UNIVERSIDAD NACIONAL

"PEDRO RUIZ GALLO"

ESCUELA DE POSTGRADO

MAESTRÍA EN CIENCIAS

"Modelos supervisados para evaluar el riesgo crediticio en la calificación de créditos en personas naturales en una Institución Financiera en Chiclayo"

TESIS

Presentada para optar el Grado Académico de Maestro en Ciencias con Mención en Proyectos de Inversión

AUTOR:

Econ. Uriol Chavez, Sebastian Javier

ASESOR:

Mg. Tesen Arroyo, Alfonso

LAMBAYEQUE – PERÚ

2019

“Modelos supervisados para evaluar el riesgo crediticio en la calificación de créditos en personas naturales en una Institución Financiera en Chiclayo”

PRESENTADO POR:

Econ. URIOL CHAVEZ SEBASTIAN JAVIER
AUTOR

Mg. ALFONSO TESEN ARROYO
ASESOR

APROBADO POR:

Dr. LUIS ANIBAL ESPINOZA POLO
PRESIDENTE

Dra. ANA BERTHA COTRINA CAMACHO
SECRETARIA

Mg. LINDON VELA MELENDEZ
VOCAL

ACTA DE SUSTENTACIÓN

ACTA DE SUSTENTACIÓN DE TESIS

080

Siendo las 10:15 horas del día 10 de diciembre del año Dos Mil diecinueve, en la Sala de Sustentación de la Escuela de Posgrado de la Universidad Nacional Pedro Ruiz Gallo de Lambayeque, se reunieron los miembros del Jurado, designados mediante Resolución N° 1492-2019-EPG de fecha 25 octubre 2019, conformado por:

Dr. Luis Anibal Espinoza Polo — PRESIDENTE (A)
Dra. Ana Bertha Cotrina Camacho — SECRETARIO (A)
Mg. Lindon Vela Melendez — VOCAL
Mg. Alfonso Tesen Arroyo — ASESOR (A)

Con la finalidad de evaluar la tesis titulada "Modelo Supervisado para Evaluar el Riesgo Crediticio en la Calificación de Crédito en Personas Naturales en una Institución Financiera en Chiclayo".

presentado por el (la) Tesista Sebastian Javier Ulloa Chavez

sustentación que es autorizada mediante Resolución N° 1658-2019 de fecha 28 noviembre 2019

El Presidente del jurado autorizó del acto académico y después de la sustentación, los señores miembros del jurado formularon las observaciones y preguntas correspondientes, las mismas que fueron absueltas por el (la) sustentante, quien obtuvo 90 puntos que equivale al calificativo de Muy bueno

En consecuencia el (la) sustentante queda apto (a) para obtener el Grado Académico de: Maestro en Ciencias con Mención en Proyectos de Inversión

Siendo las 11:20 horas del mismo día, se da por concluido el acto académico, firmando la presente acta.

PRESIDENTE | SECRETARIO

VOCAL | ASESOR

DECLARACIÓN JURADA DE ORIGINALIDAD

Econ. Uriol Chavez, Sebastian Javier, Investigador Principal y **Dr. Tesen Arroyo, Alfonso**, Asesor del Trabajo de Investigación **"Modelos supervisados para evaluar el riesgo crediticio en la calificación de créditos en personas naturales en una Institución Financiera en Chiclayo"**; declaro bajo juramento que este trabajo no ha sido plagiado, ni contiene datos falsos. En caso se demostrará lo contrario, asumo responsablemente la anulación de este informe y por ende el proceso administrativo a que hubiera lugar. Que puede conducir a la anulación del título o grado emitido como consecuencia de este informe.

Lambayeque, 10 de Diciembre de 2019.

INVESTIGADOR: Econ. Uriol Chavez, Sebastian Javier

ASESOR: Dr. Tesen Arroyo, Alfonso

DEDICATORIA

Mi tesis la dedico a mis hijos Jeannie y Javier, fuente de motivación para lograr mi superación día a día y el logro de un futuro siempre mejor.

A mi fiel compañera Sadith, motivadora también para llevar adelante cada uno de mis proyectos

A mis amados padres, quienes no estando presentes, guían cada uno de mis pasos en este sinuoso camino de la vida.

AGRADECIMIENTO

Un agradecimiento especial para mi asesor el Dr. Alfonso Tesen Arroyo, por su dedicación a la preparación y culminación del presente trabajo

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

La presente investigación tuvo como objetivo aplicar las técnicas de minería de datos para evaluar el riesgo crediticio en la calificación de créditos en personas naturales de una institución financiera de la Ciudad de Chiclayo.

La investigación es de tipo exploratorio, *tecnologíca aplicada blanda, rama de la tecnología: general, tecnología: optimización, objetivo: minimización del riesgo, maximización del rendimiento, para lo cual se trabajó con la metodología CRISP- DM.*

Cabe recordar que este tipo de investigación no es adecuada a la calificación de verdadero-falso sino mas bien a la de eficiente-deficiente, eficaz o ineficaz, es por ello que la metodología de la investigación científica plantea que se puede obviar el planteamiento de hipotesis por las consideraciones mas arriba mencionadas

Para el desarrollo de la investigación se utilizaron dos modelos supervisados tales como, Árboles de Clasificación y el modelo clásico de la Regresión Logística; la base de datos estuvo constituida por 2356 clientes, de los cuales se utilizó el 70% de la base para el entrenamiento y el 30% restante para la validación. Para la evaluación de los modelos, se utilizó la Matriz de Confusión y la curva ROC, entre otros métodos que permiten que en la etapa de modelado, estos sean representativos y permitan la predicción de tal manera que sean utilizados para minimizar el riesgo crediticio en la concesión de créditos.

Cabe mencionar que dentro de la metodología adoptada, debemos distinguir que nuestro análisis es predictivo y no de pronóstico. El pronóstico implica predecir cuantos bienes demandaran los agentes económicos para un periodo posterior, mientras que el análisis predictivo implica predecir que agentes económicos es mas probable que demanden los mencionados bienes.

Para hacer análisis predictivo hay necesidad de contar de buena cantidad de datos, actuales y pasados de tal forma que nos permitan determinar patrones de comportamiento y deducir conocimiento.

INTRODUCCIÓN

La problemática que enfrenta casi todas las entidades financieras es la existencia de un nivel de riesgo en el cual estos entes están inmersos, la presencia de morosidad por parte de los prestatarios y hasta cierto punto de incobrabilidad de las operaciones al crédito que realizan son motivos por los cuales la colocación del crédito resulta cada vez más preocupante por la incertidumbre que genera. (Sellan, 2011)

La institución financiera motivo de esta investigación no es ajena a esta problemática ya que presenta dificultades en cuando a la concesión de créditos personales; actualmente sus clientes presentan retrasos con las fechas de pago, en algunos casos se enfrentan al incumplimiento total en el pago, situación en el cual debe optar por una serie de medidas que van de las vías administrativas (notificaciones y cobranzas a domicilio) a las vías judiciales (procesos de embargo) para poder evitar el continuo sobreendeudamiento. En estos últimos años la cooperativa presentó un nivel de morosidad muy variable, el año 2009 alcanzó un 8,21% que disminuyo hasta 7,94% al cierre del 2010, para fines del 2011 presento una morosidad de 6,76% y se incrementó a 9,75% al cierre del 2012 y que fue disminuyendo hasta un 4,79% para junio del 2013. Esta variabilidad, demuestra una ineficiencia en la asignación de créditos y la necesidad de ajustar los criterios de evaluación de los clientes a fin de mantener una morosidad moderada en los próximos años.

Por diversas cuestiones existe un grupo identificable de agentes que la tecnología convencional en la producción de servicios financieros rechaza, principalmente por problemas de información. En los mercados financieros existe la asimetría de la información cuya consecuencia inmediata es el racionamiento del crédito, es en este sentido que adquiere relevancia la incertidumbre sobre el reembolso de los préstamos, cuando esta es elevada es posible que simplemente los préstamos no sean otorgados. Por otro lado, tradicionalmente se presume demasiado costoso adquirir la información que se necesita para pronosticar mejor el comportamiento del deudor. Siendo el interés del individuo no estar sujeto a dicha restricción, tiene incentivos para procurar proveer esta información al prestamista, estos problemas clasificados por la teoría como de información imperfecta, son causales de racionamiento del crédito. Mientras exista requerimiento de garantías colaterales y el costo de la información de selección y del monitoreo sea alto, los prestamistas convencionales restringirán los montos a disposición de los prestatarios o directamente podrían decidir no prestarle a determinado grupo de solicitantes. (Baca, G. A. 1997)

Así podemos identificar grupos de la población que debido a su bajo nivel de ingresos presentes y/o flujo futuro de ingresos inciertos, es decir, poseen escasos activos liquidables, encuentran restringido su acceso al crédito a cualquier tasa de interés y esto implica un límite al nivel de bienestar alcanzable. La tecnología de crédito convencional implica un alto costo financiero para suplir la demanda de algunos tipos de créditos. Es así como cada vez más, los mercados de capitales se han vuelto accesibles a las Pequeñas y Medianas Empresas PYMES, y éstas a su vez no presentan estructuras financieras que faciliten su inserción en modelos de análisis de riesgo, presentándose inicialmente como altamente riesgosas por sí mismas, por ello, a través de una adecuada determinación de perfiles, pueden verse como un mercado potencial interesante que genere un buen índice de rentabilidad al atenderlas financieramente. (Krugman, R. P. 1995)

Vistas estas problemáticas se entiende que la restricción al acceso al crédito no necesariamente reflejará falta de capacidad de pago del potencial deudor, sino un complejo entramado de relaciones entre los prestamistas y los aspirantes a crédito; para contrarrestar los efectos adversos de la operatoria tradicional sobre la distribución del ingreso, se vuelve necesario explorar y desarrollar nuevas tecnologías de crédito que superen estas barreras. (Fragoso, J. 2002)

Se ha logrado encontrar la manera de solucionar el problema técnico de producir servicios financieros para clientelas marginadas a un costo razonable y una tasa de ganancia positiva, es decir, una función de producción (tecnología) que ha posibilitado este resultado, abriendo nuevas posibilidades para relajar las restricciones de liquidez. (Baca, G. A. 1997)

Dada la necesidad de establecer mecanismos para medir la probabilidad de no pago, se deben determinar las variables significativas que expliquen el fenómeno y contribuyan a generar un modelo de medición de riesgo de crédito, a través de métodos estadísticos teórico-prácticos, que puedan ser implementados para el otorgamiento de créditos en una en la institución financiera materia de estudio, haciéndose ampliamente pertinente dado que la institución está interesada en generar modelos con los cuales logre medir la posibilidad de no pago, en las diferentes líneas de créditos establecidas. Bajo las situaciones anteriores, se presenta la oportunidad de generar un método de medición de riesgo de crédito, el cual se abordará desde los ***denominados modelos predictivos*** que permitan estimar la probabilidad de incumplimiento, con los cuales se puedan efectuar comparaciones de las bondades y desventajas que cada uno de ellos presenta. Para este efecto, la Entidad suministrará la

información acerca del perfil de los clientes actuales en cada línea de crédito que permita asumir variables de tipo tanto cualitativo como cuantitativo y así desarrollar los modelos; bajo un laborioso proceso metodológico para el diseño y selección de una muestra, la cual generará los valores para las variables exógenas y endógenas con las cuales se correrá cada uno de los modelos. (Cabrera, A. 2014)

CAPITULO I

ANÁLISIS DEL OBJETO DE ESTUDIO

CAPÍTULO. I ANÁLISIS DEL OBJETO DE ESTUDIO

3.1. Ubicación

Institución Financiera en Chiclayo, inició sus actividades con siete personas. La superintendencia de Banco y Seguros autorizó sus operaciones en diciembre de 1986. El patrimonio inicial fue de aproximadamente $US 30,000 dólares que fueron en el punto de partida al servicio de las microfinanzas en Chiclayo.

En 1986 iniciaron el trabajo con una oficina y en la actualidad cuentan con 75 puntos de atención, llevan sus servicios a trece regiones del Perú: Tumbes, Piura, Lambayeque, La Libertad, Cajamarca, Ancash, Ica, Lima, Callao, Arequipa, Moquegua, Puno y Cuzco.

Ofrecen productos de Ahorro y Crédito. En ahorros disponen de una gama de posibilidades para generar y consolidar una cultura del ahorro en nuestro país: Depósitos de Ahorro, Depósitos a Plazo, Multimás, Rinde +, Ahorro Plan, CTS y Ahorro Con órdenes de Pago, En créditos atienden de manera rápida y oportuna los requerimientos de nuestros clientes.

Cuentan con Crédito Empresarial, Crédito Pesca, Crédito Agropecuario, Crédito Personal, Cuenta Sueldo, Crédito Descuento Por Planilla, Crédito Prendario, Crédito Compuplan, Crédito Vehicular, Credigas GNV y GLP, Sully Te Presta y Vive Mejor, entre otros. Este abanico de alternativas de crédito les permite crecer junto a nuestros clientes.

Además, disponen de alta tecnología financiera como cajeros automáticos, homebanking, Tarjeta de Débito VISA para compras en el Perú y el extranjero; asimismo, operaciones a través de la Cámara de Compensación Electrónica que buscan estar cerca de sus clientes.

3.2. Cómo surge el problema

El problema que enfrentan casi todas las entidades financieras, es la existencia de un nivel de riesgo en el cual estos entes están inmersos en la presencia de morosidad por

parte de los prestatarios y hasta cierto punto de incobrabilidad de las operaciones al crédito resulta cada vez más preocupante por la incertidumbre que genera. (Marzo, C.; Wicijowski, C. & Rodriguez, L. 2008)

La gestión corporativa del riesgo se ha convertido en un elemento importante dentro de las políticas administrativas de las instituciones dedicadas al otorgamiento de créditos; de igual manera en el entorno económico peruano han cobrado especial importancia las figuras crediticias, como instrumentos para la generación de alternativas de crecimiento, teniendo claro que el crédito por sí solo no es suficiente para impulsar el desarrollo económico. (Aguilar, G. 2004)

3.3. Cómo se manifiesta y qué características tiene.

Se manifiesta de tal manera que cada día se incrementa el porcentaje de clientes que caen en impago debido a la mala evaluación de éstos para hacer acreedores de préstamos de la entidad financiera. Tanto es así que a partir de la tercera cuota comienzan los retrasos de los pagos comprometidos con la institución, situación que perjudica e incrementa la deuda.

3.4. Objetivos

1.4.1. Objetivo general

Implementar un modelo predictivo basado en las técnicas de minería de datos de forma tal que permitan discriminar y clasificar adecuadamente el riesgo crediticio teniendo en cuenta el proceso de descubrir conocimiento en base a los datos, mediante la aplicación de algoritmos de regresión logística y árboles de clasificación.

1.4.2. Objetivos específicos

- ✓ Calcular los parámetros de la regresión logística y aplicar las técnicas de minería de datos para validar el mismo
- ✓ Aplicar el algoritmo C0.1 para el diseño de un árbol de clasificación validando luego el mismo con las técnicas de minería de datos.

3.5. Descripción detallada de la metodología

Para la recolección de la base de datos, constituida por el historial crediticio de 2356 clientes que forman parte de la cartera de crédito personal de la Institución financiara, contiene en total 6 variables seleccionadas por criterio del investigador, el análisis estadístico se llevó a cabo considerando todas las variables en estudio.

Al inicio se realizó la evaluación de la relación entre la variable dependiente y las variables independientes con el objetivo de encontrar al menos una variable explicativa que permita la construcción del modelo.

