, se trata de encontrar el árbol T(α), $T(\alpha) < T_{max}$, que minimiza $R_\alpha(T)$,

$$R_\alpha(T(\alpha)) = \min_{T \prec T_{max}} \{R_\alpha(T)\}$$

2.4.6 SELECCIÓN DEL MEJOR ÁRBOL PODADO

El objetivo es seleccionar uno de los árboles de la secuencia de árboles podados como el árbol óptimo, que constituirá el árbol solución. El procedimiento de selección trata de asociar una medida de error a cada árbol de la secuencia y escoger aquel que tenga asociado el menor error:

$$\text{Escoger } T_{k0} \text{ si } \hat{R}(T_{k0}) = \min_k \hat{R}(T_k)$$

El problema que hay que resolver ahora es responder a la pregunta: ¿cómo estimar honestamente $R^*(T_k)$?. Estimadores honestos del error asociado a un árbol T son $R^{ts}(T)$, el *estimador por conjunto de prueba* y $R^{cv}(T)$, el *estimador por validación cruzada* . Su aplicación para la selección del mejor árbol podado es de la siguiente manera:

a) Estimación por conjunto de prueba

El estimador por conjunto de prueba es aplicable si el conjunto de entrenamiento es suficientemente grande y se calcula partiendo el conjunto de aprendizaje, S, en dos conjuntos disjuntos, uno destinado para aprendizaje, S^l, y otro destinado a prueba, S^t.

- El primero, S^l, se usa para construir T_{max}, y a partir de él la secuencia $T_1 > T_2 > \cdots > \{t_1\}$ como se ha descrito anteriormente.
- El segundo, S^t, se usa como conjunto de prueba, de manera que para cada T_k, se clasifican las muestras de S^t utilizando el clasificador T_k y se obtiene $R^{ts}(T_k)$.

El árbol más adecuado, T_{k0}, es el que verifica:

$R^{ts}(T_{k0}) = \min_k R^{ts}(T_k)$

b) Estimación por validación cruzada

Si el conjunto de entrenamiento tiene pocos prototipos no es aconsejable el uso del estimador por conjunto de prueba ya que reduce aún más el tamaño efectivo del conjunto de aprendizaje. Suele utilizarse validación cruzada con V = 10 conjuntos. El árbol seleccionado, T_{k0}, será aquel que verifique:

$R^{cv}(T_{k0}) = \min_k R^{cv}(T_k)$

2.4.7 REGLA DE ASIGNACIÓN DE CLASES

$$j(t) = j \text{ si } p(j|t) = \max_{i=1,2,\ldots,J} \{p(i|t)\}$$

Si el máximo se alcanza para dos o más clases, se realiza un sorteo y se asigna arbitrariamente cualquiera de ellas.

2.4.8 RESUMEN DE GANANCIAS

El resumen de ganancias proporciona estadísticos descriptivos para los nodos terminales de un árbol. Permite identificar los nodos terminales deseados a partir de sus valores de ganancia.

"Programa Cybertesis PERÚ - Derechos son del Autor"

Si la variable criterio es categórica (nominal u ordinal), el resumen de ganancias muestra el porcentaje de casos de la categoría criterio seleccionada.

2.4.9 IMPORTANCIA DE LAS VARIABLES

El aspecto crítico es como rankear aquellas variables que si bien no proporcionan la mejor división de un nodo, pueden proporcionar la segunda o tercera mejor. Por ejemplo, una variable X_1 puede nunca participar en una división en la estructura final de un árbol. Si una variable X_2 es removida, X_1 puede participar en las divisiones y el árbol resultante puede ser tan preciso como el anterior. Para tal situación se requiere de un método de ranking de variables que detecte la importancia de X_1.

La medida de la importancia de una variable es la suma en todos los nodos de la disminución de la impureza producida por la mejor división de X_m en cada nodo.

2.5 IMPORTANCIA DE LA UTILIZACIÓN DEL ARBOL DE CLASIFICACIÓN CART

Lewis, R [6]; menciona que “el algoritmo CART posee las siguientes ventajas:

- Es un método bastante flexible. Las variables independientes pueden ser una mixtura de variables cualitativas y cuantitativas.
- Permite considerar los costes de clasificación errónea en el proceso de desarrollo del árbol.
- Permite asignar la distribución de probabilidades a priori en los problemas de clasificación.
- CART no hace asunciones sobre la distribución de ningún tipo, ya sea de las variables predictoras como de la variable criterio.
- CART no está afectado por valores extremos, colinealidad, heterocedasticidad que afecten los procedimientos establecidos. Los valores outliers pueden ser aislados en un nodo y no tienen ningún efecto en la división.
- CART puede detectar y revelar interacciones en un conjunto de datos.
- CART es invariante con relación a la transformación de las variables independientes. Es decir, las transformaciones no afectan el árbol producido.
- CART con el conjunto de variables predictoras o independientes, produce resultados usando sólo las variables más importantes.”

No obstante, el mismo autor menciona que “CART presenta algunos inconvenientes. Entre sus principales desventajas se tienen:

- Al ser un algoritmo binario, tiende a generar árboles de muchos niveles. Por ello, el

árbol resultante puede que no presente los resultados de manera eficiente, sobre todo si la misma variable ha sido utilizada para la división de varios niveles consecutivos.

- También tiende a seleccionar las variables que pueden permitir un mayor número de divisiones en el proceso de crecimiento del árbol. Por tanto, las conclusiones extraídas de las estructuras de árbol pueden resultar poco fiables.
- CART es un método muy complejo. Los cálculos pueden ser muy lentos cuando se utilizan conjuntos de datos voluminosos. CART utiliza la división por substitutos para tratar los valores perdidos.
- CART no está basado en un modelo probabilístico. No existen intervalos de confianza asociados con las predicciones derivadas usando el algoritmo CART para clasificar a un conjunto de datos."

2.6 DEFINICIÓN DE TÉRMINOS BÁSICOS

Algoritmo CART (Classification and Regression Tree): algoritmo de árbol de clasificación que se caracteriza por realizar particiones binarias y por utilizar una estrategia de poda basada en el criterio de coste-complejidad.

Árbol de clasificación: es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto de prototipos. Esta partición recursiva se traduce en una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol.

Historial crediticio: representa la hoja histórica de un cliente que solicita un préstamo.

Patrón: es una entidad a la que se puede dar un nombre y que está representada por un conjunto de propiedades medidas y las relaciones entre ellas (vector de características).

Riesgo: incertidumbre respecto a la realización de ganancias futuras. No implica caos o desconocimiento. Es la posibilidad de sufrir un daño.

"Programa Cybertesis PERÚ - Derechos son del Autor"

III METODOLOGÍA DE LA INVESTIGACIÓN

3.1 TIPO DE INVESTIGACIÓN

El estudio es de tipo exploratorio y descriptivo. A través de una data histórica de clientes para un producto crediticio, del cual ya se conocen los clientes morosos y no morosos se pueden buscar patrones de comportamiento de morosidad y de no morosidad para incrementar la precisión al momento de evaluar el crédito para un nuevo cliente.

3.2 FORMULACIÓN DE LA HIPÓTESIS

Para el centro de investigación seleccionado en el trabajo, un producto crediticio de una banca de consumo, se ha definido la siguiente hipótesis: “que usando el algoritmo CART se determinará las características de la morosidad reduciendo los índices de morosidad para un producto crediticio en una banca nacional”.

3.3 IDENTIFICACION DE LAS VARIABLES

Se realizó un estudio en el período comprendido entre Agosto del 2001 y Junio del 2002 de todas las solicitudes aprobadas del producto crediticio. Se tomaron las variables almacenadas en la base de datos de la entidad bancaria y que fueron recogidas a través de los formularios de solicitud del crédito, y se seleccionaron para participar aquellas variables que tenían mayor poder predictivo. En el Cuadro Nº 2 se muestra un listado y descripción de las mismas:

Cuadro Nº 2: Descripción de las variables en estudio

Nombre de la variable	Descripción
MORA	Condición de moroso del cliente con la entidad bancaria. La entidad bancaria considera moroso a clientes con más de 90 días de atraso en una cuota. Esta variable (MORA) constituye la variable criterio.
EDAD	Edad (en años) del cliente.
SEXO	Sexo del cliente.
ESTCIVIL	Estado civil del cliente.
CARGAS	Carga familiar del cliente. Número de personas dependientes del cliente.
FONPAR	Si el cliente posee o no teléfono particular.
FONLAB	Si el cliente posee o no teléfono en su centro laboral.
CONAUTOV	Si presenta o no autovaluo al momento de solicitar el crédito.
ESAVAL	Si el cliente es aval o no de otro cliente de la entidad bancaria.
ANTLAB	La antigüedad laboral del cliente en años.
TIPREN	Si la renta del cliente es fija o variable.
UBIGEO	La ubicación geográfica (ciudad del país) desde donde se aprobó el crédito al cliente.

Fuente: elaboración propia

3.4 DESCRIPCIÓN DEL PRODUCTO CREDITICIO APLICADO

Para aplicar el algoritmo de árbol de clasificación CART se ha trabajado con un producto crediticio del sistema financiero nacional que otorga préstamos en dólares a personas naturales.

