UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA ECONÓMICA, ESTADÍSTICA Y CIENCIAS SOCIALES

"CAPACIDAD PREDICTIVA DE LOS MODELOS DE MÁQUINA DE VECTORES DE SOPORTE Y MODELO DE REGRESIÓN LOGÍSTICA EN EL ANÁLISIS DE RIESGO DE CRÉDITO - PERSONA"

TESIS

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO ESTADÍSTICO

POR LA MODALIDAD DE TESIS

ELABORADO POR:

ANGEL DE JESUS FRANCISCO REYES OBISPO
DANY ROYMER LEÓN DAVÁN

Lima - Perú

2014

A mis padres Samuel y Milagros por su gran cariño, a hermanos Alex, Samuel y Maria Isabel por su alegría; a una persona especial Gelyn. Además a un gran amigo y guía profesional Dany León por su apoyo incondicional. En memoria de mi madrina Dina por su eterno amor.

Angel

A Leandro y Robertina, mis padres.

Dany

AGRADECIMIENTOS

A nuestra alma mater y a todos aquellos que hicieron posible cumplir este gran reto.

Tabla de contenido

RESUMEN EJECUTIVO

El presente estudio, desarrollado en el ámbito de la aplicación de modelos predictivos para la evaluación de riesgo crediticio en banca personal es llevado a cabo sobre una base de datos histórica del profesor Hoffman. Siendo importante mencionar que, la investigación es realizada debido que recientes estudios han revelado que las emergentes técnicas de inteligencia artificial son más ventajosas a los modelos estadísticos en cuanto a poder de pronóstico por su alta capacidad de discernimiento de patrones. En el estudio aplicamos y comparamos los resultados encontrados de la técnica de clasificación estadística como es la Regresión Logística con la técnica computacional desarrollada de Support Vector Machine (SVM), esta última es basada en algoritmos matemáticos de aprendizaje. Los resultados experimentales serán llevados a cabo para la problemática de detección de incumplimiento de pago en riesgo de crédito. Siendo los principales hallazgos que: El modelo de SVM presenta mejores indicadores de capacidad predictiva en sus 4 indicadores de potencialidades de capacidad preditiva, con respecto a la aplicación de la Regresión Logística. Y adicionalmente encontramos que ambos modelos de propensión de riesgo de crédito identifican riesgos relativos similares entre las variables elegidas para el modelamiento del riesgo de incumplimiento de pago.

I. Planteamiento del problema

I.1 Antecedentes

Una mayor competencia en el mercado y búsqueda de incremento de rentabilidad ha conllevado a las empresas financieras (Banca Minorista, Personal, etc.) a investigar maneras efectivas para conseguir nuevos clientes que se pueda ofrecer crédito y al mismo tiempo controlar las pérdidas del incumplimiento de sus pagos (Default). Los esfuerzos del marketing agresivo han generado como resultado una profunda inserción en los grupos de riesgo de clientes potenciales, y la necesidad de procesar rápida y efectivamente ha conllevado a una creciente automatización de las postulaciones a crédito y seguro. El gerente de riesgo de crédito es ahora retado a producir soluciones en la asignación del riesgo, que no sólo evaluará la solvencia, además también debe mantener el bajo costo de procesamiento por unidad. Además la calidad del servicio al consumidor demanda que este proceso automatizado sea adecuada para minimizar la negación de créditos a clientes que sean dignos de crédito (buenos clientes).

En el pasado las instituciones financieras adquirían el puntaje de riesgo de crédito de mano de proveedores de riesgo de crédito, esto involucraba que las instituciones financieras entregasen su data a los proveedores, luego los proveedores desarrollaban un puntaje. Mientras que algunas compañías avanzadas han tenido funciones de modelamiento internos y desarrollo de puntajes por largo tiempo, la tendencia a desarrollar puntajes en la misma compañía dentro de la propia empresa se ha vuelto más popular en los últimos años.