Determinadas las variables del caso y especificado el modelo, se aplicarán las técnicas adecuadas a los modelos en su forma de muestra completa, a fin de evaluar el rendimiento, en su forma de discriminación (Regresión Logística) y clasificación (Árbol de decisión)

Posteriormente, se particióna la muestra para trabajar el entrenamiento y la validación del modelo, de esta manera se plantea que un 70% de los casos serán destinados al conjunto de entrenamiento (construcción del modelo) y el otro 30% al conjunto de validación (ajuste y comparación de los modelos); no se considerará asignar un porcentaje al conjunto de testeo por la limitación de no contar con una mayor cantidad de datos. En la modelización predictiva del conjunto de datos de entrenamiento se pueden fácilmente generar modelos que predigan el valor del TARGET (variable Respuesta) a partir de un conjunto de valores de entrada; sin embargo, estas predicciones solo son precisas para el propio conjunto de entrenamiento. El intento de generalizar las predicciones de este conjunto de datos a un conjunto independiente, pero con una distribución similar puede producir resultados con un deterioro significativo de la

precisión; con el fin de evitar este problema se utilizó un conjunto de validación como forma de evaluar independientemente la performance de un modelo.

CAPITULO II

MARCO TEÓRICO

CAPÍTULO II: MARCO TEÓRICO

3.1. Antecedentes

Cabrera, A. (2014) en su investigación titulada "Diseño de credit scoring para evaluar el riesgo crediticio en una entidad de ahorro y crédito popular", sostiene que: En caso de la institución objeto de estudio, se identificó una base de datos que no se encuentra depurada y contiene información que no aportó nada a la investigación. Las variables que se incluyeron en el modelo son propias de la institución y probablemente no sean útiles para otro modelo de credit scoring, puesto que lo que es significativo para una entidad no lo es para otra. El credit scoring representa para la institución una herramienta útil en la evaluación del sujeto a manera de sugerencias de aceptación o no de la solicitud de crédito. El nivel de riesgo que la institución esté dispuesto a correr será el indicador para aceptar o rechazar solicitudes, por otro lado, la determinación de los puntos de corte depende de cada institución. Se concluye que la hipótesis planteada se cumple al comprobarse que el perfil del cliente tiene que ver con su nivel de cumplimiento, obteniéndose que las variables que mejor pueden explicar la morosidad de la institución, de acuerdo al diseño del modelo de credit scoring obtenido son: Oficina, Buró, Producto y Estado civil. Por lo anterior, podemos determinar que dentro del credit scoring el perfil del cliente tiene que ver con su nivel de cumplimiento, puesto que logra diferenciar entre un cliente bueno y un cliente malo.

Ladino, I. (Marzo 2014) en su investigación titulada: "Comparación de modelos de riesgo de crédito: modelos logísticos y redes neuronales", sostiene que: Como resultado de comparar el desempeño de los modelos de ANN y la regresión logística en 1000 muestras con remplazo (bootstrap), los estadísticos D y C de los modelos ANN son estadísticamente mayores. Adicionalmente, la significancia económica en magnitud de esta mayor discriminación (0,0088 y 0,0052 en los estadísticos D y c respectivamente) es material, al disminuir la perdida en 4,4% para el caso analizado.

Es importante resaltar que es más difícil implementar los modelos ANN que los modelos logísticos en los sistemas de las entidades financieras. Los modelos ANN tienen un mayor poder de discriminación y un impacto económico material, sin embargo, este resultado es producto de comparar estos modelos con una función de costos simétrica. Un posible trabajo futuro consistiría en comparar los modelos logísticos y neuronales

utilizando una función de costos asimétrica que podría otorgar mayor costo al error de clasificar un cliente malo como bueno, respecto al error de clasificar un cliente bueno como malo.

Moreno, S. (2013) en su investigación titulada: "El modelo Logit Mixto para la construcción de un scoring de crédito", sostiene que: Los tres modelos estimados (Logit tradicional, probit y logit Mixto) tiene un buen poder discriminatorio, reflejado en las altas tasas de aciertos sobre todo para los clientes morosos. El modelo logit mixto resultó ser el de mayor sensibilidad, aunque también predijo el mayor número de falsos positivos. En cuanto a las variables que determinan que un cliente llegue a default, resultaron significativas las relacionadas con el factor de comportamiento crediticio, financiero y demográfico, como se esperaba. Se descartaron algunas variables como las moras mayores a 30 días, por problemas de tipificación. Las variables que explican el evento de llegar a default en los modelos ajustados, resultaron con signos acordes con la realidad de la entidad financiera. Mediante el modelo logit mixto se pudo determinar que los factores nivel de estudio, tipo de relación laboral, número de meses desde el último crédito y edad definida por grupos, tiene un efecto aleatorio en el modelo para predecir el default en una entidad financiera del sector cooperativo. Para una entidad financiera es muy importante contar con una herramienta estadística adecuada para la predicción del comportamiento de los clientes al momento de otorgarles el crédito, puesto que la rentabilidad y los flujos de caja, en gran medida corresponden al correcto pago de las obligaciones. Crediticias contraídas por parte de los clientes. El modelo logit mixto es la más potente en la predicción o detección de los clientes que llegan al estado de default, pero esta predicción está asociada a que es un modelo muy estricto de la aceptación de clientes óptimos (no default), lo que genera un gran porcentaje de rechazo de clientes que en su historial crediticio han pagado bien (Error tipo I), ocasionando en largo plazo un problema de crecimiento de mercado para la entidad financiera.

3.2. Base Teórica

La base teórica que presentamos a continuación tiene que ver con proposiciones sobre nuestro tema de estudio y los mostramos a continuación:

2.2.1. Riesgo Crediticio

La principal actividad de una entidad financiera es aquella que mejor la define y a la que dedica la mayor parte de sus esfuerzos. La actividad que genera la mayor parte de sus beneficios y los mayores riesgos, es la actividad crediticia. (Vigo, G. 2010)

Habitualmente la palabra riesgo tiene una connotación negativa: algo que debemos evitar. Sin embargo, el negocio bancario supone precisamente eso, la gestión de riesgos con el objetivo de obtener una rentabilidad que compense adecuadamente. Un banco es básicamente una máquina de gestión de riesgos en busca de rentabilidad. De todos los riesgos a los que está expuesto el negocio bancario, el principal es el riesgo de crédito. Este se define como la posibilidad de incurrir en pérdidas como consecuencia del incumplimiento por parte del deudor de sus obligaciones en las operaciones de intermediación crediticia. El más grave de los incumplimientos es el impago. (Vigo, G. 2010).

2.2.2. Análisis de Crédito

El análisis de crédito es considerado como un arte; ya que no hay esquemas rígidos y que por el contrario es dinámico y exige creatividad, Sin embargo, resulta importante dominar las diferentes técnicas de análisis de créditos y complementarla con una amplia experiencia y buen criterio, asimismo es necesario contar información disponible, necesaria y suficiente que permita minimizar el número de incógnitas para poder tomar la decisión correcta. (Nieto, S. 2010).

2.2.3. Técnicas de Clasificación

2.2.3.1. Minería de Datos

La minería de datos puede conceptualizarse como el conjunto de técnicas que permiten el análisis de conjuntos de datos, expresión de

forma de actuar de los agentes económicos, y encontrar patrones de comportamiento, asociaciones entre otras características que permiten generar conocimiento a partir de ellos.

La minería de datos tiene que ver con tres aspectos fundamentales:

- La estadística (El estudio numérico de las relaciones entre los datos)
- La inteligencia artificial
- Machine Learning (Aprendizaje automático, algoritmos variados de manipulación de datos para encontrar relaciones)

Aspectos Fundamentales de la minería de datos

Figura 1: Aspectos fundamentales de la minería de datos.

Al respecto se puede leer en la introducción del software SAS para minería de datos:

¿Entonces por qué es importante la minería de datos? Ha podido apreciar los números asombrosos – el volumen de datos producidos se duplica cada dos años. Los datos no estructurados por sí solos conforman el 90% del universo digital. Pero más información no significa necesariamente más conocimientos.

La minería de datos le permite:

- Filtrar todo el ruido caótico y repetitivo en sus datos.
- Entender qué es relevante y luego hacer un buen uso de esa información para evaluar resultados probables.
- Acelerar el ritmo de la toma de decisiones informadas.

Podemos decir que el proceso de minería de datos implica:

- Elección de los datos a analizar: tarea inicial concerniente a la selección de las variables motivo de predicción, las llamadas variables dependientes en Matemáticas y endógenas en Econometría, de igual forma, seleccionar las llamadas independientes o predeterminadas, que tiene el papel de explicativas del proceso predictivo.
- Propiedades de los datos: esto tiene una implicancia de naturaleza grafica fundamentalmente, además de algunas anomalías encontradas tales como valores atípicos o ausentes.
- Modificación de los datos a utilizar: es un paso fundamental en lo concerniente a preparar los datos para el modelado, variables cualitativas por ejemplo tienen que ser preparadas adecuadamente a fin de dar correcta interpretación luego de la aplicación de la técnica de minería de datos
- Modelación; aquí es donde se debe escoger la técnica de minería de datos en términos de predicción, clasificación o segmentación.
- Generación de conocimiento: los modelos de minería de datos nos entregan adecuadamente esquemas de comportamiento recurrentes, relaciones de asociación entre las variables, elementos de juicio necesarios para encarar toma de decisiones o predicción de comportamientos futuros.

- Evaluación: en este paso se considera necesario comprobar desde el punto de vista estadístico, la validez de los resultados del modelo aplicado de tal manera que nuestras conclusiones sean válidas.

Las técnicas de modelado en minería de datos, en su etapa de modelado, distingue tres tipos de modelos:

a) Modelos de clasificación: similar a los modelos causales, tienen en cuenta una o más variables denominadas de entrada a fin de predecir las denominadas variables de destino. Ejemplo de ello son: arboles de decisión, regresión lineal, logística, lineal generalizada, así como algoritmos de cox que tienen que ver con análisis de supervivencia, redes neuronales. Estos también se denominan modelos supervisados.

b) Modelos de asociación: son modelos que buscan patrones o asociaciones, correlaciones, así como también estructuras causales basados en la información alcanzada como datos.

c) Modelos de segmentación: dividen los datos en conjuntos de los mismos con patrones similares. En este tipo de modelos no se distingue entre variables de entrada y de destino Estos dos últimos son los modelos no supervisados.

Clasificación de las técnicas de Data Mining

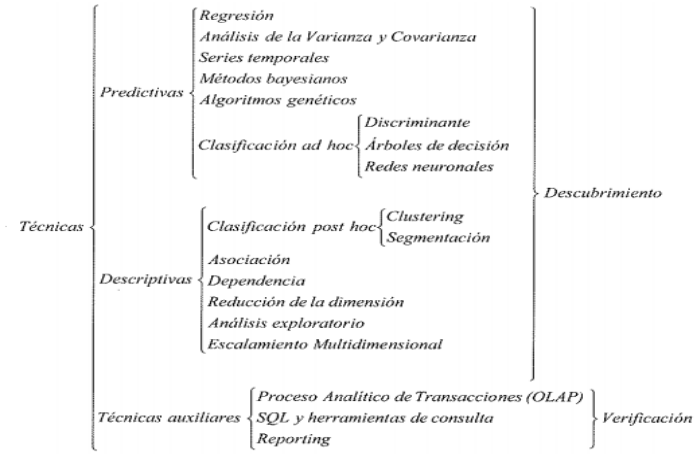


Figura 2: Clasificación de las técnicas de Data Mining (Pérez y Santín, 2008)

Las herramientas disponibles de minería de datos más utilizadas en la actualidad son:

- IBM SPSS MODELER
- SAS ENTERPRISE MINER
- SQL ANALYSIS SERVICES
- ORACLE DATA MINING

En la práctica todo el entorno para el cálculo presenta procedimientos para usar en minería de datos. Típica estructura de IBM SPSS MODELER

Estructura de IBM SPSS Modeler

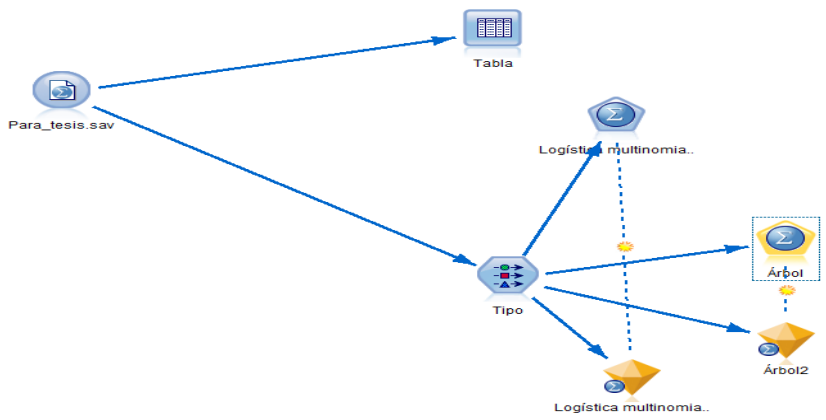


Figura 3: Estructura de IBM SPSS Modeler

2.2.3.2. Métodos para la evaluación de las técnicas de clasificación

Para medir la performance de las técnicas de clasificación se han propuesto una serie de métodos y criterios con la finalidad de validar y evaluar su bondad de ajuste al conjunto de datos, y ser utilizado para la predicción, la aplicación dependerá de la técnica de minería de datos utilizada.

Entre los métodos propuestos se tienen:

Métodos para la validación de modelos supervisados. Son métodos que permiten evaluar la performance de los modelos, cuya finalidad es realizar una evaluación sobre su bondad de ajuste al conjunto de datos. Los métodos consisten en dividir el conjunto total de observaciones en tres subconjuntos: conjunto de entrenamiento (usado para el proceso de aprendizaje o estimación del modelo), conjunto de validación (usado para validar el modelo) y conjunto de prueba (usado para la inferencia de nuevas observaciones); el ultimo conjunto es opcional y generalmente se usa datos no incluidos en la base de datos.

La regla de clasificación en los modelos supervisados se encuentra al determinar un punto de corte entre 0 y 1. Si para un elemento la probabilidad de que Y=1 es mayor que el punto de corte; se considera que este se ubica en la clase de interés, determinada por Y=1; de otro modo el elemento se ubica en la clase determinada por Y=0. (Hernández, O. R. 2015)

Existen varios métodos para validación de los modelos tales como: el método de validación cruzada, la tabla de confusión, la curva ROC, entre otras.

Los Métodos de validación cruzada (Croos – Validation): en el aprendizaje supervisado, consiste en dividir aleatoriamente el conjunto de entrenamiento D en k subconjuntos (k-folds) mutuamente excluyentes $\{D_1, D_2, \dots, D_k\}$ de similar tamaño. El proceso de validación cruzada es repetido durante k iteraciones, de tal manera que en cada iteración el modelo usa un subconjunto para la validación (D_V) y es entrenado con los k – 1 subconjuntos $(D - D_V)$, el error de clasificación se calcula como la media aritmética de los errores de cada iteración. Un caso particular de la validación cruzada dejar – uno – afuera (Leave – one–out), implica que en cada iteración se tenga un solo dato de prueba y el resto para entrenamiento, el error se calcula como el promedio de los errores cometidos. (Peña, D. 2002).