El objetivo del producto crediticio consiste en otorgar líneas de crédito a todas aquellas personas naturales dependientes que cumplan con los requisitos establecidos

"Programa Cybertesis PERÚ - Derechos son del Autor"

por la entidad bancaria.

El producto crediticio se ofreció a personas dependientes de empresas o instituciones privadas y gubernamentales que se encontraron en situación de activos (empleados, obreros, etc.) o jubilados cuyos ingresos mensuales individuales sean mayores a US $ 700.00 o su equivalente en moneda nacional.

La colocación de este producto se realizó a través de las captadoras de préstamos de las agencias y fuerzas de venta del banco.

Los montos de préstamo mínimo y máximo se otorgaron de acuerdo al Cuadro Nº 3.

Cuadro Nº 3: Montos de préstamo mínimo y máximo

Presenta fiador solidario	Préstamo mínimo	Préstamo máximo
No	US $ 500,00	US $ 4 000,00
Si	US $ 4 001,00	US $ 7 000,00

Fuente: elaboración propia

El plazo de devolución del préstamo estuvo dentro del rango indicado en el Cuadro Nº 4.

Cuadro Nº 4: Plazo del préstamo

Plazo mínimo	Plazo máximo
6 meses	36 meses

Fuente: elaboración propia

Los requisitos para el otorgamiento de un crédito fueron:

- Ser persona natural y dependiente de una institución privada y/o gubernamental (activo o jubilado).
- Edad mínima de 25 y máxima de 65 años.
- Continuidad laboral mínima de 12 meses para los activos.
- Ingreso mínimo líquido mensual de $ US 700,00 o su equivalente en moneda nacional (se admite complemento de renta siempre que el cónyuge sea dependiente y cumpla los requisitos establecidos en este producto).
- El fiador solidario podrá ser persona natural dependiente, profesional independiente o comerciante independiente y cumplir los requisitos establecidos en la pauta que le corresponda.
- La empresa o institución de la que sean dependientes los participantes de la operación, deberán estar registradas en la base de datos de la entidad bancaria.

Las condiciones especiales para otorgar créditos en dólares son las siguientes:

- Los solicitantes menores de 30 años, deberán presentar un fiador solidario (de preferencia familiar) que sea mayor de 30 años y que tenga por lo menos 1 año de antigüedad laboral.
- El fiador solidario deberá ganar igual o más que el solicitante del crédito.
- Si el préstamo es menor a US $ 4 000,00 y el solicitante es mayor a 30 años y su empresa está en amarillo no será necesario la presentación del fiador solidario. El color califica el estado financiero de la empresa donde el cliente labora. El color amarillo se aplica a empresas que presentan dificultades y cuyos trabajadores deben presentar un aval, en cambio el color verde de la empresa es para empresas solventes cuyos trabajadores no necesitan presentar aval.

Las restricciones para otorgar el producto:

- Las que se indican en las pautas generales sobre los productos de la entidad bancaria.
- Solicitante y fiador solidario no deberán pertenecer a la misma empresa, si esta se encuentra clasificada en amarillo.

La documentación a reunir por el cliente o solicitante que presentó una solicitud son:

- Formularios del banco
  - Solicitud de línea de crédito.
  - Contrato de agencia y uso de línea de crédito.
  - Pagaré.

- Documentos personales del titular
  - Fotocopia de documento de identidad (y del cónyuge de ser el caso).
  - Fotocopia del último recibo de luz o agua de su domicilio cancelado del mes o mes anterior a la presentación de la solicitud del crédito.
  - Fotocopia de autovaluo del año o año anterior o ficha registral del inmueble con un plazo máximo de 5 días útiles de antigüedad a la fecha de presentación de la solicitud de crédito (si solicita línea de crédito sobre los US $ 5 000,00).

3.5 DISTRIBUCIÓN DE DATOS EN EL PRODUCTO CREDITICIO APLICADO

El objetivo de la investigación es analizar el comportamiento de los morosos y de los no morosos basándose en las variables predictoras.

"Programa Cybertesis PERÚ - Derechos son del Autor"

En el Cuadro Nº 5 se presenta la distribución de todos los clientes que recibieron el crédito del producto crediticio durante el período Agosto del 2001 a Junio del 2002.

Cuadro Nº 5: Distribución de los clientes según la morosidad y no morosidad

Categorías de la variable: MORA	No. de clientes	%
Moroso	32	4,42%
No Moroso	692	95,58%
Total	724	100,00%

Fuente: elaboración propia

En el Cuadro Nro. 5 se aprecia que el nivel de morosidad del producto crediticio es del 4,42% mientras que el nivel de no morosidad es del 95,58%.

3.6 METODOLOGÍA APLICADA

Para probar la hipótesis, se aplicó la siguiente metodología de investigación:

1. Seleccionar las variables predictoras más importantes.

2. Descripción univariada y bivariada de las variables descriptivas y de la variable criterio (morosidad).

3. Aplicación del algoritmo de árbol de clasificación CART.

4. Determinación de la pertinencia de las variables predictoras.

5. Obtención de las reglas de clasificación de la morosidad y de la no morosidad.

Serán considerados clientes morosos a aquellos clientes que habiendo obtenido un crédito no han realizado pago alguno por un período de 90 días o más.

3.7 POBLACIÓN Y MUESTRA

No se va a trabajar con ninguna muestra sino con la base de datos histórica de todos los clientes que accedieron al producto crediticio.

"Programa Cybertesis PERÚ - Derechos son del Autor"

IV PROCEDIMIENTO DE LA TÉCNICA DEL ÁRBOL DE CLASIFICACIÓN CART

4.1 DESCRIPCIÓN DEL SOFTWARE UTILIZADO

Para el procesamiento de la información se trabajó con la información de todos los clientes que obtuvieron el producto crediticio. El trabajo se realizó en dos etapas:

Etapa I: Descripción univariada y bivariada de las variables descriptivas y de la variable criterio, para lo cual se utilizó el software estadístico SPSS ver 10.0.

Etapa II: Aplicación del algoritmo de árbol de clasificación, para lo cual se utilizó el software CART 4.0 for Windows de la compañía Salford Systems [7]; para generar los resultados de la clasificación usando como medida de impureza el índice de Gini con costos de mala clasificación iguales.

4.2 ANÁLISIS DESCRIPTIVO UNIVARIADO Y BIVARIADO

4.2.1 ANÁLISIS DESCRIPTIVO UNIVARIADO

Antes de aplicar el algoritmo CART, se debe explorar previamente la información de la matriz de datos. El análisis univariado va a permitir describir a los clientes que solicitaron el crédito en función a cada una de las variables predictoras así como en función a la variable criterio de morosidad (MORA).

En el Cuadro Nº 5 se observa que el producto crediticio durante el período de evaluación (Agosto 2001- Junio 2002) ha dado como resultado un porcentaje de 4,4% de morosidad, lo que constituye 32 clientes de un total de 724 clientes que solicitaron el crédito.

Cuadro Nº 6: Distribución de los clientes según las variables predictoras cuantitativas

	Nro.	Mínimo	Máximo	Media
Edad	724	25	66	42,31
Carga Familiar	724	0	4	1,45
Antigüedad Laboral	724	1	40	9,90

Fuente: elaboración propia

En el Cuadro Nº 6 se realiza el análisis univariado de las variables predictoras cuantitativas. Se obtienen los siguientes resultados:

- La edad de los clientes que obtuvieron el crédito varía de 25 años hasta 66 años con una edad promedio de 42,31 años.
- La carga familiar de los clientes que obtuvieron el crédito varía de 0 (no tiene ningún compromiso ni hijos) a 4 con una carga familiar promedio de 1,45.
- La antigüedad de los clientes que obtuvieron el crédito varía de 1 año hasta 40 años con una antigüedad promedio de 9,9 años.

Cuadro Nº 7: Distribución de los clientes según las variables predictoras cualitativas

"Programa Cybertesis PERÚ - Derechos son del Autor"

		Nro. de clientes	%
Sexo	Femenino	147	20,3
	Masculino	577	79,7
Estado Civil	Casado	460	63,5
	Divorciado	10	1,4
	Soltero	242	33,4
	Viudo	12	1,7
¿Posee teléfono particular?	No	128	17,7
	Si	596	82,3
¿Posee teléfono en el trabajo?	No	47	6,5
	Si	677	93,5
¿Presenta autovaluo?	No	617	85,2
	Si	107	14,8
¿Es aval?	No	535	73,9
	Si	189	26,1
Tipo de renta	Fija	692	95,6
	Variable	32	4,4
Ubigeo	Ancash	3	0,4
	Apurimac	1	0,1
	Arequipa	60	8,3
	Ayacucho	5	0,7
	Cajamarca	2	0,3
	Callao	28	3,9
	Cuzco	8	1,1
	Huanuco	16	2,2
	Ica	7	1,0
	Junín	8	1,1
	La Libertad	40	5,5
	Lambayeque	21	2,9
	Lima	277	38,3
	Loreto	56	7,7
	Moquegua	13	1,8
	Piura	97	13,4
	Puno	14	1,9
	San Martín	16	2,2
	Tacna	42	5,8
	Tumbes	4	0,6
	Ucayali	6	0,8
Total		724	100

Fuente: elaboración propia

En el Cuadro Nº 7 se realiza el análisis univariado de las variables predictoras

cualitativas. Se obtienen los siguientes resultados:

- Los clientes que obtuvieron el crédito están constituidos en su mayoría (79,7%) por personas del sexo masculino
- Los clientes que obtuvieron el crédito están constituidos en su mayoría (63,5%) por personas casadas.
- Los clientes que recibieron el crédito, en su mayoría (82,3%) poseen teléfono particular.
- Los clientes que recibieron el crédito, en su mayoría (93,5%) poseen teléfono en su centro de trabajo. Esto indicaría que esta variable a pesar de su importancia no discriminaría la situación de moroso y no moroso.
- La mayoría de los clientes (85,2%) que obtuvieron el crédito han presentado el autovaluo al momento de presentar la solicitud.
- Los clientes que obtuvieron el crédito, en su mayoría (73,9%) son avales de otros clientes de la entidad crediticia.
- Casi todos los clientes (95,6%) que obtuvieron el crédito tenían renta fija. Esto indicaría que esta variable a pesar de su importancia no discriminaría la situación de moroso y no moroso.
- La mayoría de clientes que obtuvieron el crédito provienen de Lima (38,3%).