También es posible y el error absoluto medio (EAM) como indicadores de la capacidad predictiva del modelo.

a) Validación cruzada (k-folds): Para cada iteración se encuentra el error y luego se evalúa de acuerdo a la siguiente formula que no es otra cosa que la media de los K errores

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i .$$

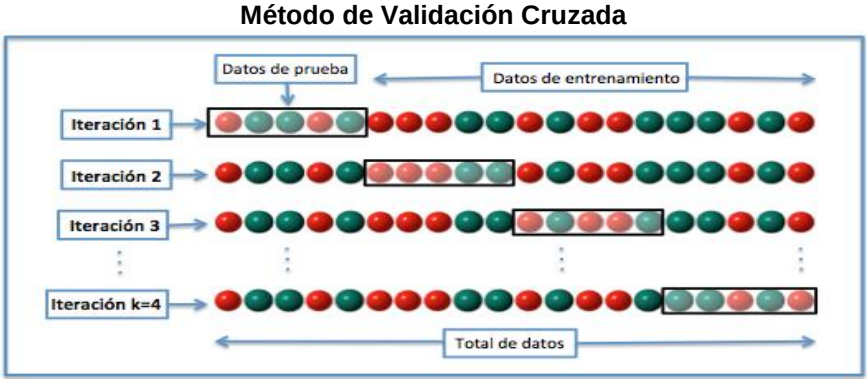


Figura 4:Método de Validación Cruzada K folds

b) Validación Cruzada (Leave – one–out): Para este caso, la formula final a utilizar es también una media, pero para N muestras.

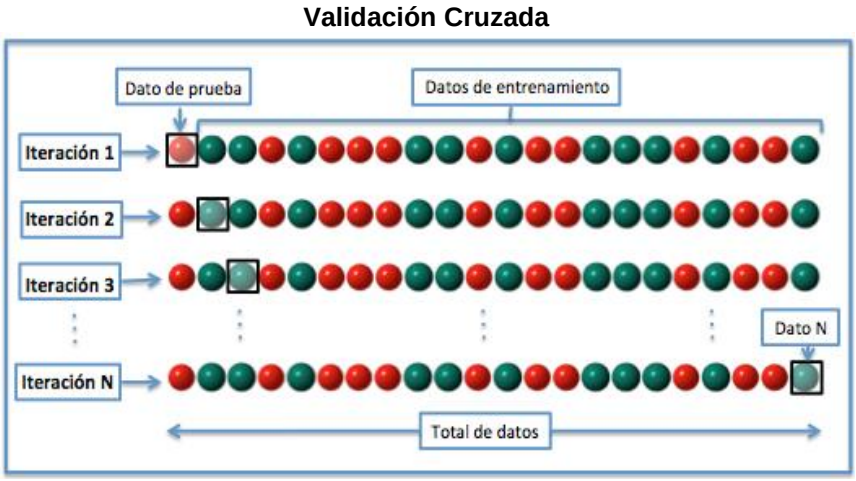


Figura 5: Validación Cruzada Leave-One-out

c) Método Holt – Out. Este método particiona aleatoriamente el conjunto de datos D en dos conjuntos mutuamente excluyentes: conjunto de entrenamiento ($\boldsymbol{D_E}$) y conjunto de validación ($\boldsymbol{D_V}$). El tamaño de $\boldsymbol{D_E}$ generalmente es mayor al $\boldsymbol{D_V}$ en proporciones 2/3 y 1/3, 4/5 y 1/5, etc. respectivamente. Los elementos del $\boldsymbol{D_E}$ suelen obtenerse mediante muestreo sin reemplazo de todo el conjunto de datos, mientras que el conjunto $\boldsymbol{D_V}$ lo conforma las observaciones restantes que no pertenecen al $\boldsymbol{D_E}$. Suele ser aplicado a un conjunto de datos. (Perez, M. 2014)

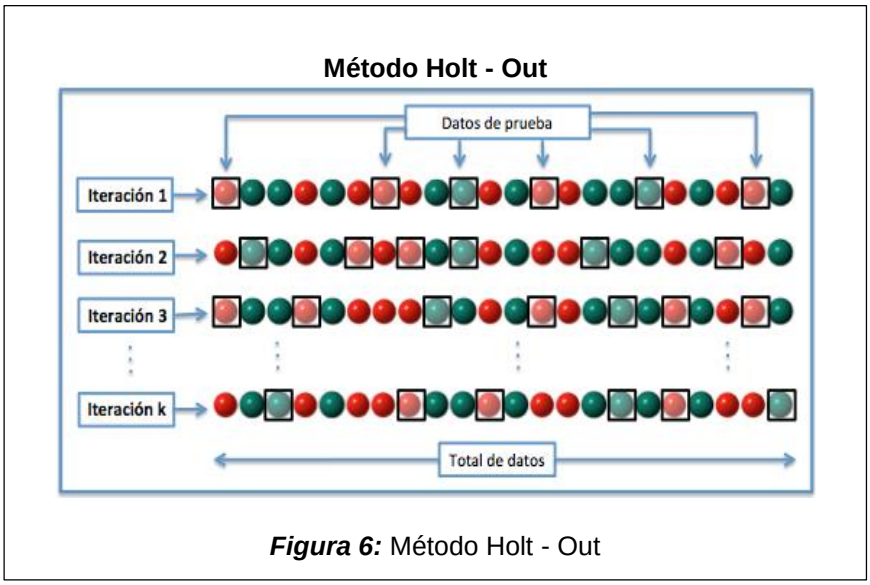


Figura 6: Método Holt - Out

Bondad Predictiva: Tabla de clasificación. Conocida también como matriz de confusión, se convierte en un instrumento de suma importancia a la hora de medir la denominada bondad predictiva de un modelo. Es una tabla de contingencia que permite ver los errores producidos por el modelo a la hora de clasificar a la variable dependiente, la misma que tiene que ser categórica.

Aquí es donde debemos tener en cuenta el esquema de la variable destino, tal como lo nombra SPSS MODELER, dentro de la minería de datos. La mencionada variable es una variable categórica que toma valores asignados a una característica o atributo del objetivo a estudiar.

Si para ejemplificar, suponemos que nuestra variable asume:

1: Presencia del atributo o característica

0: Ausencia del atributo o característica

0 es la categoría base la que nos sirve de punto de comparación, además de ello nos movemos en el campo de las probabilidades 1 sería la condición de éxito y 0 se correspondería con fracaso en un evento probabilístico. El análisis se completa con la determinación del denominado punto de corte, de tal manera que para las predicciones del modelo, se establece como punto de corte 0.5, entonces valores >0.5 se consideran 1 y menores que 0.5 se consideran con valor 0

Esto es lo que resume la matriz de confusión:

Tabla 1:

Tabla de Clasificación

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fuente:

VP: Se denomina verdaderos positivos, es decir valores de predicción concordantes con los valores observados, es decir equivalentes a 1 (Paga)

VN: Se denomina verdaderos negativos, es decir valores de predicción equivalentes a los observados o lo que es lo mismo 0 (No paga para nuestro caso)

FN: Nombramos de esta manera a los positivos que erróneamente fueron clasificados como negativos en la predicción

FP: Se designa así a la cantidad de negativos que fueron erróneamente como positivos en la predicción realizada.

De esta matriz de confusión, se establecen un conjunto de indicadores a fin de dar consistencia a la evaluación del modelo predictivo. Estos indicadores son:

a) Exactitud: Es una medida que permite determinar el porcentaje de clasificación correcta hecha por el modelo en su etapa de predicción, en términos de formula implica:

$$Exactitud = \frac{VP + VN}{(VP + FN + FP + VN)}$$

b) Tasa de Error: Permite determinar en qué porcentaje, en su etapa, de predicción, el modelo clasifica erróneamente a la data.

$$Tasa\ de\ error = \frac{FP + FN}{(VP + FN + FP + VN)}$$

c) Sensibilidad: Es una medida de la capacidad de acierto de un evento y se define como el número de categorías positivas bien predichas dividido por el total de categorías positivas.

$$x$$

$$Sensibilidad = \frac{VP}{(VP + FN)}$$

d) Especificidad: Es una medida de la capacidad de acierto del evento complementario al anterior. Se define como el número de categorías falsas bien predichas (el modelo también dice que son falsas) dividido por el total de categorías falsas.

$$Especificidad = \frac{\mathrm{VN}}{\mathrm{FP} + \mathrm{VN}}$$

e) Precisión Positiva: En el proceso de predicción, el modelo en estudio, que porcentaje es clasificado correctamente como positivo

$$Precision\ Positiva = \frac{VP}{(VP + FP)}$$

f) Precisión Negativa: En el esquema anterior que porcentaje es clasificado correctamente como negativo

$$Precicion\ Negativa = \frac{VN}{FN + VN}$$

La curva ROC. Constituye una herramienta estadística fundamental en la determinación de la capacidad de clasificación en un modelo con variable dependiente dicotómica. La curva propiamente dicha resulta de la representación, por cada punto de corte, de sensibilidad y especificidad. El área bajo

La curva (AUC)

$$AUC = \int_0^1 ROC(p)dp$$

permite determinar la capacidad del modelo de distinguir entre los valores que toma la variable endógena binaria. Si el área bajo la curva tendría el valor de 1 (100% de la curva ROC se orientaría hacia la esquina superior izquierda). Significaría que el modelo clasificaría el atributo de la variable destino 100% como que existe y 100% como que no existe. Vale decir que la clasificación seria exacta.

Si el modelo no ayuda en la clasificación, la sensibilidad es igual a la tasa de falsos positivos (1 - especificidad) y la curva es la diagonal que va de (0, 0) a (1, 1).

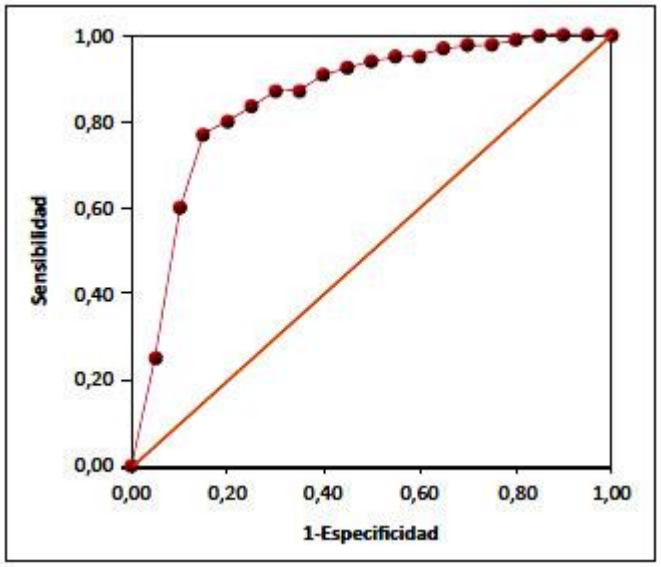


Figura 7: Curva ROC

El modelo tiene mejor desempeño si la curva ROC correspondiente se aleja más de la diagonal principal.

Tabla 2:

La Curva ROC

AREA	CAPACIDAD DE CLASIFICACION
ROC=0.5	NULO
0.6<=AREA ROC<0.8	ACEPTABLE
0.8<=AREA ROC<0.9	EXCELENTE
AREA ROC>=0.9	EXCEPCIONALMENTE BUENA

Fuente: Elaboración propia

Otras medidas de evaluación de los modelos. Para evaluar la capacidad predictiva de un modelo se utilizan varios estadísticos alternativos. (Pérez, M. 2014) Siendo n el horizonte de predicción, los estadísticos más habituales para la evaluación de la capacidad predictiva son los siguientes:

a) Error Cuadrático Medio: sirve para determinar la medida en las predicciones del modelo se ajustan a la información real

$$ECM = \sum_{i=1^{\circ}}^{n} (\widehat{Y}_i - Y_i)^2$$

b) Raíz del error cuadrático medio (Rott Mean Squared Error): es la raíz de ECM y mide con más precisión el grado de ajuste de los valores de predicción con respecto a los reales. El modelo "discrimina" mejor cuanto menor es RECM

$$RECM = \sqrt{\frac{\sum_{i=1^{\circ}}^{n} (\widehat{Y}_i - Y_i)^2}{n}}$$

c) Error Absoluto medio (Mean Absolute Error):es un promedio de los errores absolutos de la predicción y de la misma manera cuanto menor mejor ajusta el modelo a los predichos con el valor real.

$$EAM = \frac{\sum_{i=1^{\circ}}^{n} \left|\widehat{Y}_i - Y_i\right|}{n}$$

d) Error Absoluto Medio Porcentual es la suma del porcentaje que representa cada error respecto al valor real. Este porcentaje debe asumir un valor lo más bajo para tener en consideración que el modelo está prediciendo adecuadamente.

$$EAMP = \frac{1}{n} \sum_{i=1^{\circ}}^{n} \left| \frac{\widehat{Y}_i - Y_i}{Y_i} \right|$$

e) Coeficiente de desigualdad de theil: generalmente se toma en consideración a este coeficiente como indicador de la exactitud de la predicción, 0 <= CDT <=1, la predicción será perfecta cuando CDT=0, por lo que es conveniente que este coeficiente sea lo más cercano a 0.

$$CDT = \frac{\sqrt{\frac{\sum_{i=1^{\circ}}^{n}(\hat{Y}_i - Y_i)^2}{n}}}{\sqrt{\frac{\sum_{i=1^{\circ}}^{n}\hat{Y}_i^2}{n}} + \sqrt{\frac{\sum_{i=1^{\circ}}^{n}Y_i^2}{n}}}$$

Habría que acotar que todos estos indicadores tendrían que ser aplicados a la muestra de validación, puesto que es allí donde estamos probando la capacidad y bondad predictiva del modelo

Medidas para la selección de modelos. Si se tiene un conjunto de modelos M1, M2, … con parámetros K1, K2, … respectivamente, dos medidas para compararlos son el Criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC). Ambos criterios usan el Log-Verosimilitud y penalizan la complejidad del modelo. En términos prácticos la función de verosimilitud se estima de la siguiente manera:

$$£= \frac{-\mathrm{I}}{2}[1 + \ln(2\pi) + ln\frac{\sum e^2}{\mathrm{I}}]$$

Donde I es el tamaño de la muestra y e son los residuos.

a) El criterio de información de Akaike (AIC): Akaike (1973) propuso un enfoque para resolver el problema de seleccionar el modelo suponiendo que el objetivo es hacer predicciones tan precisas como sea posible. Plantea de la siguiente manera:

$$AIC = -2 * £ + 2 * K$$

Donde K es el número de parámetros del modelo M_i, £ la función de verosimilitud estimada.

b) Criterio de BIC: El criterio de BIC (Bayesian Information Criterion) está planteado de la siguiente manera:

$$BIC = * 2 * £ + K * \ln(N)$$

Lo nuevo aquí es N que es el tamaño de la muestra utilizado para estimar el modelo. Para escoger el modelo adecuado, utilizando

cualquiera de estos indicadores, se escogerá el modelo que tenga el menor AIC o BIC.

2.2.3.3. Tipos de técnicas de clasificación

Regresión logística. Esta técnica, es muy útil cuando el modelo, en su variable destino (endógena para la Econometría) tiene una estructura dicotómica o politómica y representa la probabilidad de ocurrencia de un suceso.

El modelo de regresión logística, surge en respuesta a la especificación del denominado modelo lineal de probabilidad, con el cual se tiene problemas en lo referente al intervalo de probabilidad 0, 1 además de la linealidad inherente al mismo que no concuerda con aspectos reales, no se puede decir por ejemplo que ante mayor ingreso de un agente económico la probabilidad de mayor consumo seria la misma. Esto motivo a tener que modelar tomando en consideración distribuciones de probabilidad tal como la logística para el modelo logit y la normal para el modelo probit, cuyo comportamiento ya no es lineal:

$$Yi = \frac{1}{1 + e^{-X_i B}} + U_i = \frac{e^{X_i B}}{1 + e^{X_i B}} + U_i$$

Funcionalmente el modelo puede escribirse:

$$Y_i = \Lambda(X_i B) + u_i$$

Donde:

Λ : se refiere a la distribución logística

X: Vector de variables predeterminadas (entradas)

u: Variable aleatoria que se distribuye normalmente con media cero y varianza constante.