4.2.2 ANÁLISIS DESCRIPTIVO BIVARIADO

Se efectuó un análisis bivariado de cada una de las variables predictoras con la variable criterio (MORA). Estos análisis previos permiten observar las características de los morosos y no morosos por cada variable predictora.

Cuadro Nº 8: Distribución de los clientes según la morosidad para las variables predictoras cuantitativas

	Morosidad				**Total**	
	Moroso		No moroso		Nro.	Media
	Nro.	Media	Nro.	Media	Nro.	Media
Edad	32	39,59	692	42,43	724	42,31
Carga Familiar	32	,69	692	1,49	724	1,45
Antigüedad Laboral	32	5,97	692	10,08	724	9,90

Fuente: elaboración propia

En el Cuadro Nº 8 se realiza el análisis bivariado de las variables predictoras cuantitativas según la morosidad. Se obtienen los siguientes resultados:

- No se puede afirmar que la edad influya sobre la morosidad o no morosidad. Se observa que la edad promedio de los clientes morosos y no morosos es similar (39,59 y 42,43 años).

"Programa Cybertesis PERÚ - Derechos son del Autor"

- La carga familiar de los morosos es de 0,69 personas, mientras que de los no morosos es de 1,49 personas.
- Los clientes no morosos tienen una mayor antigüedad laboral que los morosos. Se observa que la antigüedad promedio de los morosos es de 6 años, mientras que de los no morosos es de 10 años.

Cuadro Nº 9: Distribución de los clientes según la morosidad para las variables predictoras cualitativas

MorosoNo morosoTotal		Morosid					
		Nro.	%	Nro.	%	Nro.	%
Sexo	Femenino	7	4,8	140	95,2	147	100
	Masculino	25	4,3	552	95,7	577	100
Estado Civil	Casado	10	2,2	450	97,8	460	100
	Divorciado			10	100,0	10	100
	Soltero	22	9,1	220	90,9	242	100
	Viudo			12	100,0	12	100
¿Posee teléfono particular?	No	11	8,6	117	91,4	128	100
	Si	21	3,5	575	96,5	596	100
¿Posee teléfono en el trabajo?	No	2	4,3	45	95,7	47	100
	Si	30	4,4	647	95,6	677	100
¿Presenta autovaluo?	No	32	5,2	585	94,8	617	100
	Si			107	100,0	107	100
¿Es aval?	No	28	5,2	507	94,8	535	100
	Si	4	2,1	185	97,9	189	100
Tipo de renta	Fija	30	4,3	662	95,7	692	100
	Variable	2	6,3	30	93,8	32	100
Ubigeo	Ancash			3	100,0	3	100
	Apurimac			1	100,0	1	100
	Arequipa	3	5,0	57	95,0	60	100
	Ayacucho			5	100,0	5	100
	Cajamarca			2	100,0	2	100
	Callao	3	10,7	25	89,3	28	100
	Cuzco			8	100,0	8	100
	Huanuco			16	100,0	16	100
	Ica			7	100,0	7	100
	Junín	2	25,0	6	75,0	8	100
	La Libertad	3	7,5	37	92,5	40	100
	Lambayeque			21	100,0	21	100
	Lima	18	6,5	259	93,5	277	100
	Loreto	1	1,8	55	98,2	56	100
	Moquegua			13	100,0	13	100
	Piura			97	100,0	97	100
	Puno			14	100,0	14	100
	San Martin			16	100,0	16	100
	Tacna	2	4,8	40	95,2	42	100
	Tumbes			4	100,0	4	100
	Ucayali			6	100,0	6	100
Total		32	4,4	692	95,6	724	100

Fuente: elaboración propia

En el Cuadro Nº 9 se realiza el análisis bivariado de las variables predictoras

"Programa Cybertesis PERÚ - Derechos son del Autor"

cualitativas según la morosidad. Se obtienen los siguientes resultados:

- No se puede decir que los hombres o las mujeres son más morosos. Se observa que el porcentaje de morosidad es similar en ambos sexos (4,8% y 4,3%).
- Los solteros son más morosos que cualquier otra categoría del estado civil. Se observa que el porcentaje de morosidad de los solteros es el mayor de todos los estados civiles considerados (9,1%).
- Los clientes que no poseen teléfono particular son más morosos que los que si lo poseen. Se observa que su porcentaje de morosidad es el mayor (8,6% vs. un 3,5%).
- No se puede decir que los clientes que poseen teléfono en su trabajo o los que no lo poseen son más morosos. Se observa que el porcentaje de morosidad es similar en ambos casos (4,3% y 4,4%).
- Los clientes que no presentaron autovaluo al momento de solicitar el crédito son más morosos que los que si lo presentaron. Se observa que su porcentaje de morosidad es el mayor (5,2% vs. un 0,0%).
- No se puede afirmar que los clientes con renta fija sean más morosos que los que poseen renta variable. Se observa que el porcentaje de morosidad es similar en ambos casos (4,3% y 6,3%).
- La morosidad está muy relacionada con el ubigeo donde se otorgó el crédito. En el caso de los morosos Junín posee la tasa más elevada de morosidad (25,0%).

4.3. ANÁLISIS DEL ÁRBOL GENERADO POR EL ALGORITMO CART

Usando el software CART (ver manual en el Anexo No. 1) y usando la variable MORA como variable criterio y a las demás variables como variables predictoras se obtiene el árbol de clasificación presentado en el Gráfico Nº 5

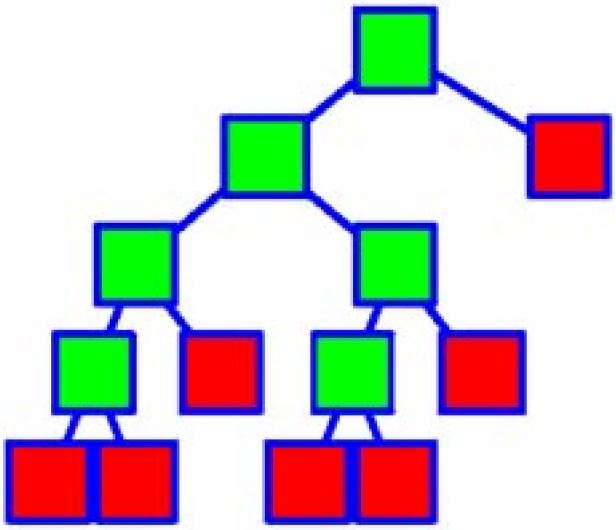


Gráfico Nº 5: Árbol completo generado por el algoritmo CART

El costo relativo con que se obtuvo el árbol fue de 0,554 que corresponde a 7 nodos, tal como puede apreciarse en el Gráfico Nº 6.

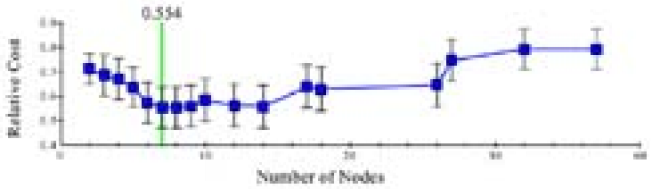


Gráfico Nº 6: Costo Relativo vs. Número de Nodos

Antes de empezar a analizar el árbol de clasificación obtenido, se debe evaluar la pertinencia o no de cada variable predictora. El Cuadro Nº 10 proporciona la importancia de cada una de las variables predictoras. Un score es asignado a cada variable, y está basado en la mejora que hace en la división del árbol. El score da un reconocimiento a las variables cuya importancia es marcada.

Cuadro Nº 10 Importancia de cada variable predictora

"Programa Cybertesis PERÚ - Derechos son del Autor"

Variable	Score	
UBIGEO	100,00						

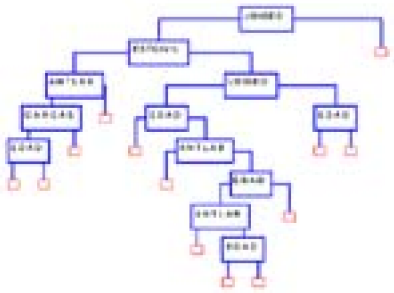


Gráfico Nº 8: Árbol de clasificación con las variables predictoras que generaron el árbol

Del Gráfico Nº 8 se pueden obtener los siguientes resultados:

- Es un árbol con 13 nodos terminales, mostrados de un color diferente (rectángulo rojo pequeño) de los nodos parentales (rectángulo azul grande).
- La primera variable predictora que va a discriminar a los morosos y los no morosos es el ubigeo o ciudad donde el cliente solicitó el crédito (UBIGEO).
- Las demás variables predictoras que van a discriminar a los morosos y no morosos son: el estado civil (ESTCIVIL), la antigüedad laboral (ANTLAB), la carga familiar (CARGAS) y la edad del cliente (EDAD).

Un árbol más disgregado que presenta además el número de casos que hay en cada nodo parental y el número de casos que hay en cada nodo terminal se presenta en el Gráfico Nº 9.



Gráfico Nº 9: Árbol de clasificación con las variables predictoras que generaron el árbol y el número de casos por nodo

4.4 ANÁLISIS PARA LOS MOROSOS Y NO MOROSOS

4.4.1 ANÁLISIS PARA LOS MOROSOS

En el Gráfico Nº 10 se puede apreciar el árbol de clasificación para los clientes que fueron morosos. En dicho gráfico los nodos terminales de color rojo intenso indican una mejor clasificación que los nodos terminales de color rojo claro.

"Programa Cybertesis PERÚ - Derechos son del Autor"

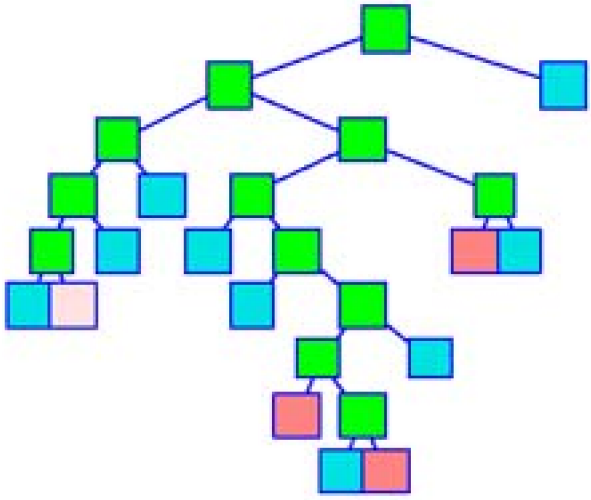


Gráfico Nº 10: Árbol de Clasificación para los morosos

CUADRO DE GANANCIAS PARA LOS MOROSOS

Cuando la variable criterio es categórica, el Cuadro Nº 11 proporciona estadísticos de los nodos terminales que describen el árbol de clasificación respecto a la categoría criterio de la variable criterio (MORA) para los morosos (MORA=0).

Cuadro Nº 11: Cuadro de ganancias para los morosos (MORA = 0)

Node	Cases Class 0	% of Node Class 0	% Class 0	Cum. % Class 0	Cum. % Pop	% Pop	Cases in Node	Cum. Lift	Lift Index
7	5	50,000	15,63	15,63	1,38	1,38	10	11,313	11,313
9	2	33,333	6,25	21,88	2,21	0,83	6	9,898	7,542
11	1	33,333	3,13	25,00	2,62	0,41	3	9,526	7,542
2	22	20,952	68,75	93,75	17,13	14,50	105	5,474	4,740
12	1	0,565	3,13	96,88	41,57	24,45	177	2,330	0,128
13	1	0,372	3,13	100,00	78,73	37,15	269	1,270	0,084
5	0	0,000	0,00	100,00	84,39	5,66	41	1,185	0,000
4	0	0,000	0,00	100,00	88,95	4,56	33	1,124	0,000
3	0	0,000	0,00	100,00	91,71	2,76	20	1,090	0,000
6	0	0,000	0,00	100,00	94,34	2,62	19	1,060	0,000
8	0	0,000	0,00	100,00	96,69	2,35	17	1,034	0,000
1	0	0,000	0,00	100,00	98,34	1,66	12	1,017	0,000
10	0	0,000	0,00	100,00	100,00	1,66	12	1,000	0,000

Fuente: elaboración propia

El cuadro de ganancias para los morosos muestra la siguiente información para cada nodo terminal:

Node	Número de referencia del nodo terminal.
Cases Class 0	Número de casos en el nodo que son de la Class 0.
% of Node Class 0	% de los casos en el nodo que son de la Class 0.
% Class 0	Número de casos de la Class 0 en el nodo como % del total de casos de la Class 0.
Cum % Class 0	Número acumulativo de casos de la Class 0 como % del número total de casos de la Class 0.
Cum % Pop	Número acumulativo de casos como % del número total de casos en el análisis.
% Pop	% del número total de casos en el análisis que están contenidos en el nodo.
Cases in Node	Número total de casos en el nodo.
Cum. Lift	% acumulativo de Class 0 dividido por el acumulativo del número total de casos compartido.
Lift Index	% de la Class 0 en el nodo dividido por el % del número total de casos en el nodo.

Esto significa que todos los clientes morosos han sido clasificados en los nodos terminales 7, 9, 11, 2, 12 y 13 (Cum. % Class 0 igual a 100%). También se observa en el Cuadro Nº 11 que el mayor porcentaje de clientes morosos se han obtenido en los nodos terminales 7 (15,63% de los morosos) y 2 (68,75%), por lo que se puede usar las reglas de decisión de estos nodos para explicar la clasificación de los morosos. El Gráfico Nº 11 muestra el ploteo del porcentaje del total de casos de los morosos para cada nodo terminal.

"Programa Cybertesis PERÚ - Derechos son del Autor"

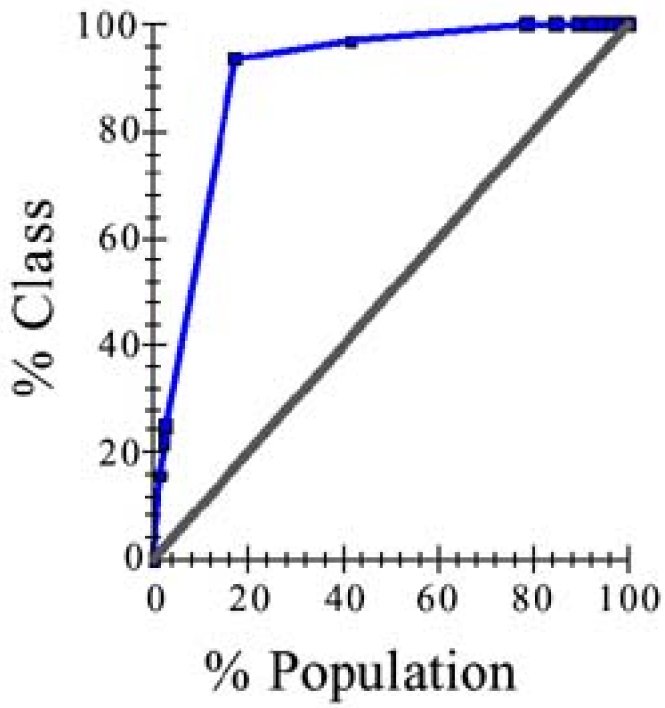


Gráfico Nº 11: Ganancias para los morosos

Según lo apreciado en la Cuadro Nº 11 y el Gráfico Nº 11, el total de casos de dicha categoría se alcanza recién en el nodo terminal 13; sin embargo, con el nodo terminal 7 y 2 se tiene explicado la clasificación del 84,38% (15,63% y 68,75%) del total de los morosos.

4.4.2 ANÁLISIS PARA LOS NO MOROSOS

En el Gráfico Nº 12 se puede apreciar el árbol de clasificación pero para los clientes que fueron morosos. En dicho gráfico los nodos terminales son todos de color rojo intenso lo que indica que todos ellos hacen una buena clasificación de los no morosos.

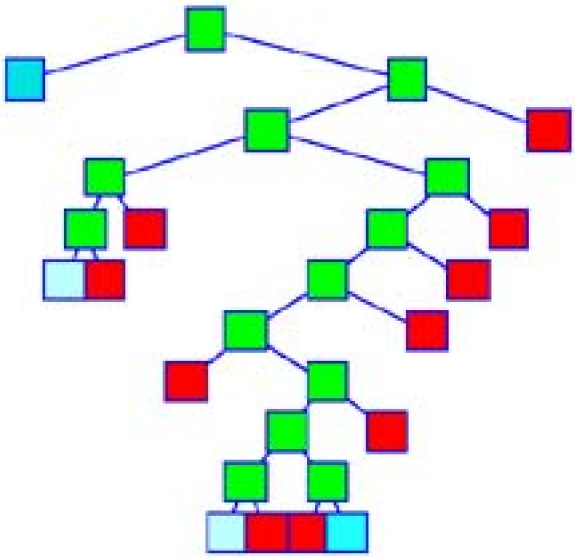


Gráfico Nº 12: Árbol de Clasificación para los no morosos

CUADRO DE GANANCIAS PARA LOS NO MOROSOS

El Cuadro Nº 12 proporciona estadísticos de los nodos terminales que describen el árbol de clasificación respecto a la categoría criterio de la variable criterio (MORA) para los no morosos (MORA=1).