B: Vector de parámetros desconocidos, materia de estimación.

Y: toma valores cero o la unidad

Para el ajuste de este modelo y la estimación del vector de parámetros no puede aplicarse los mínimos cuadrados, pues la relación ya no es lineal, entonces este problema se soluciona aplicando el procedimiento de estimación denominado de máxima verosimilitud La estimación pasa por maximizar la siguiente función de verosimilitud:

$$\mathcal{L}(\beta) = \prod_{i\ /\ y_i=1} \frac{1}{1+e^{\beta_0+\beta^t.x}} \prod_{i\ /\ y_i=0} \left(1 - \frac{1}{1+e^{\beta_0+\beta^t.x}}\right)$$

Con esta función se logran los estimadores máximos verosímiles los mismos que cumplen con las propiedades de todo buen estimados.

En el uso práctico, se selecciona un punto de corte, de tal forma que valores mayores al mencionado punto, equivalen a 1 y valores menores son 0, por lo que claramente queda establecido el esquema de discriminación que es el uso que se le da a estos modelos.

Para la interpretación del modelo se pasa por la siguiente expresión:

$\text{Prob}(= 1/X_i\ ) = P_i$

Es decir, conocidos los valores del vector de variables X, asignamos una probabilidad de que la variable Y valga la unidad. Gráficamente está representado de la siguiente manera:

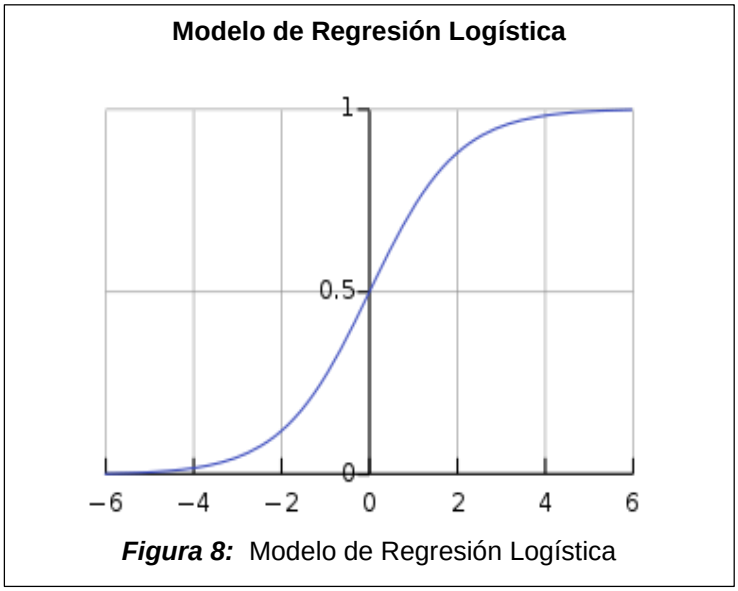


Figura 8: Modelo de Regresión Logística

Contraste de hipótesis para la regresión logística. Teniendo en consideración la estimación de parámetros del modelo Logit, metodológicamente se continua con lo que se denominaría las pruebas preliminares al modelo.

Por la naturaleza el modelo las pruebas son las siguientes:

a) La razón de verosimilitud: Se calcula a partir de la función de verosimilitud, para ello se establece el procedimiento que consiste en:

- Estimar un primer modelo sin variable alguna, calcular su función de verosimilitud, le denominamos modelo con restricciones: L_{CR}
- Estimar un segundo modelo con todas las variables del caso, se denomina modelo sin restricciones: L_{SR}
- Se establece el estadístico de prueba de la siguiente manera:

LR= L (Modelo sin variables)- L (Modelo con variables)

Donde L representa las funciones de verosimilitud para cada modelo. Es decir:

$$LR = -2\ln\left(\frac{L_{CR}}{L_{SR}}\right) = -2(lnL_{CR} - lnL_{SR})$$

Este estadístico se distribuye como una X^2 con grados de libertad igual a número de restricciones.

b) Relevancia Individual: Una de las formas más comunes de contrastar la hipótesis en lo que se denomina relevancia individual es plateando que el parámetro de regresión es cero ($\boldsymbol{H_0}: \boldsymbol{B_k} = \mathbf{0}$). Siendo el estadístico de contraste:

$$\text{Prob}\left(-N_{\frac{\alpha}{2}} < \frac{\hat{B}_k - B_k}{S_{\hat{B}_k}} < N_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

N representa a la distribución normal y se acepta la hipótesis nula planteada si se cumple la desigualdad para uno de los parámetros del modelo.

c) Prueba de la ji – cuadrado. Esta prueba es medida de ajuste de bondad que se basan en comparar los valores observados y los estimados por el modelo que se desea evaluar (valores esperados), todo ello, una vez más, bajo la H_0 de que dicho modelo ajusta bien los datos observados. La prueba se basa en la obtención de un estadístico X^2 que mide el nivel de discordancia que puede existir al comprar, para cada uno de los diferentes patrones de predictores existentes, el número de respuesta (afirmativas) observadas con la probabilidad estimada por el modelo. Por patrón de predictores se entiende cada una de las diferentes combinaciones de valores que pueden adoptar las variables independientes incluidas en el modelo (Jovell, 1995). El estadístico X^2, cuando el número de patrones de predictores M<N, es:

$$X^2 = \sum_{i=1}^{M} \frac{m_i(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)}$$

Donde m_i es el número de casos incluidos en cada patrón de predictores, y_i la opción de la variable respuesta y $\hat{p}_i$ la probabilidad estimada por el modelo para el patrón de covariables i. Para grandes muestras el estadístico se distribuye, como una ji-cuadrado con $M - p$ grados de libertad.

En presencia de variables continuas, el número de patrones de predictores es muy probable que sea igual al número de observaciones muestrales $M \approx N$. En estos casos la prueba X^2 tomaría la expresión:

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)}$$

Puesto que n_i seria igual a 1. Hosmer y Lemeshow (1989) advierten de la obtención de valores p incorrectos cuando $M \approx N$; sin embargo, en

los casos en que el modelo ajustado es el correcto, se puede utilizar la prueba X^2 con $N - p$ grados de libertad con unos resultados razonables.

d) Prueba de Hosmer – Lemeshow: Esta prueba es especialmente adecuada para evaluar la bondad del ajuste de aquellos modelos que incluyan una o varias variables independientes de tipo continuo y que cuenten con un número de patrones de predictores prácticamente igual al número de casos observados ($M \approx N$). Estos autores proponen ordenar de menor a mayor las N probabilidades estimadas (una para cada caso observados) y a continuación agruparlas en diez grupos de tal modo que en el primero de ellos se encuentren los $n_1 = N/10$ sujetos que tengan las probabilidades estimadas más bajas y en el ultimo los $n_{10} = N/10$ sujetos con las probabilidades estimadas mas elevadas. Estos grupos son conocidos Hosmer – Lemeshow, como "deciles de riesgo" (Sharma, 1996). El estadístico de bondad del ajuste de $\hat{C}$, se obtiene calculando el estadístico ji-cuadrado de Pearson de una tabla de 2x10 referida a las frecuencias observadas y estimadas para cada uno de los diez grupos. Aunque los principales paquetes estadísticos que desarrollan la regresión logística ofrecen una salida con el resultado de esta prueba, reproducimos la fórmula de cálculo de $\hat{C}$:

$$\hat{C} = \sum_{k=1}^{10} \frac{(o_k - n_k \bar{p}_k)^2}{n_k \bar{p}_k (1 - \bar{p}_k)}$$

Donde n_k es el número de patrones de predictores del grupo $k - ésimo$,

$$o_k = \sum_{i=1}^{n_k} y_i$$

Es decir, el número de respuesta afirmativas registradas para la variable respuesta (Y=1) para los $n_k$patrones de predictores, y la media de la probabilidad estimada.

$$\bar{p}_k = \sum_{i=1}^{n_k} \frac{m_i \hat{p}_i}{n_k}$$

Arboles de decisión: Esta técnica, implica un análisis no simultaneo de las variables denominadas variables de entrada, predeterminadas o explicativas, sino un examen una a una. Estos árboles de decisión también suelen denominarse arboles de clasificación y constituyen particiones secuenciales del conjunto de información a disposición del investigador a fin de hacer máxima las diferencias de la variable respuesta, comienzan con un denominado nodo para luego ramificarse adecuadamente.

En la minería de datos, los árboles de decisión describen datos, y los clasifican, de tal manera que los resultados son utilizados justamente en la toma de decisiones.

En minería de datos los árboles de decisión se clasifican en:

- Arboles de clasificación: Cuando la variable destino es de naturaleza discreta, es decir de naturaleza binomial.

- Arboles de regresión: Cuando la variable destino está compuesta por números reales, como por ejemplo precio de un automóvil.

En Minería de Datos (DM): Los árboles de decisión son usados en este ámbito abordando problemas como predicción, clasificación y segmentación de datos con el fin de convertirlos en información valiosa para el análisis y toma de decisiones. Lograr una buena minería de datos depende de algoritmos, algunos más y otros menos sofisticados que se aplican a los árboles de decisión para obtener la respuesta la información óptima de los datos. (Witten & Eibe, 20015)

La metodología a seguir en el análisis con árboles de decisión lo podemos resumir de la siguiente manera:

➢ Especificación de criterios: se trata de minimizar costes en el sentido de proporción de casos reales clasificados mal, uso no adecuado del

cálculo de probabilidades. Pero toda vez que los algoritmos utilizados en minería de datos tienen que ver con el aprendizaje automático y la inteligencia artificial automáticamente estos costos son minimizados.

- Método de división: se trata de escoger el mejor método para llevar adelante el proceso de división a los distintos niveles. Los métodos son básicamente estadísticos y tienen que ver con los denominados métodos exhautivos y discriminante, en el primero de los casos entra en funcionamiento las técnicas estadísticas de comparación basados en la bondad de ajuste mediante una serie de medidas estadísticas, X^2 p_value entre otras de interés estadístico. el otro método denominado discriminante tiene que ver con la utilización previamente del método anterior y posteriormente aplicar la técnica de minería de datos denominada de discriminación.

- Tamaño adecuado: si es que no queda establecido límite alguno para el numero de divisiones dentro del árbol de decisiones, se corre el riesgo de tener un árbol en el cual su estructura frondosa no permitirían análisis y conclusiones de la mejor manera. Este problema es tratado con lo que se denomina conjunto de reglas, las mismas que tienen que ver el con el algoritmo utilizado y automáticamente permite determinar el máximo número de niveles y nodos para un árbol que se está utilizando en la investigación.

a) Tipos de Arboles de decisión:

- Arboles CHAID: Estos árboles identifican variables importantes que tienen que ver con la variable destino (dependiente), esta puede ser cualitativa (nominal u ordinal) o cuantitativa. Para el caso de variables cualitativas, el proceso utiliza básicamente X^2 para encontrar relación entre la dependiente y las predictoras (independientes). En el caso de variable cuantitativa se utilizan métodos de análisis de varianza fundamentalmente, para determinar de forma óptima las divisiones de tal manera que minimicen la varianza de la variable dependiente. Mediante procesos iterativos y

de comparación de indicadores o pruebas estadísticas se continua hasta que se activa la regla de parada del proceso.

- Arboles cart: Este tipo de árbol, surge en virtud a algunas deficiencias del CHAID antes que su variante exhautiva, lo caracteriza una estructura estadística más fuerte que el método chaid, por lo que es muy usual en investigaciones médicas, y de finanzas. El proceso empieza dividiendo la data en varios subconjuntos de tal manera que estos son materia de análisis utilizando para ello los indicadores estadísticos tales como el error cuadrático medio para el caso de variable dependiente cuantitativa y el coeficiente de Gini para el caso de variable dependiente cualitativa, evaluando la probabilidad de una mala clasificación.

- Arboles QUEST: Los arboles QUEST son un algoritmo de clasificación, calcula en cada nodo la asociación entre la dependiente y el predictor mediante la prueba F del análisis de la varianza para predictores de naturaleza continua y ordinales o mediante X^2 para predictores de naturaleza nominal.

El aspecto fundamental que tienen en cuenta este tipo de árboles es el sesgo en la selección de variables, elige el predictor que esta más asociado con la dependiente y para hallar el punto de corte optimo se recurre al análisis discriminante, esto constituye un proceso recursivo que termina cuando entra en juego las reglas de parada.

Otros Algoritmos. En lo que respecta a minería de datos, hay que tener en cuenta que en la actualidad se ha desarrollado software especializados que contienen algoritmos especializados también, tale es el caso del software IBM SPSS MODELER que contiene además de los anteriores el algoritmo denominado C5.0 que tiene que ver con el aprendizaje automático y la inteligencia artificial, además de ello es preciso anotar que se cuenta con un algoritmo que se denomina clasificador automático y que determina a partir de la información cuál es el modelo o modelos adecuados a nuestros datos disponibles.

3.3. Variables

Las variables básicas consideradas en el estudio son las que se muestrean en el siguiente cuadro.

Tabla 3:

Descripción de variables

Variables	Descripción	Valores
X_1 (Monto)	***Monto del préstamo***	***Soles***
X_2 (Estciv)	***Estado Civil***	***Soltero=0, Casado=1***
X_3 (Tarjet)	***Tarjetas***	***Menos de 3=0, 3 o más=1***
X_4 (Ingresot)	***Ingresos***	***Menores=0,Mayores=1***
X_5 (Edadt)	***Edad***	***Menores o iguales a 30 años =0; mayores de 30=1***
Y (Pago)	***Condición de Pago***	***No paga= 0 Paga=1***

CAPITULO III

RESULTADOS

CAPÍTULO III: RESULTADOS

3.1. Análisis y Discusión de los Resultados de o los Instrumento utilizados

3.1.1. Materiales, Técnicas e Instrumentos de Recolección de datos.

Para la obtención de datos se obtuvo el historial crediticio de los clientes pertenecientes a la cartera de crédito personal obtenido de la base de datos de la institución crediticia objeto de estudio que fueron de 2356 clientes. Cabe resaltar que la información brindada a los investigadores no incumple o infringe la ley del secreto bancario, pues no se muestra nombres, apellidos o cualquier otro dato que identifique los clientes.

3.1.1.1. Análisis estadísticos de los datos.

En el modelamiento de datos se utilizó las técnicas de Regresión Logística y Árboles de Clasificación, una vez hallado los modelos predictivos se procedió a comparar y evaluar utilizando las tablas de clasificación, área bajo la curva ROC para poder determinar el mejor modelo predictivo para el otorgamiento del crédito personal.

El análisis se realizó mediante el software estadístico SPSS V.25, STATA 15, IBM SPSS MODELER, este último contiene algoritmos exclusivos de la minería de datos.

3.2. Análisis, Interpretación y discusión de resultados

3.2.1. Validación Regresión Logística y árbol de decisión: muestra completa

3.2.1.1. Validación Aparente Regresión Logística

Esta forma de medir el rendimiento de un modelo se conoce como validación aparente, es decir que, en el modelo predictivo, estamos usando toda la información del mismo para evaluar su capacidad de predecir comportamientos futuros de agentes económicos.

En lo que a la probabilidad de ocurrencia se refiere, en nuestro caso la variable dicotómica Pago (1: paga 0: no paga), siendo nuestro objetivo la probabilidad de pago.

Claramente podemos observar:

✓ La razón de verosimilitud, que presenta el cuadro # 9 de STATA explica claramente que a nivel de significación global el modelo discrimina adecuadamente entre los que pagan y no pagan puesto que el valor es:

LR CH(5)=287.8

Prob > CHI=0.00000

Toda vez que el denominado valor P (P_Value) es menor que 0.0 5 (5% de nivel de significación) se va a rechazar cualquier hipotesis relacionada a la no discriminación adecuada.