Cuadro Nº 12: Tabla de ganancias para los no morosos (MORA=1)

Node	Cases Class 1	% of Node Class 1	% Class 1	Cum. % Class 1	Cum. % Pop	% Pop	Cases in Node	Cum. Lift	Lift Index
5	41	100,000	5,92	5,92	5,66	5,66	41	1,046	1,046
4	33	100,000	4,77	10,69	10,22	4,56	33	1,046	1,046
3	20	100,000	2,89	13,58	12,98	2,76	20	1,046	1,046
6	19	100,000	2,75	16,33	15,61	2,62	19	1,046	1,046
8	17	100,000	2,46	18,79	17,96	2,35	17	1,046	1,046
1	12	100,000	1,73	20,52	19,61	1,66	12	1,046	1,046
10	12	100,000	1,73	22,25	21,27	1,66	12	1,046	1,046
13	268	99,628	38,73	60,98	58,43	37,15	269	1,044	1,042
12	176	99,435	25,43	86,42	82,87	24,45	177	1,043	1,040
2	83	79,048	11,99	98,41	97,38	14,50	105	1,011	0,827
9	4	66,667	0,58	98,99	98,20	0,83	6	1,008	0,697
11	2	66,667	0,29	99,28	98,62	0,41	3	1,007	0,697
7	5	50,000	0,72	100,00	100,00	1,38	10	1,000	0,523

Fuente: elaboración propia

Esto significa que los clientes no morosos han sido clasificados en todos los nodos terminales (% of Node Class 1 igual a 100%). También se observa en el Cuadro Nº 12

"Programa Cybertesis PERÚ - Derechos son del Autor"

que el mayor porcentaje de clientes morosos se han obtenido en los nodos terminales 13 (38,73% de los no morosos) y 12 (25,43%), por lo que se puede usar las reglas de decisión de estos nodos para explicar la clasificación de no los morosos. El Gráfico Nº 13 muestra el ploteo del porcentaje del total de casos de los no morosos para cada nodo terminal.

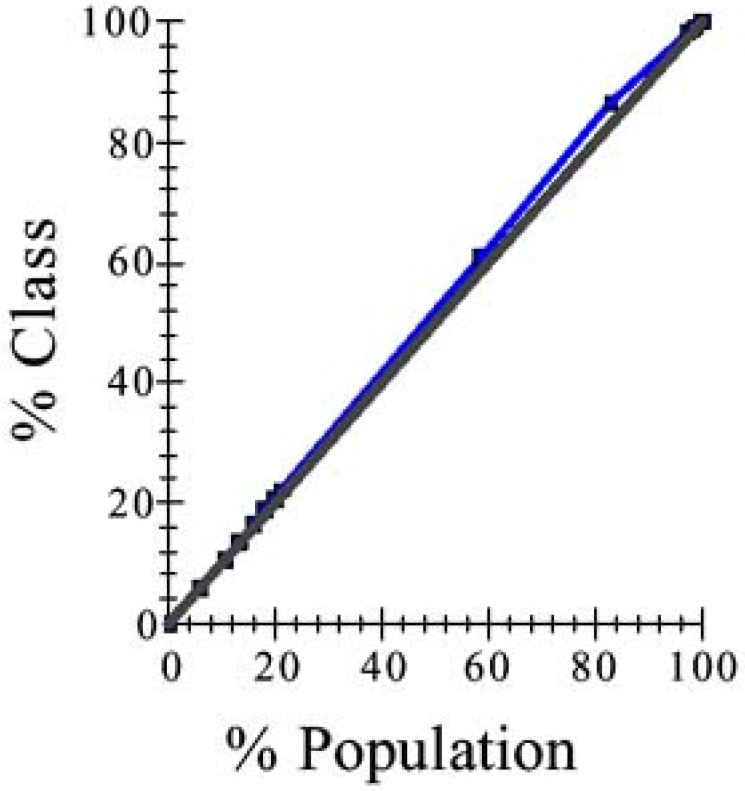


Gráfico Nº 13: Ganancias para los no morosos

Según lo apreciado en la Cuadro Nº 12 y el Gráfico Nº 13, el total de casos de dicha categoría se alcanza recién en el último nodo terminal; sin embargo, con los nodos terminal 13 y 12 se tiene explicado la clasificación del 64,16% del total de los no morosos.

4.5 IMPORTANCIA DE LAS VARIABLES PARA EL PRODUCTO CREDITICIO

El Cuadro Nº 13 proporciona la importancia de todas las variables predictoras. Un score es asignado a cada variable, y está basado en la mejora que hace en la división del árbol. El score da un reconocimiento a las variables cuya importancia es marcada.

Cuadro Nº 13: Importancia de cada variable

Variable	Score	
UBIGEO	100,00						

V. RESULTADOS

5.1 REGLAS DE CLASIFICACIÓN PARA LOS MOROSOS

Las características o el patrón que poseen los morosos se pueden apreciar en el Anexo No. 2 y se resumen en lo siguiente:

- Nodo 2: Que él estado civil del cliente sea soltero y que el crédito lo haya obtenido en Arequipa, Callao, Junín, La Libertad, Lima o Tacna y la antigüedad laboral sea menor a 11 años y que la carga familiar sea menor de 2 hijos y que la edad del cliente sea mayor de 28 años. (representa el 68,75% de los clientes morosos); ó
- Nodo 7: Que el crédito lo haya obtenido en Arequipa, Junín, La Libertad o Tacna y que el estado civil del cliente sea casado, divorciado o viudo y que la edad del cliente sea mayor a 42 años pero menor a 54 años y que la antigüedad laboral sea mayor a 6 años pero menor de 8 años. (representa el 22,73% de los clientes morosos); ó
- Nodo 9: Que el crédito lo haya obtenido en Arequipa, Junín, La Libertad o Tacna y que el estado civil del cliente sea casado, divorciado o viudo y que la edad del cliente sea mayor a 49 años pero menor de 55 años y que la antigüedad laboral sea mayor a 8 años. (representa el 6,25% de los clientes morosos); ó

- Nodo 11: Que el crédito lo haya obtenido en Callao ó Lima y que el estado civil del cliente sea casado, divorciado o viudo y que la edad del cliente sea menor de 29 años. (representa el 3,13% de los clientes morosos)

5.2 REGLAS DE CLASIFICACIÓN PARA LOS NO MOROSOS

Las características o el patrón que poseen los no morosos se pueden apreciar en el Anexo No. 3 y se resumen en lo siguiente:

- Nodo 13: Que el lugar donde obtuvo el crédito sea Ancash, Apurímac, Ayacucho, Cajamarca, Cuzco, Huanuco, Ica, Lambayeque, Loreto, Moquegua, Piura, Puno, San Martín, Tumbes ó Ucayali (representa el 38,73 % de los clientes no morosos); ó
- Nodo 12: Que el lugar donde obtuvo el crédito sea Callao o Lima y que el estado civil sea casado, divorciado o viudo y que la edad sea mayor de 29 años (representa el 25,32% de los clientes no morosos)
- Nodo 5: que el estado civil del cliente sea casado, divorciado ó viudo y que el lugar donde obtuvo el crédito sea Arequipa, Junín, La Libertad ó Tacna y que la edad del cliente sea menor de 42 años. (representa el 5,92% de los clientes no morosos); ó
- Nodo 4: que el estado civil del cliente sea soltero y que el lugar donde obtuvo el crédito sea Arequipa, Callao, Junín, La Libertad, Lima ó Tacna y que la antigüedad laboral sea mayor de 10 años. (representa el 4,77% de los clientes no morosos); ó
- Nodo 3: que el estado civil del cliente sea soltero y que el lugar donde obtuvo el crédito sea Arequipa, Callao, Junín, La Libertad, Lima ó Tacna y que la antigüedad laboral sea menor de 11 años y que la carga familiar sea mayor de 1 hijo. (representa el 2,89% de los clientes no morosos); ó
- Nodo 6: que el estado civil del cliente sea casado, divorciado ó viudo y que el lugar donde obtuvo el crédito sea Arequipa, Junín, La Libertad ó Tacna y que la edad del cliente sea mayor de 41 años y la antigüedad laboral del cliente sea menor de 6 años. (representa el 2,75% de los clientes no morosos); ó
- Nodo 8: que el estado civil del cliente sea casado, divorciado ó viudo y que el lugar donde obtuvo el crédito sea Arequipa, Junín, La Libertad ó Tacna y que la edad del cliente sea mayor de 41 años pero menor de 50 años y la antigüedad laboral del cliente sea mayor de 8 años. (representa el 2,46% de los clientes no morosos); ó
- Nodo 1: Que el estado civil sea soltero y que el lugar donde obtuvo el crédito sea Arequipa, Callao, Junín, La Libertad, Lima ó Tacna y que la antigüedad laboral sea menor de 11 años y que la carga familiar sea menor de 2 hijos y que la edad del cliente sea menor de 29 años. (representa el 1,73% de los clientes no morosos); ó
- Nodo 10: que el estado civil del cliente sea casado, divorciado ó viudo y que el lugar

"Programa Cybertesis PERÚ - Derechos son del Autor"

donde obtuvo el crédito sea Arequipa, Junín, La Libertad ó Tacna y que la edad del cliente sea mayor de 54 años y la antigüedad laboral del cliente sea mayor de 5 años. (representa el 1,73% de los clientes no morosos)

5.3 VALIDACIÓN DE LOS RESULTADOS

Para poder evaluar la capacidad del algoritmo CART a la hora de predecir la morosidad o no morosidad se puede examinar el Cuadro Nº 14 que compara el tipo asignado mediante el árbol con la morosidad/no morosidad realmente registrada. Se observa que la proporción de casos clasificados de manera incorrecta es bastante baja: en total 2 de 32 morosos y 94 de 692 no morosos. Esto da como resultado un porcentaje de clasificación correcta del 86,74% de todos los clientes que solicitaron el crédito, lo que indica que el algoritmo CART obtenido tiene poder predictivo.