✓ De la misma manera la salida de IBM SPSS MODELER en el cuadro # para la prueba de *Hosmer y Lameshow* , se presentan los siguientes resultados:

Chi Cuadrado = 32.941

siendo el tabulado (de las tablas)

Chi Cuadrado (8 g.l.) =15.5072

Por lo que, por regla general, si el calculado es mayor que el tabulado, se rechaza cualquier hipotesis contraria a la discriminación adecuada en este caso.

✓ Otra forma de probar la capacidad de discriminación del modelo logístico es lo que se denomina la CURVA ROC, mostramos la siguiente figura:

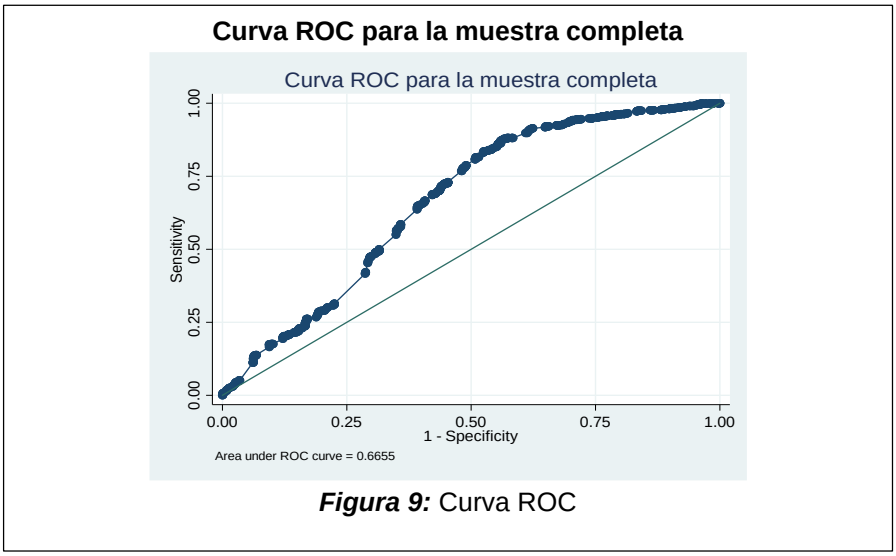


Figura 9: Curva ROC

Lo que se puede observar en el grafico es que el área bajo la curva equivale a 0.6565, lo que para muchos autores representa un nivel de discriminación aceptable

✓ En lo que respecta a la denominada "significación individual", propia de un análisis de modelos explicativos, debemos mencionarla pero sin que esta signifique eliminación de variables que no son estadísticamente significativas. Tenemos entonces dos esquemas al respecto que mostramos a continuación:

Tabla 4:

Significación Individual

VARIABLE	***STATA*** Z	***SPSS MODELER*** Wald	***Valor P STATA Y SPSS MODELER***
Monto	-0.56	0.309	0.578
Estciv	-6.03	36.395	0.000
Tarjet	-13.8	190.339	0.000
Ingresot	0.18	0.033	0.855
Edadt	2.87	8.214	0.004

Fuente: Elaboración Propia

Visto desde el punto de vista de los modelos explicativos, del cuadro anterior concluiríamos que las variables que no deben estar en el

modelo son el moto y el ingreso puesto que su P-Value es mayor que 0.05, no siendo este el caso con las demás variables del modelo.

- ✓ En este caso también es necesario tener en cuenta la denominada matriz de confusión, que es fundamental para determinar con mayor precisión el grado de discriminación en el modelo logístico. De la Tabla 5 del anexo de SPSS MODELER y 9 de STATA , extraemos lo siguiente:

Tabla 5:

Matriz de Confusión (Tabla de Clasificación)

	NO PAGA	NO PAGA
NO PAGA	541	637
PAGA	186	992

Fuente: Elaboración propia

Esta tabla explica con claridad la forma como el modelo ha discriminado entre las dos alternativas de la variable Pago. La diagonal principal muestra los valores exactamente predichos por el modelo. Cabe mencionar que esta predicción es dentro del periodo muestral, es decir corresponde a cada uno de los individuos: La predicción de los que no pagan observada es 541 y la estimación (predicción) es 541; de igual manera de los catalogados como que pagan la predicción es 992 y efectivamente fueron 992 que pagan (=1). Cabe precisar que las predicciones están en probabilidades cuyo intervalo es: 0<= Pr<=1, se toma entonces un punto de corte que generalmente es 0.5, valores mayores o iguales de la predicción se asumen como 1, es decir la probabilidad de ocurrencia del suceso y valores menores al punto de corte toman el valor cero. Se extrae entonces un conjunto de indicadores a partir de la matriz de confusión, siendo el más comúnmente utilizado el denominado exactitud que se calcula como:

$$Exactitud = \frac{541 + 992}{541 + 637 + 186 + 992} = 0.650679 = 65.07\%$$

Es decir, el modelo predice con exactitud 65.07%. de las observaciones. A la diferencia, como es lógico, se denomina tasa de error en nuestro caso sería:

Tasa de error=1-0.6507=34.93%

El cálculo de la sensibilidad será el siguiente:

$$Sensibilidad = \frac{992}{992 + 186} = 0.8421 = 84.21\%$$

992 (Pr=1) son los clasificados correctamente, 186 son positivos (Pr=1) que el modelo clasifico erróneamente como negativos (Pr=0). 637 son negativos (Pr=0) que el modelo clasifico erróneamente como positivos (Pr=1).

A la sensibilidad se denomina la capacidad de acierto, y en este caso esa capacidad es considerable 84.21%.

$$Especificidad = \frac{541}{541 + 637} = 0.45925 = 45.93\%$$

Como se puede apreciar, este indicador está mostrando la capacidad de acierto del evento complementario del modelo, en este caso Pago=0. Para el caso, la capacidad es de 45.93%

En el eje de las X de la curva ROC, aparece 1-especificidad, que se puede calcular de la siguiente manera:

$$1 - Especificidad = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{VN}}$$

FP: se denominan falsos positivos, pues siendo negativos, el modelo los ha clasificado como positivos

VN: se denominan verdaderos negativos y el modelo los ha clasificado como tal.

En lo que respecta a la interpretación de los parámetros del modelo, partimos de las tablas mostradas a continuación:

Tabla 6:

Resultados de la Regresión Logística

```
. logit Pago Monto Estciv Tarjet Ingresot Edadt

Iteration 0:   log likelihood = -1633.0548
Iteration 1:   log likelihood = -1489.9433
Iteration 2:   log likelihood = -1489.1268
Iteration 3:   log likelihood = -1489.1256
Iteration 4:   log likelihood = -1489.1256

Logistic regression                             Number of obs     =      2,356
                                                LR chi2(5)        =     287.86
                                                Prob > chi2       =     0.0000
Log likelihood = -1489.1256                     Pseudo R2         =     0.0881

------------------------------------------------------------------------------
        Pago |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Monto |  -.0000203   .0000366    -0.56   0.578    -.0000921    .0000514
      Estciv |  -.5874092   .0973684    -6.03   0.000    -.7782478   -.3965706
      Tarjet |  -1.967942   .1426423   -13.80   0.000    -2.247515   -1.688368
    Ingresot |   .0255312   .1396465     0.18   0.855    -.2481708    .2992332
       Edadt |   .2626608   .0916459     2.87   0.004     .0830381    .4422835
       _cons |   .4020465   .0849965     4.73   0.000     .2354565    .5686365
------------------------------------------------------------------------------
```

Fuente: Elaboración propia

***Tabla 7*:**

Regresión Logística teniendo en cuenta la estructura binaria de las variables independientes.

```
. logit Pago Monto i.Estciv i.Tarjet i.Ingresot i.Edadt

Iteration 0:   log likelihood = -1633.0548
Iteration 1:   log likelihood = -1489.9433
Iteration 2:   log likelihood = -1489.1268
Iteration 3:   log likelihood = -1489.1256
Iteration 4:   log likelihood = -1489.1256

Logistic regression                             Number of obs     =      2,356
                                                LR chi2(5)        =     287.86
                                                Prob > chi2       =     0.0000
Log likelihood = -1489.1256                     Pseudo R2         =     0.0881

------------------------------------------------------------------------------
        Pago |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Monto |  -.0000203   .0000366    -0.56   0.578    -.0000921    .0000514
             |
      Estciv |
      casado |  -.5874092   .0973684    -6.03   0.000    -.7782478   -.3965706
             |
      Tarjet |
3 O mas tarjetas |  -1.967942   .1426423   -13.80   0.000    -2.247515   -1.688368
             |
    Ingresot |
Ingresos altos |   .0255312   .1396465     0.18   0.855    -.2481708    .2992332
             |
       Edadt |
Mayor a 36 años |   .2626608   .0916459     2.87   0.004     .0830381    .4422835
       _cons |   .4020465   .0849965     4.73   0.000     .2354565    .5686365
------------------------------------------------------------------------------
```

Fuente: Elaboración propia

Como se trata de un modelo cuya variable destino (Endógena), binaria, no se puede interpretar como lo haríamos con un modelo explicativo de

regresión lineal lo que se puede decir es que teniendo como base los signos interpretar el tipo de relación de la probabilidad de ocurrencia del evento, en este caso **Pago=1** el cliente si paga. A ese respecto, podemos decir

- El monto es variable continua, si aumenta disminuye la probabilidad de pago en 0.0000203
- En lo que respecta al estado civil, este tiene una relación inversa con la probabilidad de **Pago=1**
- De la misma manera, el número de tarjetas se relaciona en forma inversa con la probabilidad de pago
- Los ingresos tienen una relación directa con la probabilidad de ocurrencia del suceso igual a 1
- La variable edad se relaciona directamente con la probabilidad de pago.

Para tener en cuenta los efectos marginales, ya cuantificados de las variables explicativas hacia la variable de destino (endógena), presentamos el siguiente cuadro:

***Tabla 8*:**

Efectos marginales de las variables explicativas

```
. mfx

Marginal effects after logit
      y  = Pr(Pago) (predict)
         =  .48753034
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
Monto	-5.08e-06	.00001	-0.56	0.578	-.000023	.000013	1717.06
Estciv*	-.1448809	.02347	-6.17	0.000	-.190881	-.098881	.310272
Tarjet*	-.4154761	.02131	-19.50	0.000	-.457247	-.373706	.17657
Ingresot*	.0063802	.0349	0.18	0.855	-.062029	.074789	.114177
Edadt*	.0655578	.02282	2.87	0.004	.020839	.110276	.418081

(*) dy/dx is for discrete change of dummy variable from 0 to 1

En este caso, podemos afirmar lo siguiente:

- Si el monto aumenta en una unidad (Miles de soles) la probabilidad del suceso 1 (Paga) disminuye en 0.0000508
- Siendo 0 la categoría base (soltero), pasar a casado disminuye la probabilidad del suceso 1 en 0.1448809
- Del mismo modo, pasar de tener menos de 3 tarjetas a 3 o mas disminuye la probabilidad en 0.4154761 de ocurrencia del suceso pago=1
- Si el agente pasa de la categoría ingresos bajos a ingresos altos, la probabilidad de Pago=1 aumenta en 0.0063802
- En lo que respecta a la edad, si esta pasa de 36 años a mas, la probabilidad de pago aumenta en 0.0655578

La interpretación en términos de ODDS la extraemos del siguiente cuadro:

Tabla 9:

Odss Ratio en la regresión Logística

```
. logit Pago Monto i.Estciv i.Tarjet i.Ingresot i.Edadt,or

Iteration 0:   log likelihood = -1633.0548
Iteration 1:   log likelihood = -1489.9433
Iteration 2:   log likelihood = -1489.1268
Iteration 3:   log likelihood = -1489.1256
Iteration 4:   log likelihood = -1489.1256

Logistic regression                             Number of obs     =      2,356
                                                LR chi2(5)        =     287.86
                                                Prob > chi2       =     0.0000
Log likelihood = -1489.1256                     Pseudo R2         =     0.0881

------------------------------------------------------------------------------
        Pago | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Monto |   .9999797   .0000366    -0.56   0.578     .9999079    1.000051
             |
      Estciv |
      casado |   .5557653    .054114    -6.03   0.000     .4592099    .6726228
             |
      Tarjet |
3 O mas tarjetas |   .1397442   .0199334   -13.80   0.000     .1056614    .1848209
             |
    Ingresot |
Ingresos altos |    1.02586   .1432577     0.18   0.855     .7802266    1.348824
             |
      Edadt |
Mayor a 36 años |   1.300386    .119175     2.87   0.004     1.086583    1.556257
       _cons |   1.494881   .1270596     4.73   0.000     1.265486    1.765858
------------------------------------------------------------------------------
```

Fuente: Elaboración propia

Veamos un cuadro alterno para interpretación adecuada:

Tabla 10:

ODDS Ratio

VARIABLE	ODDS	INVERSA	RESULTADO
Estcivil	0.5557653	1/0.5557653	1.79932
Tarjet	0.1397442	1/0.1397442	7.15593
Ingresot	1.02586	----------	1.02586
Edadt	1.300386	----------	1.300386

Fuente: Elaboración propia

A los ODDS ratios menores que uno, se les encuentra su inversa para una mejor interpretación, no así los que son mayores a uno, entonces:

- Si el agente económico es casado, la oportunidad de pago es 1.79932 veces menor que si es soltero
- Si el solicitante tiene 3 o más tarjetas, la oportunidad de pago es 7.15593 veces menor que si tuviera menos de tres tarjetas.
- Perteneciendo el agente económico a la categoría altos ingresos, la oportunidad de pago será 1.02586 veces mayor que si perteneciera a la categoría ingresos bajos.
- Si el solicitante del préstamo tiene una edad mayor a 30 años la oportunidad de pago será 1.30 veces mayor que si tuviera 30 años o menos.

3.2.1.2. Validación Árbol de decisión

El siguiente modelo, que vamos a utilizar es el árbol de decisión cuyas características mencionamos más arriba, la evaluación de esta técnica lo haremos mediante la tabla de clasificación y la curva ROC y otros indicadores. Recordamos que este análisis es a nivel de muestra competa

Tabla 11:

Resultados para el campo de resultado 1 Paga 0 No paga

Comparando $R-Pago con 1 Paga 0 No paga		
Correctos	1.561	66,26%
Erróneos	795	33,74%
Total	2.356	

Es notorio que este modelo clasifica correctamente 66.26% y la curva ROC es la siguiente:

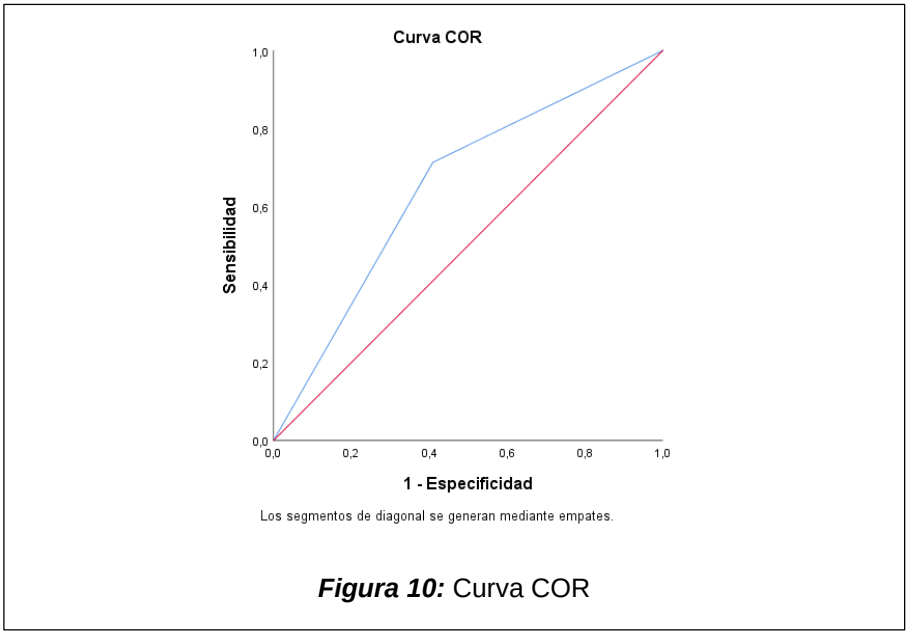


Figura 10: Curva COR

Tabla 12:

Área bajo la Curva

Variables de resultado de prueba: X$X_Pago

Área	Desv. Error[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,652	,011	,000	,630	,674

Fuente: Elaboración propia

Observamos con claridad que el área bajo la curva es 0.65, lo que representa una clasificación aceptable.

3.2.2. Validación por división de datos

Estos métodos consisten en dividir la muestra original en dos submuestras, una denominada de entrenamiento y otra de validación, el tamaño de ambas submuestras no está estandarizada y generalmente es 70% muestra de entrenamiento y 30% para la muestra de validación, del total muestral:

Entrenamiento: se refiere a los datos con los que llevo a cabo la construcción del modelo.

Validación: Parte de los datos que sirven para validar el modelo

3.2.2.1. Validación Regresión Logística: Etapa de Entrenamiento

La tabla mostrada a continuación, ilustra la validación en la etapa de entrenamiento de la Regresión logística:

Tabla 13:

Tabla de Clasificación Regresión Logística: Etapa de Entrenamiento

'Partición'	1_Entrenamiento	
Correctos	**1.109**	**66,33%**
Erróneos	**563**	**33,67%**
Total	**1.672**	

Fuente: Elaboración propia

Como se puede observar en la Tabla 12, la tabla de clasificación en la etapa de entrenamiento de la Regresión Logística, donde se tiene 1672 casos que corresponde el (70%) del total, 1109 han sido correctamente clasificados , por lo tanto, el modelo de la Regresión Logística clasifica en un 66.33%, de forma correcta, esto se considera un porcentaje apropiado.

De la misma manera, podemos observar la curva ROC, para la mencionada etapa:

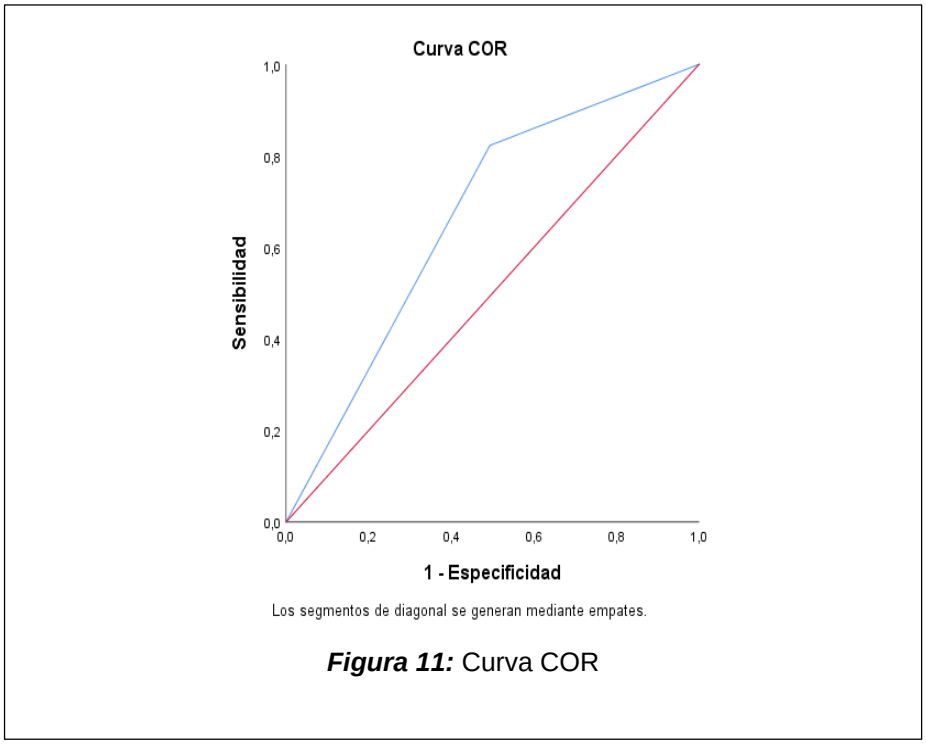


Figura 11: Curva COR

Tabla 14:

Área bajo la Curva Roc

Área bajo la curva

Variables de resultado de prueba: 1 Paga 0 No paga

Área	Desv. Error[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,665	,013	,000	,638	,691

Las variables de resultado de prueba: 1 Paga 0 No paga tienen, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

El complemento de la curva ROC lo constituye el cuadro 13 donde nos da el porcentaje bajo el área observando que este es 0.665, que es un valor que podemos considerar aceptable

3.2.2.2. Validación Regresión Logística: Etapa de Validación

La etapa de validación, implica tener en cuenta la capacidad predictiva del modelo en un "espacio" para el cual los parámetros no son tenidos en cuenta, equivale a decir como predice el modelo los valores conocidos para en este caso el 30% de la información restante ya que el 70% de la información sirvió para cuantificar el modelo en estudio. E nuestro caso vamos a utilizar las mismas herramientas de evaluación que en la etapa anterior, es decir la denominada tabla (matriz) de clasificación

Para nuestro caso, la mencionada tabla es:

Tabla 15:

Tabla de Clasificación Regresión Logística: Etapa de validación

'Partición'	2_Comprobación	
Correctos	450	65,79%
Erróneos	234	34,21%
Total	684	

Es notorio que, en esta etapa de validación, el modelo ha hecho predicciones correctas del 65.79%, lo que ubica como un modelo aceptable, es decir que de cada 100 casos la predicción ha sido correcta 65.79% de las veces. La curva ROC del caso lo mostramos a continuación:

Grafico Curva ROC Regresión Logística: Etapa de validación

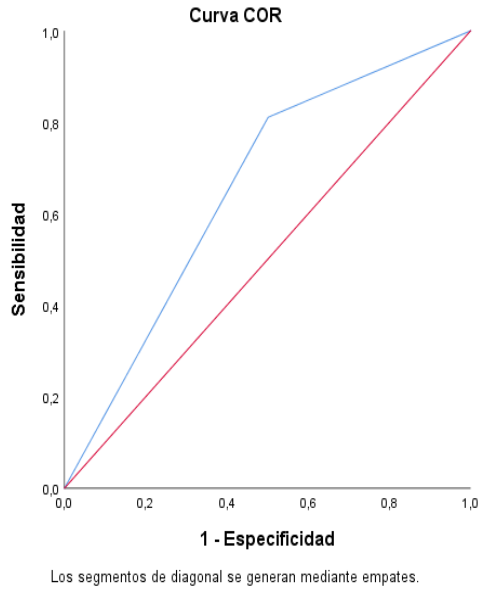


Figura 12: Curva COR

De la misma manera, el cuadro complementario es ilustrativo al respecto:

Tabla 16:

Área bajo la Curva

Área bajo la curva

Variables de resultado de prueba: 1 Paga 0 No paga

Área	Desv. Error[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,655	,021	,000	,613	,696

Las variables de resultado de prueba: 1 Paga 0 No paga tienen, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

Mediante este instrumento, podemos corroborar que la predicción es bastante aceptable puesto que si la curva coincidiera con la línea recta, las predicciones serian solo 50% y si la curva estuviera debajo de la curva, el modelo no nos serviría de nada para la predicción.

3.2.2.3. Validación cruzada e importancia de los predictores

En lo que respecta a la validación cruzada aplicamos la técnica leave-one-out y k-folds, cuyos algoritmos viene implementados en el software STATA.

En lo que respecta a la primera técnica tenemos los siguientes resultados

Leave-One-Out Cross-Validation Results

Method	Value
Root Mean Squared Errors	.47166819
Mean Absolute Errors	.44410709
Pseudo-R2	.11014102

Para el caso k-folds tenemos:

	RMSE
est1	.476542
est2	.483979
est3	.4613502
est4	.4642241
est5	.4706968
est6	.4612808
est7	.4750079
est8	.4774943
est9	.4843078
est10	.4597445

En lo referente a la influencia de cada predictor, se ha usado la técnica boost y tenemos los siguientes resultados:

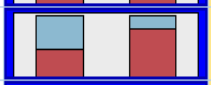
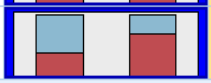
```
e(influence)[5,1]
                  c1
   Monto  58.225893
  Estciv  7.5260126
  Tarjet  23.054635
Ingresot  2.8660233
   Edadt  8.3274362
```

3.2.3. Validación arboles de clasificación

La otra técnica utilizada por la minería de datos a fin de "extraer conocimiento" a partir de los datos es la que se refiere a los denominados arboles de clasificación, para lo cual existen variados algoritmos que permiten su diseño y como es lógico su interpretación y a partir de allí la predicción del caso. Observemos como es el software quien indicaría cual modelo aplicar en este cao de las técnicas de clasificación.

Tabla 17:

Selección del Modelo de Clasificación a utilizar

¿Utilizar?	Gráfico	Modelo	Tiempo de generación (min)	Precisión general (%)
☑		C5 1	< 1	66,553
☑		Red neuronal 1	< 1	66,553
☑		CHAID 1	< 1	66,256

La indicación del software, es que los tres modelos mostrados son adecuados para este tipo de información, cabe destacar que en orden de precisión empatan el algoritmo C5.0 y la RED NEURONAL, seguido de otro árbol de clasificación denominado CHAID

Cabe mencionar que en minería de datos, el algoritmo C5.0 es el mas usual que construye los arboles a partir de un conjunto de datos de entrenamiento cuya decisión es la ganancia de información, tiene ventaja sobre otros algoritmos en lo que respecta al tiempo para construir el árbol, menor uso de memoria además de árboles con más precisión que sus antecesores como el C4.5 y otros tipos de algoritmos generadores de árboles tales como el CHAID por ejemplo en su versión mejorada: CHAID EXAHUSTIVO

Árbol de clasificación: muestra completa

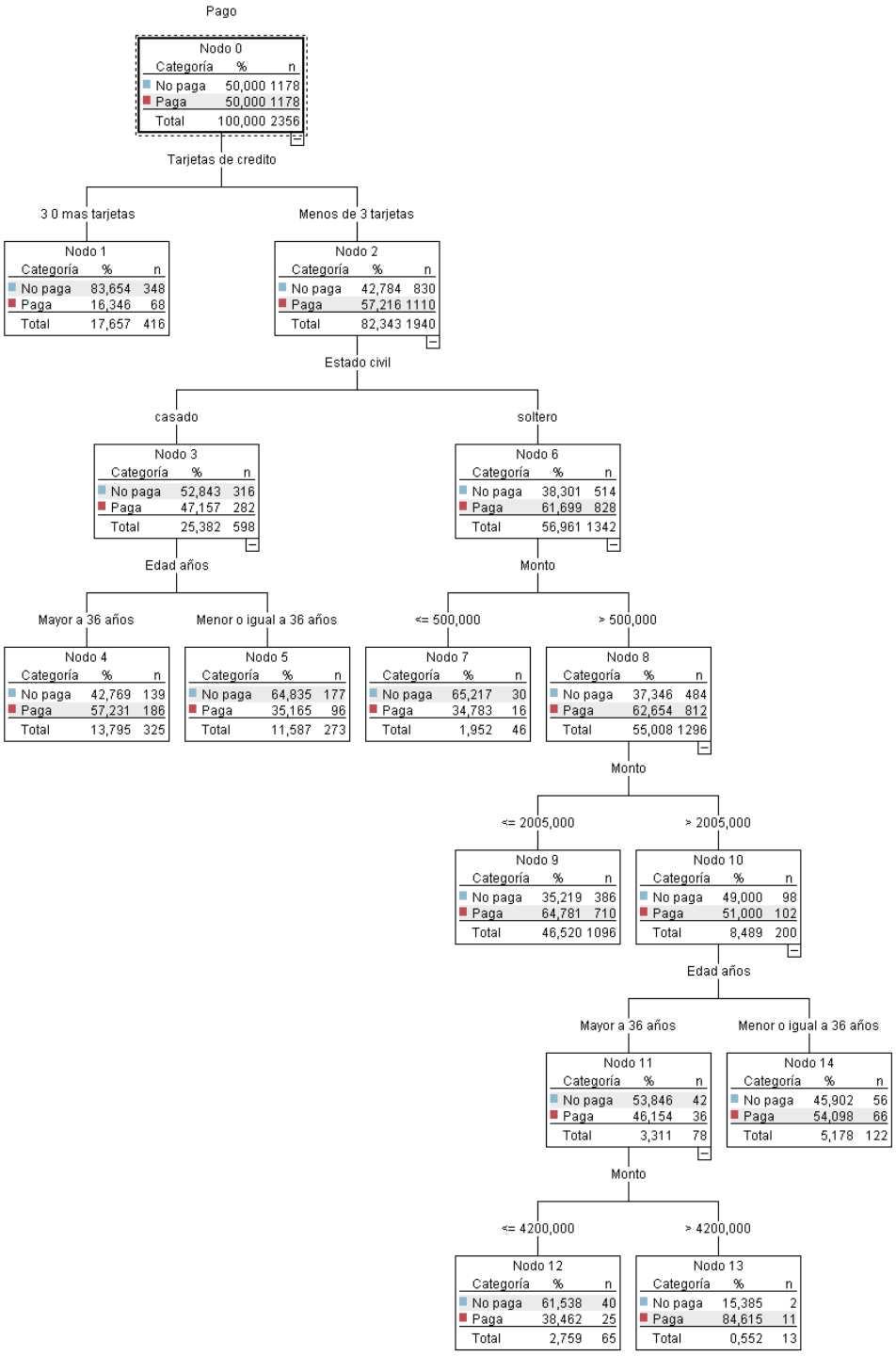


Figura 13: Árbol de clasificación: muestra completa

3.2.3.1. Validación árbol de clasificación: etapa de entrenamiento

En la etapa de entrenamiento, con el 70% de muestra seleccionada, del total de datos se encontró el poder predictivo del árbol de clasificación mostrado líneas arriba, en primer lugar, mostramos la matriz de clasificación o lo que es lo mismo la denominada matriz de confusión

Árbol de Clasificación: Etapa de Entrenamiento

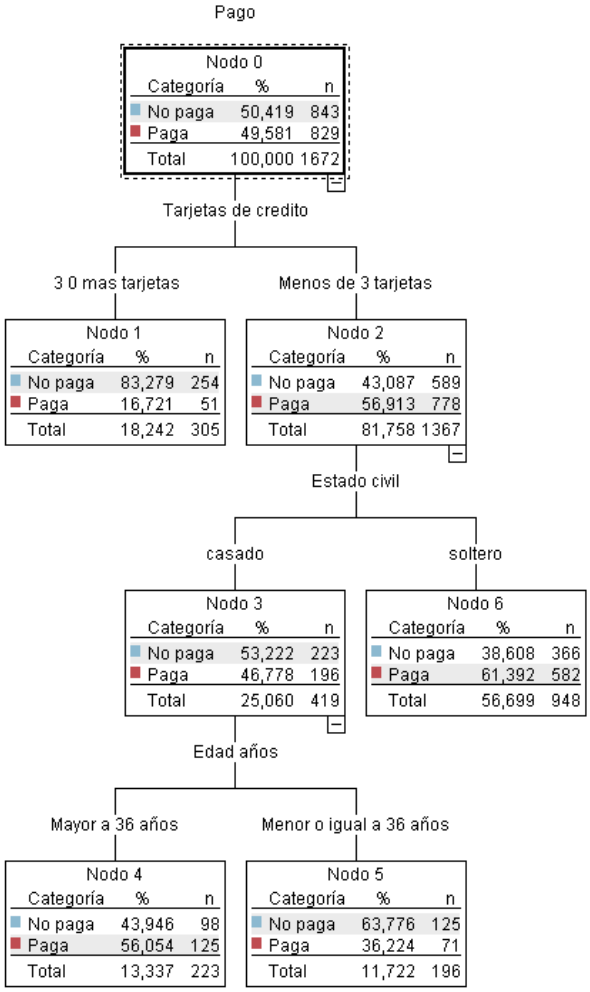


Figura 14: Árbol de Clasificación: Etapa de Entrenamiento

Tabla 18:

Tabla de clasificación árbol de clasificación: etapa de entrenamiento

'Partición'	1_Entrenamiento	
Correctos	**1.086**	**64,95%**
Erróneos	**586**	**35,05%**
Total	**1.672**	

Los resultados, son evidentes, el 65% de las predicciones son correctas para un total de 1672 observaciones que representan el 70% del total de los datos, reiteramos que la predicción es contra las observaciones utilizadas también para el modelado, por lo que es necesario validar el mismo para observaciones no consideradas en la obtención del modelo.

La denominada curva ROC, es presentada a continuación:

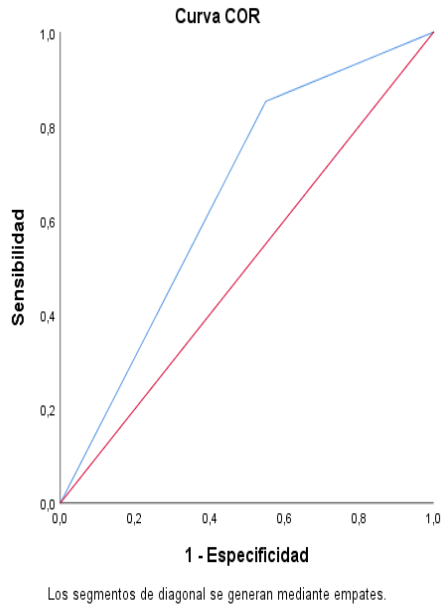


Figura 15: Curva ROC Árbol de clasificación: Etapa de entrenamiento

Mostramos a continuación la tabla que complementa al grafico anterior:

Tabla 19:

Área bajo la Curva ROC

Área bajo la curva

Variables de resultado de prueba: 1 Paga 0 No paga

Área	Desv. Error[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,651	,013	,000	,625	,678

Las variables de resultado de prueba: 1 Paga 0 No paga tienen, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

El área bajo la curva ROC es de 0.65 considerada desde un punto de vista representativo es aceptable, pero nos interesa la etapa de validación.

3.2.3.2. Árbol de clasificación: etapa de validación

Tenemos a continuación el árbol de clasificación en esta etapa de la validación:

Árbol de Clasificación: Etapa de Validación

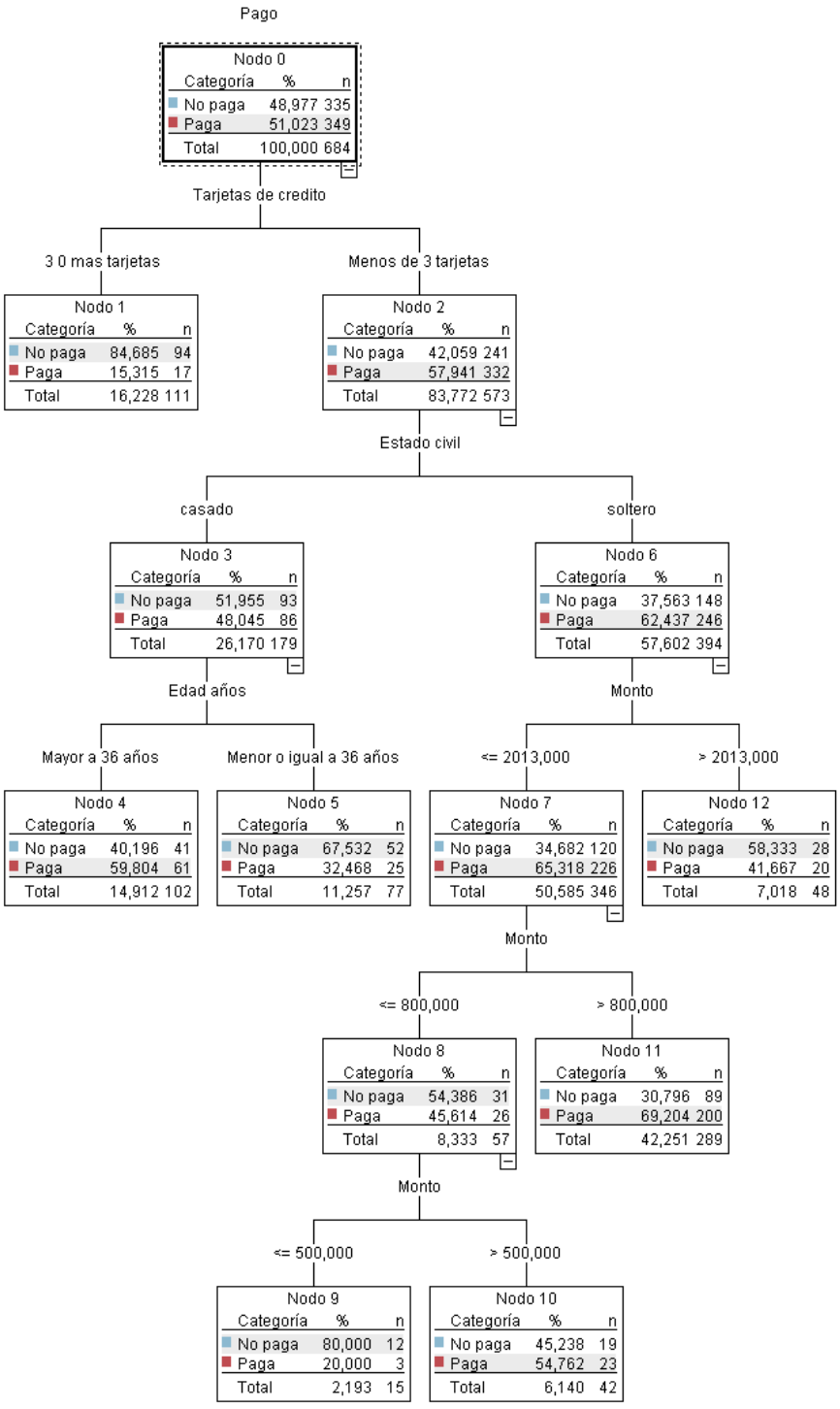


Figura 16: Árbol de Clasificación: Etapa de Validación

La tabla de clasificación, referida al árbol de clasificación en su etapa de validación es la que mostramos a continuación:

Tabla 20:

Tabla de clasificación árbol de clasificación: etapa de entrenamiento

'Partición'	2_Comprobación	
Correctos	470	68,71%
Erróneos	214	31,29%
Total	684	

En esta etapa de la validación, podemos observar que el 69% de los casos tienen una predicción correcta hecha por el algoritmo C5.0, lo que ubica a este adecuadamente para hacer las predicciones.

La curva ROC es presentada a continuación:

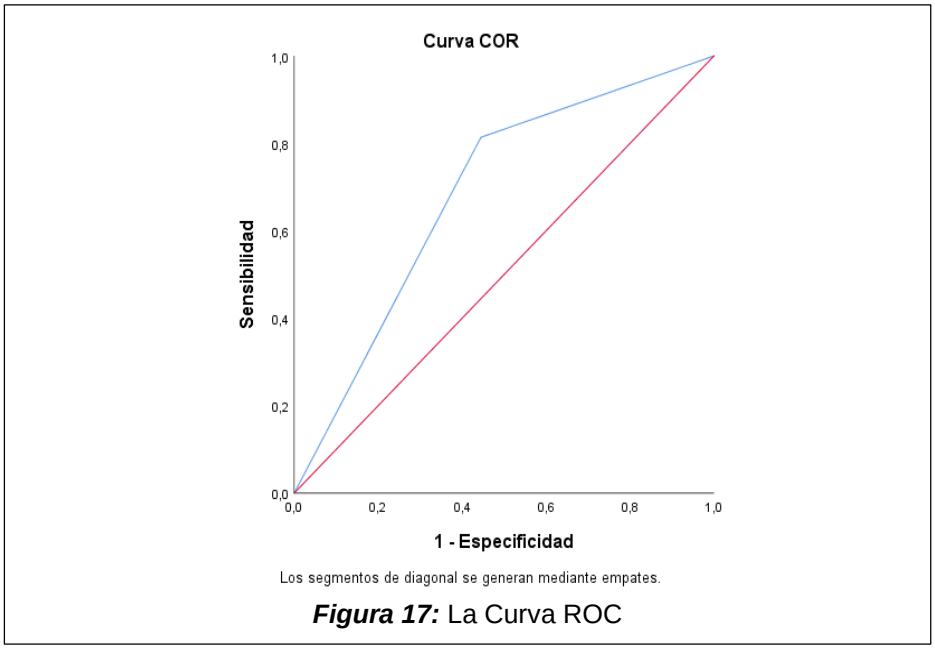


Figura 17: La Curva ROC

Tabla 21:

Área bajo la Curva

Área bajo la curva

Variables de resultado de prueba: 1 Paga 0 No paga

Área	Desv. Error[a]	Significación asintótica[b]	95% de intervalo de confianza asintótico	
			Límite inferior	Límite superior
,684	,021	,000	,644	,725

Las variables de resultado de prueba: 1 Paga 0 No paga tienen, como mínimo, un empate entre el grupo de estado real positivo y el grupo de estado real negativo. Las estadísticas podrían estar sesgadas.

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

El complemento de la curva ROC está mostrando con claridad que el área bajo la mencionada es de 0.68, bastante representativa en esta etapa de valuación o también denominada de convalidación.

3.2.4. Tabla Resumen

Tabla 22:

Predicciones Correctas y Área Bajo la Curva ROC

	ENTRENAMIENTO		VALIDACION	
	MATRIZ	CURVA ROC	MATRIZ	CURVA ROC
Regresión Logística	*66.3%*	*0.665*	*65.79%*	*0.655*
Árbol de clasificación	*64.9%*	*0.651*	*68.71%*	*0.684*

Fuente: Elaboración propia

CONCLUSIONES

- Al hacer el análisis de las dos principales técnicas de minería de datos, regresión logística y árboles de clasificación, se concluye que la aplicación de aquellas resulta adecuadas y eficientes para cuantificar la probabilidad de pago (P=1) así como su complemento (1-P), con lo cual se descubre conocimiento mediante los datos.