Cuadro Nº 14: Cuadro de clasificación errónea

Clase	Número de casos	Número de casos mal clasificados	% de clasificación errónea	Número de casos correctamente clasificados	% de clasificación correcta
Moroso	32	2	6,25	30	93,75
No Moroso	692	94	13,58	598	86,42
Total	724	96	13,26	628	86,74

Fuente: elaboración propia

En el Cuadro Nº 15 se presentan algunos casos de clientes que solicitan el crédito y la predicción de su situación (moroso o no moroso) según el algoritmo CART desarrollado.

Cuadro Nº 15: Cuadro de simulación de predicciones

Eda	Estado civil	Carga familiar	Antigüedad laboral	Ubigeo	Nod	Predicción según CART
33	Casado	2	5	Lima	12	No Moroso
44	Divorciado	1	11	Lima	12	No Moroso
32	Casado	1	5	Lima	12	No Moroso
43	Soltero	1	7	Callao	2	Moroso
37	Soltero	0	10	Lima	2	Moroso
36	Soltero	1	2	Ucayali	13	No Moroso
54	Viudo	0	22	Arequipa	8	No Moroso
34	Soltero	1	2	Lima	2	Moroso
33	Soltero	0	4	Lima	2	Moroso
36	Soltero	0	7	Tacna	2	Moroso
36	Soltero	0	9	Lima	2	Moroso
54	Casado	0	29	Lambayeque	13	No Moroso
44	Casado	3	7	Tacna	7	Moroso

Fuente: elaboración propia

5.4 DISCUSIÓN DE RESULTADOS

El análisis descriptivo univariado a permitido describir a los clientes que solicitaron el crédito en función a cada una de las variables predictoras: edad, sexo, estado civil, carga familiar, posesión de teléfono particular, posesión de teléfono laboral, si presenta o no autovaluo, si el cliente es aval o no, la antigüedad laboral, el tipo de renta y el ubigeo.

Para poder encontrar en forma exploratoria y descriptiva un posible patrón se optó por un análisis descriptivo bivariado. Este análisis permitió observar las características de los morosos y de los no morosos para cada variable predictora. La desventaja de este análisis es que solamente realizó un análisis por cada variable predictora y no en forma conjunta. Se encontró que las variables que discriminaban la morosidad son: la carga familiar, la antigüedad laboral, el estado civil, la posesión o no de teléfono particular, si presenta o no autovaluo y el ubigeo.

Para poder encontrar un patrón que analice las interrelaciones entre las variables predictoras se optó por el algoritmo de árbol de clasificación CART. Con este algoritmo se logró seleccionar previamente a las variables predictoras que discriminaban la morosidad y la no morosidad. Estas variables son: el ubigeo, el estado civil, la antigüedad laboral, la carga familiar y la edad del cliente. Asimismo, el algoritmo CART determinó la importancia de cada una de estas variables al momento de discriminar la morosidad y la no morosidad. Las variables según el orden de importancia son: el ubigeo, el estado civil, la antigüedad laboral, la carga familiar y la edad del cliente

Finalmente, el algoritmo CART encontró las reglas de clasificación para morosos y no morosos, logrando un 86,74% de clasificación correcta.

"Programa Cybertesis PERÚ - Derechos son del Autor"

5.5 MEDIDAS CORRECTIVAS Y CONTROL DE MEJORAS

Los resultados obtenidos con el algoritmo CART parten de la premisa de que las solicitudes de crédito ya han pasado una evaluación estándar. La metodología sólo funciona para este tipo de créditos y en general para este producto crediticio.

El poder predictivo de los resultados del algoritmo CART supone que el desempeño de los créditos futuros tendrá las mismas características o condiciones que los créditos pasados con los que se obtuvieron los resultados del algoritmo CART. Si las condiciones del producto crediticio cambia, el futuro no es como el pasado -como a menudo sucede a las instituciones financieras que desarrollan productos y nichos nuevos, enfrentan la competencia, u operan en mercados cambiantes- entonces los resultados del algoritmo CART no funcionarán bien, y en todo caso se deberá aplicar un nuevo algoritmo CART.

"Programa Cybertesis PERÚ - Derechos son del Autor"

CONCLUSIONES

1. Usando el algoritmo CART se ha conseguido detectar las variables que más influyen sobre la morosidad/no morosidad del producto crediticio y se las aisló de las variables que no influyen. Estas variables son: el ubigeo, la antigüedad laboral, la edad, la carga familiar y el estado civil.

2. Usando el algoritmo CART se detectaron patrones diferentes para los morosos y los no morosos. Para los morosos el principal patrón detectado es que el estado civil del cliente sea soltero y que el crédito lo haya obtenido en Arequipa, Callao, Junín, La Libertad, Lima o Tacna y la antigüedad laboral sea menor a 11 años y que la carga familiar sea menor de 2 hijos y que la edad del cliente sea mayor de 28 años. Este patrón caracteriza al 68,75% de los clientes morosos. Para los no morosos el principal patrón detectado es que el lugar donde obtuvo el crédito sea Ancash, Apurímac, Ayacucho, Cajamarca, Cuzco, Huanuco, Ica, Lambayeque, Loreto, Moquegua, Piura, Puno, San Martín, Tumbes ó Ucayali.

3. Usando el algoritmo CART se ha automatizado el proceso crédito, discriminando con gran precisión si el cliente es sujeto a crédito. Se obtuvo un 93,75% de clasificación correcta para los morosos y un 86,42% de clasificación correcta para los no morosos. En general se logró un 86,74% de clasificación correcta. Además bajo este análisis de sistema de información la banca puede tener proyecciones más seguras y a tiempo real de quienes son sus mejores clientes crediticios.

"Programa Cybertesis PERÚ - Derechos son del Autor"

RECOMENDACIONES

1. Se pueden comparar los resultados obtenidos con el algoritmo CART con los resultados obtenidos con otros algoritmos de árboles de decisión entre los cuales se mencionan el CHAID (detector automático de interacciones de chi-cuadrado), CHAID exhaustivo y el algoritmo QUEST.

2. El algoritmo CART constituye una técnica para explorar datos que está limitado por la falta de un modelo probabilístico, por lo que sus resultados deberían ser confirmados con técnicas estadísticas multivariadas.

3. Para poder realizar una eficiente explotación de la propia cartera de colocaciones, hoy la entidad crediticia debería plantearse la posibilidad, de contar con una herramienta estadística que en base al comportamiento de sus clientes, le permita segmentarlos y de esta forma ofrecer nuevas líneas de productos acorde a sus necesidades.

4. La capacidad de pronóstico aumenta con el número de características disponibles (variables independientes o explicativas) que se usen del formulario de solicitud del crédito.

"Programa Cybertesis PERÚ - Derechos son del Autor"

BIBLIOGRAFÍA

Acuña, E. Curso internacional: “Métodos de análisis discriminante”. Universidad Nacional Agraria La Molina. Departamento de Estadística e Informática. 1999.

Bonson, E. Tecnologías inteligentes para la gestión empresarial. Editorial ra-ma. 1999.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. Belmont, Calif: Wadsworth. 1984.

Hair-Anderson-Tatham-Black. Análisis Multivariante. Prentice Hall. Quinta Edición. 1999.

Johnson, D. E. Métodos multivariados aplicados al análisis de datos. International Thomson Eds. 2000.

Lewis, R. An Introduction to Classification and Regression Tree (CART) Analysis. Annual Meeting of the Society for Academy Emergency Medicine. San Francisco, California. 2000

Salford Systems. Cart for Windows. User’s guide. A Salford systems implementation of the original CART program. 2000.

Soler Ramos, José y Otros. Gestión de Riesgos Financieros. Un enfoque práctico para países latinoamericanos. Banco Interamericano de Desarrollo. Grupo Santander. 1999.

SPSS, Inc. Data Mining with Confidence. 1998.

Tau, J.T and Gonzales, R.C. Pattern Recognition principles. Addison-Wesley. 1974.

Watanabe, S. Pattern Recognition: Human and Mechanical. Wiley. New York. 1985.