- El porcentaje de clasificación de personas naturales de una Institución Financiera de Chiclayo mediante el modelo supervisado de la Regresión Logística en la etapa de entrenamiento fue del 66.31% mientras que en la etapa de validación el porcentaje de clasificación fue del 65.79%. En lo referente al área bajo la curva ROC fue 0.665 y 0.655 para las etapas respectivas

- El porcentaje de clasificación de personas naturales de una Institución Financiera de Chiclayo mediante el modelo supervisado Árboles de Clasificación en la etapa de entrenamiento fue del 64.9% mientras que en la etapa de validación o prueba fue del 68.71%. siendo del mismo modo para la curva ROC 0.651 y 0.684 respectivamente para las etapas del caso.

- Después de realizar la evaluación se comparó y concluyo que el modelo supervisado de Arboles de clasificación proporciona un mayor grado de exactitud para identificar la probabilidad de pago de las personas naturales que solicitarían préstamos a una institución financiera en la ciudad de Chiclayo.

- Del conjunto de predictores se observa con claridad que los que tienen mayor influencia en la probabilidad de pago de los clientes de la institución financiera, esta en primer lugar el monto cuyo indicador de influencia es 58.22, seguida de el número de tarjetas cuyo indicador es 23.05 siendo los otros predictores de menor importancia, por lo que a la hora de decidir la aprobación de la solicitud de crédito tiene que tener en cuenta la evaluación de estas variables.

RECOMENDACIONES

- Se deben realizar o poner en práctica otras herramientas de minería de datos con la intención de explotar adecuadamente estas nuevas técnicas.

- Se debe implementar el modelo a fin de ayudar a tomar la decisión de otorgar o no el prestamo, de tal manera de servir como complemento a las políticas de préstamos que la institución financiera tiene como guía.

- Como se evidencia que ambos modelos se ubican en la categoría de aceptables podemos recomendaría hacer uso de técnicas de minería de datos a fin de obtener resultados adecuados y minimizar el riesgo inherente a una institución del sistema financiero.

- Puesto que estas técnicas de minería de datos, constituyen una nueva forma de análisis de los mismos, es conveniente el desarrollo adecuado de asignaturas que tiendan al conocimiento de los planteamientos fundamentalmente a los estudiantes de ciencias económicas y afines.

REFERENCIAS BIBLIOGRÁFICAS

Aguilar, A. G., & Camargo, C. G. (2004). *Análisis de la Morosidad de las Instituciones Microfinancieras (IMF) en el Perú.* (Documento de Trabajo 133), Instituto de Estudios Peruanos IEP, Lima. Obtenido de http://biblioteca.clacso.edu.ar/Peru/iep/20190802040639/aguilar.pdf

Aldrich, J., & Nelson, F. D. (1984). *Linear Probability. Logit and Probit Models* (Vol. 1° edición). USA: Sage publications.

Alvarado, M. (2002). *Evaluación y Manejo del Riesgo Crediticio en el Ámbito Agrícola.* Perú.

Baca, G. A. (1997). La Administración de Riesgos Financieros. *Ejecutivos de Finanzas*.

Bensic, M., Sarlija, N., & Zekic, S. (2005). Modelling Small-Business Credit Scoring by Using Logistic, Neural Networks and Decision Trees. *International Journal of Intelligent Systems in Accounting and Finance Management*, 133 - 150.

Bodie, Z., & Merton, R. C. (1999). *Finanzas.* Mexico: Pretince Hall.

Cabrera, C. A. (2014). *"Diseño de Credit Scoring para Evaluar el Riesgo Crediticio en una Entidad de Ahorro y Crédito Popular".* (Tesis de Maestría), Universidad Tecnológica de la Mixteca, Mexico. Obtenido de http://jupiter.utm.mx/~tesis_dig/12221.pdf

Cardona, H. P. (2004). Aplicación de Árboles de decisión en Modelos de Riesgo Crediticio. *Revista Colombiana de Estadística, 27*, 139 - 151. Obtenido de https://www.emis.de/journals/RCE/V27/V27_2_139Cardona.pdf

Escalona, C. A. (2011). *Uso de los Modelos Credit Scoring en Microfinanzas.* (Tesis de Maestría), Universidad Autónoma Chapingo, Mexico. Obtenido de http://www.lareferencia.info/vufind/Record/MX_ff2cebc4a3a295a69f9a9c3819538c97

Fragoso, C. (2002). *Análisis y Administración de Riesgos Financieros* (Vol. Capítulo 13 Mercado de Derivados). Xalpa .

Galicia, R. M. (2003). *Nuevos Enfoques de Riesgo de Crédito.* Mexico.

Hernandez, O. R. (2015). *Introducción a la MInería de Datos.*

Herrán, A. L. (2009). *"Evaluación Crediticia aplicando un Modelo de Credit Scoring en el Ámbito Microempresaria. Caso CMAC Paita".* (Tesis Licenciatura), Universidad de Piura, Facultad de Ciencias Económicas y Empresariales, Piura. Obtenido de https://pirhua.udep.edu.pe/bitstream/handle/11042/1325/ECO_030.pdf?sequence=1&isAllowed=y

Jorion, P. (1999). *Valor en riesgo.* Mexico: Limusa.

Katare, A., & Athavale, V. (2011). *Behavior Analysis of Different Decision Tree Algorithms* (Vol. 1). International Journal of Computer Technology and Electronics Engineerin (IJICTEE).

Krugman, R. P. (1995). *Economía Internacional* (3° Edición ed.). España: McGraw - Hill.

Ladino, B. I. (2014). *Comparación de Modelos de Riesgo de Crédito: Modelos Logísticos y Redes Neuronales.* (Tesis de Maestría), Universidad Javierana, Facultad de Ciencias Económicas y Administrativas, Bogotá. Obtenido de https://repository.javeriana.edu.co/bitstream/handle/10554/14857/LadinoBecerraIvanCamilo2014.pdf?sequence=1&isAllowed=y

Levi, D. M. (1997). *Finanzas Internacionales* (Vol. 3° Edición). McGraw -Hill.

Lewent, J. C., & Kearney, J. A. (1990). Identifying, Measuring and Hedging Currency Risk at Merck. *Journal of Applied Corporate Finance, 2*, 19 - 28.

Liao, T. F. (s.f.). *Interpreting Probability Models, Logit, Probit and other Generalizes linear models.* Sage Publications.

Long, J. S. (s.f.). *Regression models for categorical and limited dependet variables.* Sage Publications.

Mallo, F. F. (2011). *"Modelos Multivariantes Internos de Medición de Riesgo de Crédito, acordes con Basilea II".* (Tesis Doctoral), Universidad de Salamanca, Departamento de Estadística, Salamanca. Obtenido de

file:///C:/Users/User/Downloads/DES_Mallo_Fernandez_F_ModelosMultivariantes.pdf

Marzo, M. C., Wicijowski, C., & Rodriguez, Z. L. (2008). *"Prevención y cura de la Morosidad. Análisis y Evolución futura de la Morosidad en España"*. (Curso Master en Mercats Financers), HAS Agencia de Investigación Privada . Obtenido de https://www.bsm.upf.edu/documents/mmf/07_03_prevencion_morosidad.pdf

Moreno, V. S. (2013). *"El Modelo Logit Mixto para la construcción de de un Scoring de Crédito"*. (Tesis de Mestría), Universidad Nacional de Colombia, Facultad de Ciencias, Escuela de Estadística, Medellin. Obtenido de http://bdigital.unal.edu.co/39466/1/43596322.2014.pdf

Nieto, M. S. (2010). *"La Estadística Aplicada a un problema de Riesgo Crediticio"*. Universidad Autónoma Metropolitana, Mexico. Obtenido de http://mat.izt.uam.mx/mcmai/documentos/tesis/Gen.07-O/Nieto-S-Tesis.pdf

Patil, N., Lathi, R., & Chitre, V. (2012). Customer Card Classification Based on C5.0 & CART Algorithms. International Journal of Engineering Research and Applications (IJERA). Obtenido de https://pdfs.semanticscholar.org/87b9/df85c16d81c09d399f15f06aacfce8021df8.pdf?_ga=2.68957138.27754705.1572987145-735371681.1572987145

Peña, D. (2002). *Análisis de datos Multivariantes.* S.A. MCGRAW-HILL / INTERAMERICANA DE ESPAÑA.

Pérez, C. (2011). *Técnicas de Segmentación, Concepto, Herramientas y Aplicaciones* (Vol. 1° edición). (G. E. C.V., Ed.) Mexico, Mexico: Alfaomega.

Pérez, C. (2014). *Técnicas estadísticas con variables categóricas. IBM SPSS* (Vol. 1°Edición). Madrid: Ibergarceta Publicaciones, S.L.

Pérez, M. (2014). *Minería de datos a través de ejemplos.* México: Alfaomega Grupo Editor, S.A de C.V.

Quilan, J. R. (1986). *Induction of decision trees* (Vol. 1). Machine Learning.

Quilan, J. R. (1993). *C4.5: Programs for Machine Learning.* Douglas Sery.

Resendiz, T. J. (2006). *"Las maquinas de vectores de soporte para identificación en línea"*. (Tesis de Maestría), Centro de Investigación y Estudios Avanzados del Instituto Politécnico Nacional , Departamento de Control Automático, Mexico. Obtenido de https://www.ctrl.cinvestav.mx/~yuw/pdf/MaTesJAR.pdf

Rodríguez, R. J. (2010). *Fundamentos de Minería de datos* (Vol. 1° edición). Bogotá: BOGOTA D.C. Universidad Distrital Francisco José de Caldas.

Rojas, M. R. (2001). *"Aplicaciones Estadísticas en la Evaluación Financiera de Proyectos"*. (Tesis Licenciatura), Universidad Autónoma de Manizales, Colombia. Obtenido de http://bdigital.unal.edu.co/9236/1/ricardoalfredorojasmedina.2001.pdf

Rosillo, J. (2002). Modelo de Predicción de quiebras de las empresas colombianas. *Revista de Ciencias Administrativas y Sociales*. Obtenido de https://www.redalyc.org/pdf/818/81801908.pdf

Salinas, Á. J. (2005). *"Metodologías de Medición del Riesgo de Mercado en INstituciones de Fomento y Desarrollo Territorial"*. (Tesis de Maestría), Universidad Nacional de Colombia, Facultad de Ciencias y Administración, Manizales. Obtenido de http://bdigital.unal.edu.co/1156/1/johnjairosalinasavila.2005.pdf

Soltan, A. A., & Mohammadi, M. M. (2012). A hybrid model using decision tree and neural network for credit scoring problem. *Management Science Letters* . Obtenido de https://core.ac.uk/download/pdf/26797962.pdf

Valderrey, S. P. (2011). *Segmentación de Mercados* (Vol. 1° edición). Bogotá, Colombia: Ediciones de la U.

Véliz, C. C. (2016). *Análisis Multivariante: Métodos Estadísticos Multivariantes para la investigación* (Vol. 1° edición). Buenos Aires, Argentina: CENGAGE Learning.

Venkata, K., Kumar, S., & Kiruthika, P. (2015). An Overview of Classification Algorithm in Data mining. *International Journal of Advanced Research in Computer and Communication Engineering, 4*, 255 - 257.

Vigo, C. G. (2010). *"Método de Clasificación para Evaluar el Riesgo Crediticio"*. (Tesis Licenciatura), Universidad Nacional Mayor de San Marcos, Facultad de Ciencias

Matemáticas E.A.P. de Estadística. Obtenido de http://cybertesis.unmsm.edu.pe/bitstream/handle/cybertesis/3327/Vigo_cg.pdf?sequence=1&isAllowed=y

ANEXOS

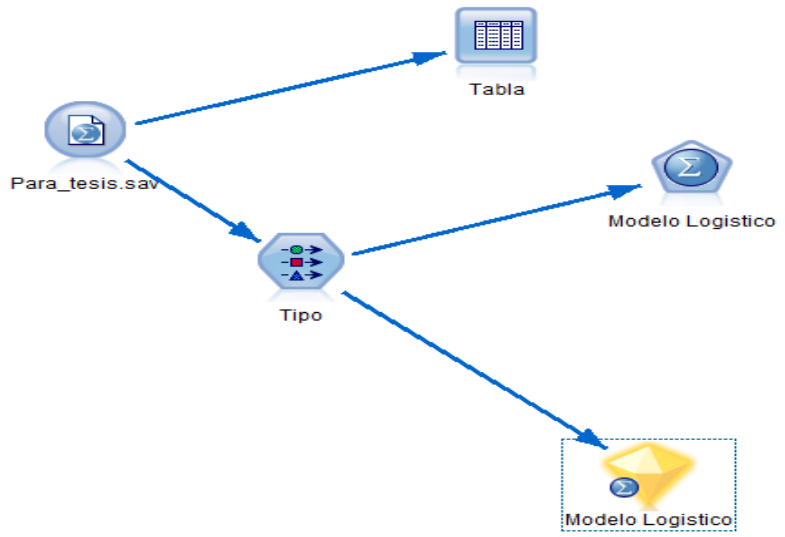


Figura 18: Ruta para la Regresión Logística (IBM SPS MODELER)

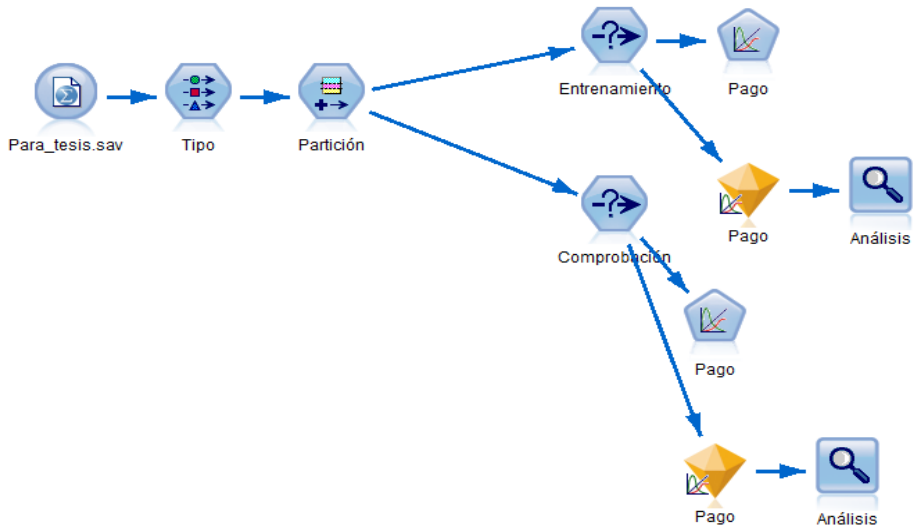


Figura 19: Ruta para la Regresión Logística (IBM SPS MODELER)

Resultados de la Regresión Logística (IBM SPSS MODELER)

Tabla 23:

Codificación de Variable Dependiente

Valor original	Valor interno
No paga	0
Paga	1

Fuente: Elaboración propia

Tabla 24:

Tabla de Clasificación

			Pronosticado		
			1 Paga 0 No paga		Porcentaje
	Observado		No paga	Paga	correcto
Paso 0	1 Paga 0 No paga	No paga	0	1178	,0
		Paga	0	1178	100,0
	Porcentaje global				50,0

Fuente: Elaboración propia

Tabla 25:

Variables en la Ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	,000	,041	,000	1	1,000	1,000

Tabla 26:

Resumen del Modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	2978,251[a]	,115	,153

a. La estimación ha terminado en el número de iteración 4 porque las estimaciones de parámetro han cambiado en menos de .001.

Fuente: Elaboración propia

Tabla 27:

Prueba de Hosmer y Lemeshow

Paso	Chi-cuadrado	gl	Sig.
1	32,941	8	,000

Fuente: Elaboración propia

Tabla 28:

Tabla de Contingencia para la prueba de Hosmer y Lemeshow

		1 Paga 0 No paga = No paga		1 Paga 0 No paga = Paga		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	197	198,334	33	31,666	230
	2	182	179,095	56	58,905	238
	3	150	129,603	86	106,397	236
	4	115	126,419	143	131,581	258
	5	112	120,192	170	161,808	282
	6	78	95,401	156	138,599	234
	7	79	98,590	164	144,410	243
	8	110	92,759	126	143,241	236
	9	83	67,717	112	127,283	195
	10	72	69,890	132	134,110	204

Fuente: Elaboración propia

Tabla 29:

Tabla de Clasificación

			Pronosticado		
			1 Paga	0 No paga	Porcentaje
	Observado		No paga	Paga	correcto
Paso 1	1 Paga 0 No paga	No paga	541	637	45,9
		Paga	186	992	84,2
	Porcentaje global				65,1

Fuente: Elaboración propia

Tabla 30:

Variables en la Ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Monto	-0,000203	,000	,309	1	,578	1,000
	Estado civil	-,587	,097	36,395	1	,000	,556
	Tarjetas de credito	-1,968	,143	190,339	1	,000	,140
	Ingresos	,026	,140	,033	1	,855	1,026
	Edad años	,263	,092	8,214	1	,004	1,300
	Constante	,402	,085	22,374	1	,000	1,495

Fuente: Elaboración propia

Tabla 31:

Variables en la Ecuación

		95% C.I. para EXP(B)	
		Inferior	Superior
Paso 1[a]	Monto	1,000	1,000
	Estado civil	,459	,673
	Tarjetas de crédito	,106	,185
	Ingresos	,780	1,349
	Edad años	1,087	1,556
	Constante		

a. Variables especificadas en el paso 1: Monto, Estado civil, Tarjetas de crédito, Ingresos, Edad años

.

Fuente: Elaboración propia

Resultados de la Regresión Logística STATA 15

Tabla 32:

Regresión Logística

```
. logit Pago Monto Estciv Tarjet Ingresot Edadt

Iteration 0:   log likelihood = -1633.0548
Iteration 1:   log likelihood = -1489.9433
Iteration 2:   log likelihood = -1489.1268
Iteration 3:   log likelihood = -1489.1256
Iteration 4:   log likelihood = -1489.1256

Logistic regression                             Number of obs     =      2,356
                                                LR chi2(5)        =     287.86
                                                Prob > chi2       =     0.0000
Log likelihood = -1489.1256                     Pseudo R2         =     0.0881
```

Pago	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Monto	-.0000203	.0000366	-0.56	0.578	-.0000921	.0000514
Estciv	-.5874092	.0973684	-6.03	0.000	-.7782478	-.3965706
Tarjet	-1.967942	.1426423	-13.80	0.000	-2.247515	-1.688368
Ingresot	.0255312	.1396465	0.18	0.855	-.2481708	.2992332
Edadt	.2626608	.0916459	2.87	0.004	.0830381	.4422835
_cons	.4020465	.0849965	4.73	0.000	.2354565	.5686365

```
. estat classification

Logistic model for Pago
```

	True		
Classified	D	~D	Total
+	992	637	1629
-	186	541	727
Total	1178	1178	2356

```
Classified + if predicted Pr(D) >= .5
True D defined as Pago != 0
```

Sensitivity	Pr(+| D)	84.21%
Specificity	Pr(-|~D)	45.93%
Positive predictive value	Pr(D| +)	60.90%
Negative predictive value	Pr(~D| -)	74.42%
False + rate for true ~D	Pr(+|~D)	54.07%
False - rate for true D	Pr(-| D)	15.79%
False + rate for classified +	Pr(~D| +)	39.10%
False - rate for classified -	Pr(D| -)	25.58%
Correctly classified		65.07%

Logistic model for Pago, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.1699	40	36.4	218	221.6	258
2	0.4437	49	55.9	165	158.1	214
3	0.4939	89	107.6	149	130.4	238
4	0.5142	140	128.6	112	123.4	252
5	0.5894	170	161.8	112	120.2	282
6	0.5932	113	100.6	57	69.4	170
7	0.5943	207	182.4	100	124.6	307
8	0.6004	99	104.5	76	70.5	175
9	0.6548	115	149.0	115	81.0	230
10	0.6638	156	151.2	74	78.8	230

number of observations = 2356
number of groups = 10
Hosmer-Lemeshow chi2(8) = 45.56
Prob > chi2 = 0.0000