"Programa Cybertesis PERÚ - Derechos son del Autor"

ANEXO

Anexo Nº 1

Manejo del Software CART

Tipos de Datos

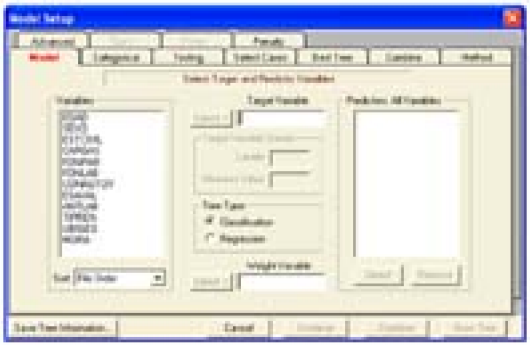
El software CART no posee un editor de datos donde se pueda ingresar datos, en cambio, puede trabajar con datos ingresados en diferentes programas como Excel, Minitab, SPSS, FoxPro, etc. Debe tomarse en cuenta que la información debe ser ingresada en formato numérico y no usar formato caracter, texto o String para ninguna variable. Por ejemplo, a pesar que la variable SEXO es una variable categórica, se la debe codificar previamente e ingresar únicamente valores numéricos como 1 si es masculino y 2 si es femenino, y no ingresar los textos de M y F por ejemplo.

Abriendo un archivo de datos

Para abrir el archivo **BaseCumbre.sav** se selecciona la opción **Open…** del menú File y se selecciona el archivo que contiene los datos a trabajar.

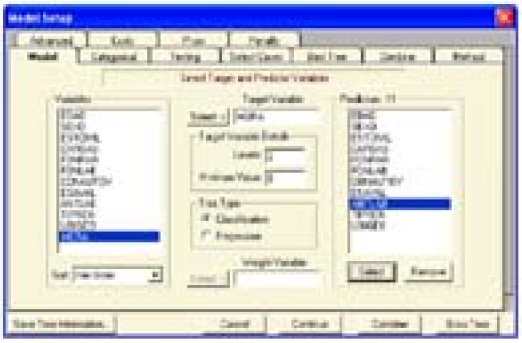


Una vez abierto el archivo BaseCumbre.sav la ventana de diálogo **Model Setup** aparece automáticamente.



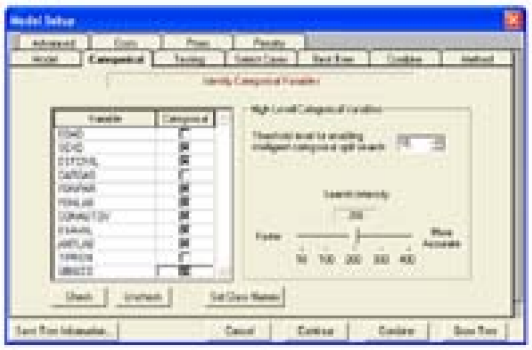
Especificación del modelo

Se debe especificar la variable criterio (**Target Variable**), el número de categorías que tiene y el valor mínimo de la variable criterio (MORA). Asimismo, se deben establecer las variables predictoras. (**Predictors**)



Para las variables predictoras se debe especificar usando la opción de **Categorical** cuales de las variables predictoras se van a considerar como categóricas.

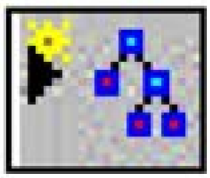
"Programa Cybertesis PERÚ - Derechos son del Autor"

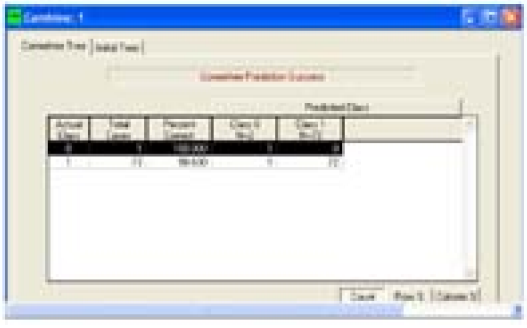


De igual manera, usando la opción de **Set Class Names**, se pueden poner las etiquetas a cada uno de los valores de las variables predictoras categóricas.



Una vez especificado el modelo (variable criterio y variables predictoras) se empieza a realizar el análisis CART usando el botón

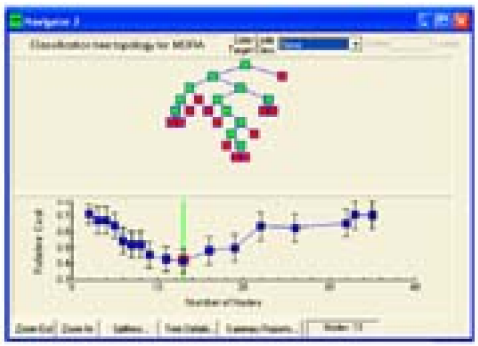




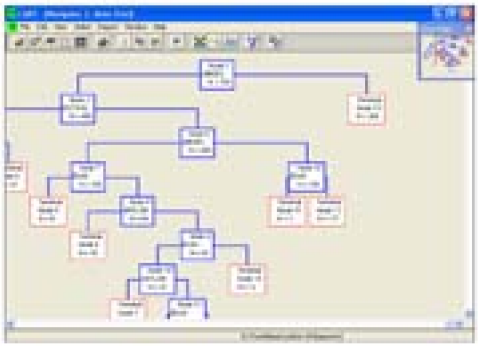
Navegación por el árbol de clasificación

El árbol mostrado por defecto en la ventana de **Navigator**, nos presenta el árbol en forma resumida y la profundidad de dicho árbol. El costo mínimo o el óptimo y el número

de nodos terminales es presentado.



Se puede visualizar el árbol de clasificación completo usando la opción de **Tree Details**

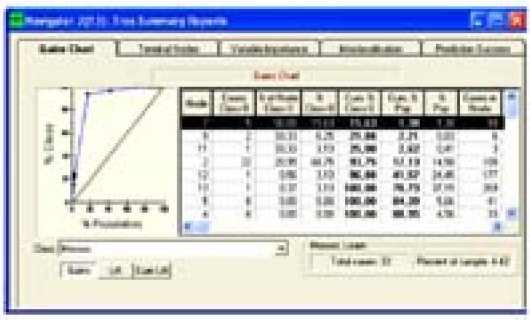


Reportes resumen del árbol

Usando la opción de **Summary Reports** se pueden obtener los siguientes reportes:

- Gráfico de Ganancias.
- La distribución de nodo terminales.
- La medición de la importancia de cada variable predictora.
- Las tablas de clasificación errónea
- La tabla de predicción.

El gráfico de ganancias se puede obtener usando la opción de **Gains Chart**

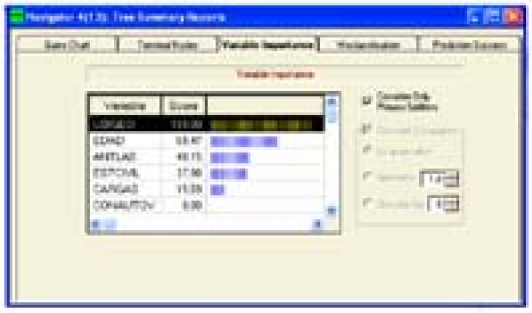


Con la opción de **Terminal Nodes** se puede obtener una representación gráfica de

los nodos terminales.



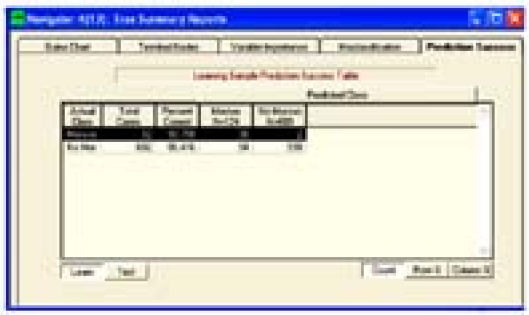
Con la opción de **Variable Importance** se puede obtener un score de la importancia de cada variable predictora. El score representa la contribución de cada variable en la clasificación y predicción de la variable criterio.



La opción de **Misclassification** reporta cuantos casos fueron incorrectamente clasificados en el árbol en la muestra de aprendizaje y en la muestra de prueba.



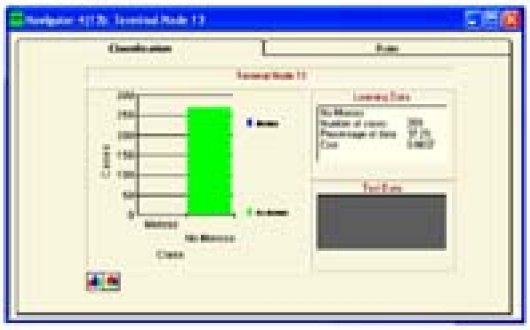
La opción de **Prediction Success** muestra si el algoritmo CART tiende a concentrar su clasificación errónea en categorías específicas de la variable criterio y, si este fuera el caso, indica donde está ocurriendo la clasificación errónea.



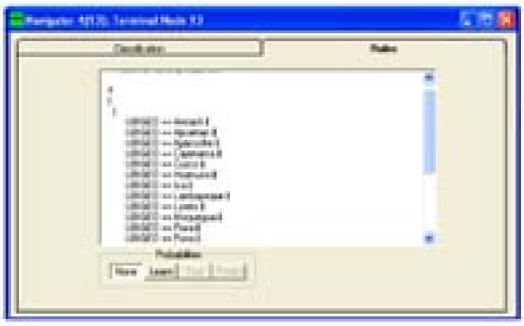
Obtención de las Reglas de Clasificación

Se pueden obtener las reglas de clasificación por cada nodo terminal o de todos los nodos terminales del árbol.

Para obtener las reglas de clasificación de un nodo terminal se hace clic en el nodo terminal

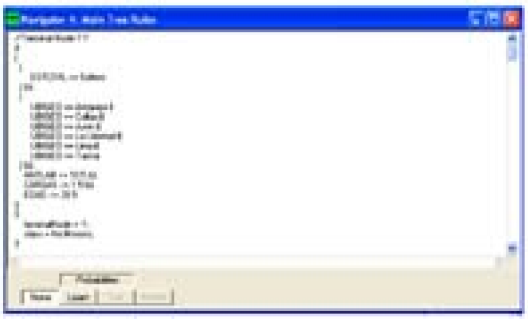


y luego se escoge la opción de **Rules**



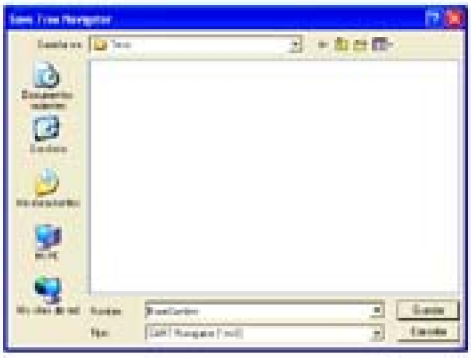
Para obtener las reglas de clasificación de todos los nodos terminales se escoge la opción **Rules...** del menú **View**.

"Programa Cybertesis PERÚ - Derechos son del Autor"



Grabando la salida del modelo

Se puede grabar la salida del modelo (árbol, reportes, reglas, etc.) usando la opción **Save Navigator...** del menú **File**.



Anexo Nº 2

Reglas de Clasificación para los morosos

```
/*Terminal Node 1*/
if
(
  (
      ESTCIVIL == Soltero
  ) &&
  (
      UBIGEO == Arequipa ||
      UBIGEO == Callao ||
      UBIGEO == Junin ||
      UBIGEO == La Libertad ||
      UBIGEO == Lima ||
      UBIGEO == Tacna
  ) &&
    ANTLAB <= 10.5 &&
    CARGAS <= 1.5 &&
    EDAD <= 28.5
)
{
    terminalNode = -1;
    class = No Moroso;
}
```

```
/*Terminal Node 7*/
if
(
  (
      UBIGEO == Arequipa ||
      UBIGEO == Junin ||
      UBIGEO == La Libertad ||
      UBIGEO == Tacna
  ) &&
  (
      ESTCIVIL == Casado ||
      ESTCIVIL == Divorciado ||
      ESTCIVIL == Viudo
  ) &&
    EDAD > 41.5 &&
    EDAD <= 54.5 &&
    ANTLAB > 5.5 &&
    ANTLAB <= 8.5
)
{
    terminalNode = -7;
    class = Moroso;
}
```

```
/*Terminal Node 9*/
if
(
  (
      UBIGEO == Arequipa ||
      UBIGEO == Junin ||
      UBIGEO == La Libertad ||
      UBIGEO == Tacna
  ) &&
  (
      ESTCIVIL == Casado ||
      ESTCIVIL == Divorciado ||
      ESTCIVIL == Viudo
  ) &&
    ANTLAB > 8.5 &&
    EDAD > 49.5 &&
    EDAD <= 54.5
)
{
    terminalNode = -9;
    class = Moroso;
}
```


"Programa Cybertesis PERÚ - Derechos son del Autor"

```
/*Terminal Node 11*/
if
(
  (
      UBIGEO == Callao ||
      UBIGEO == Lima
  ) &&
  (
      ESTCIVIL == Casado ||
      ESTCIVIL == Divorciado ||
      ESTCIVIL == Viudo
  ) &&
    EDAD <= 28.5
)
{
    terminalNode = -11;
    class = Moroso;
}
```

Anexo Nº 3

Reglas de Clasificación para los no morosos

```
/*Terminal Node 1*/
if
(
  (
     ESTCIVIL == Soltero
  ) &&
  (
     UBIGEO == Arequipa ||
     UBIGEO == Callao ||
     UBIGEO == Junin ||
     UBIGEO == La Libertad ||
     UBIGEO == Lima ||
     UBIGEO == Tacna
  ) &&
    ANTLAB <= 10.5 &&
    CARGAS <= 1.5 &&
    EDAD <= 28.5
)
{
    terminalNode = -1;
    class = No Moroso;
}
```


"Programa Cybertesis PERÚ - Derechos son del Autor"

```
/*Terminal Node 3*/
if
(
  (
     ESTCIVIL == Soltero
  ) &&
  (
     UBIGEO == Arequipa ||
     UBIGEO == Callao ||
     UBIGEO == Junin ||
     UBIGEO == La Libertad ||
     UBIGEO == Lima ||
     UBIGEO == Tacna
  ) &&
    ANTLAB <= 10.5 &&
    CARGAS > 1.5
)
{
    terminalNode = -3;
    class = No Moroso;
}
```

```
/*Terminal Node 4*/
if
(
  (
     ESTCIVIL == Soltero
  ) &&
  (
     UBIGEO == Arequipa ||
     UBIGEO == Callao ||
     UBIGEO == Junin ||
     UBIGEO == La Libertad
     UBIGEO == Lima ||
     UBIGEO == Tacna
  ) &&
    ANTLAB > 10.5
)
{
   terminalNode = -4;
   class = No Moroso;
}
```

```
/*Terminal Node 5*/
if
(
  (
      UBIGEO == Arequipa ||
      UBIGEO == Junin ||
      UBIGEO == La Libertad ||
      UBIGEO == Tacna
  ) &&
  (
      ESTCIVIL == Casado ||
      ESTCIVIL == Divorciado ||
      ESTCIVIL == Viudo
  ) &&
    EDAD <= 41.5
)
{
    terminalNode = -5;
    class = No Moroso;
}
```

```
/*Terminal Node 6*/
if
(
  (
      UBIGEO == Arequipa ||
      UBIGEO == Junin ||
      UBIGEO == La Libertad ||
      UBIGEO == Tacna
  ) &&
  (
      ESTCIVIL == Casado ||
      ESTCIVIL == Divorciado
      ESTCIVIL == Viudo
  ) &&
    EDAD > 41.5 &&
    ANTLAB <= 5.5
)
{
    terminalNode = -6;
    class = No Moroso;
}
```


"Programa Cybertesis PERÚ - Derechos son del Autor"

```
/*Terminal Node 8*/
if
(
  (
      UBIGEO == Arequipa ||
      UBIGEO == Junin ||
      UBIGEO == La Libertad ||
      UBIGEO == Tacna
  ) &&
  (
      ESTCIVIL == Casado ||
      ESTCIVIL == Divorciado ||
      ESTCIVIL == Viudo
  ) &&
    ANTLAB > 8.5 &&
    EDAD > 41.5 &&
    EDAD <= 49.5
)
{
    terminalNode = -8;
    class = No Moroso;
}
```

```
/*Terminal Node 10*/
if
(
  (
      UBIGEO == Arequipa ||
      UBIGEO == Junin ||
      UBIGEO == La Libertad ||
      UBIGEO == Tacna
  ) &&
  (
      ESTCIVIL == Casado ||
      ESTCIVIL == Divorciado ||
      ESTCIVIL == Viudo
  ) &&
    ANTLAB > 5.5 &&
    EDAD > 54.5
)
{
    terminalNode = -10;
    class = No Moroso;
}
```


"Programa Cybertesis PERÚ - Derechos son del Autor"

```
/*Terminal Node 12*/
if
(
  (
      UBIGEO == Callao ||
      UBIGEO == Lima
  ) &&
  (
      ESTCIVIL == Casado ||
      ESTCIVIL == Divorciado ||
      ESTCIVIL == Viudo
  ) &&
    EDAD > 28.5
)
{
    terminalNode = -12;
    class = No Moroso;
}
```

```
/*Terminal Node 13*/
if
(
  (
      UBIGEO == Ancash ||
      UBIGEO == Apurimac ||
      UBIGEO == Ayacucho ||
      UBIGEO == Cajamarca ||
      UBIGEO == Cuzco ||
      UBIGEO == Huanuco ||
      UBIGEO == Ica ||
      UBIGEO == Lambayeque ||
      UBIGEO == Loreto ||
      UBIGEO == Moquegua ||
      UBIGEO == Piura ||
      UBIGEO == Puno |
      UBIGEO == San Martín ||
      UBIGEO == Tumbes ||
      UBIGEO == Ucayali
  )
)
{
   terminalNode = -13;
   class = No Moroso;
}
```


"Programa Cybertesis PERÚ - Derechos son del Autor"

TEXTO COMPLETO EPUB

Descargar texto completo en formato EPUB en la dirección:

http://www.cybertesis.edu.pe/sisbib/2005/salinas_fj/epub/salinas_fj.epub