En Perú existe como precedente el estudio presentado por la Universidad ESAN Un *modelo de CreditScoring para instituciones micro financieras basados en la normativa Basilea II (Junio 2010)* Salvador Canton, Juan Rubio y David Blascopor los

profesoresde las universidades de Granada y Carlos III de España. En este documento tienen como objetivo hacer una presentación metodológica de un modelo predictivo de riesgo de crédito banca persona para analizar el proceso de calificación de riesgo mediante modelos internos, específicamente aplicando el modelo de regresión logística. Finalmente el estudio concluye de la siguiente manera:

"La estimación del modelo de creditscoring se realizó mediante el método de introducción por pasos, y aplicando la técnica paramétrica de regresión logística de las variables explicativas sobre la base de las fases y estudios obtenidos en el proceso de concesión de un microcrédito. De esta forma, la investigación realizada diseña un modelo de calificación estadística capaz depredecir correctamente en 78.3% de los créditos de la cartera de la EdpymeProempresa, corroborado por un porcentaje similar en el proceso de validación del modelo. A este respecto, las medidas de valoración del modelo globalmente indican un ajuste aceptable en regresión logística."(pág.27 Canton-Rubio-Blasco.)

Pero existen estudios internacionales tal como: *Creditscoringwith a data miningapproachbasedonsupport vector machines* que promuevan la aplicación de técnicas de optimización como es la Máquina de Vectores de Soporte como modelo predictivo para la evaluación de riesgo crediticio.Estudiocomo el planteado en el 2007por Chen-Lung Huang, Mua-Chen Chen y Chieh-Jen Wand delasuniversidades de National Kaosiungng First University of Science and Technology, Institute of Traffic and Transportation, National Chiao Tung University y Department of Information Management, Huafan Universityrespectivamente.El documento se da inicio con la propuesta comparativa de técnicas de minería de datos para el contexto de la evaluación crediticia en riesgo de banca persona:

"La aplicación de software ha permitido a los usuarios a desarrollar puntajes sin recurrir en infraestructura ni avanzados programadores. Complejas funciones de minería de datos están disponibles para su uso de forma sencilla, permitiendo al analista manejar sus propios modelos valoración de riesgo de crédito interna".

El estudio tiene como objetivo sustentar la afirmación que el modelo de Máquina de Vectores de Soporte que es actualmente investigado, presenta mejores resultados en cuanto a capacidad predictiva que la aplicación de otras técnicas de modelamiento del riesgo crediticio de banca persona. Y como resultados del estudio detalla lo siguiente:

"El modelo de riesgo de crédito puede clasificar a los aspirante a la obtención de un crédito financiero de manera adecuada minimizando el riesgo y detectando clientes con buen comportamiento de pago.

Es evidente que el modelo de SVM es muy competitivo con igual o mejor capacidad predictiva que las técnicas de Algoritmo Genético o el Backpropagation de Redes Neuronales en cuanto a la evaluación de crediticia de banca personal".(pág. 8Huang ,Chen y Wand)

La propuesta de aplicar la técnica de Máquina de Vectores de Soporte como modelo predictivo para mejorar los indicadores de capacidad predictiva es reforzado también por el estudio *CreditRiskEvaluationwithLeastSquareSupport Vector Machine (2006) de los autores KinKeungLai, Lean Yu, LigangZhou y Shouyang Wang de las universidades de Hunan, Universidad de Hong Kong, del instituto de ciencias de sistemas de la academia de ciencias chinas respectivamente.*

Los autores dan inicio a su estudio mostrado describiendo la diversidad de técnicas de minería de datos y teniendo como objetivo la comparación con la técnica de Máquina de Vectores de Soporte:

"Las modernas técnicas de minería de datos, han tenido contribuciones significativas para el campo de la ciencia de la información, las cuales pueden ser adoptados para construir un modelo scoring de crédito. Analistas en la práctica e investigadores han desarrollado una gran variedad de modelos estadísticos tradicionales como, modelos discriminantes lineales, modelos logísticos, modelos de k vecinos cercanos, modelos de árboles de decisiones. Pero los resultados computacionales de las redes neuronales son más precisos en una predicción de falla, que los modelos antes mencionados. Esto debido a que las redes puedes ser más robustas y precisas. Las técnicas de minería de datos más recientes tales como las redes neuronales, programación genética y las Máquinas de Vectores de Soporte (SVM) (Vapnik 1995) pueden mejorar la tarea de clasificación sin limitaciones."

Después de desarrollo del modelamiento de riesgo crediticio financiero de una entidad financiera la conclusión de los autores fue la siguiente:

"En este estudio, se propone el método de estimación de la función de una poderosa clasificación los Mínimos cuadrados de la Máquina de Vectores de Soporte. A través de data experimental se obtuvo buenos resultados en la capacidad predictiva en comparación con el performance alcanzado por la validación de otros modelos predictivos."(pág. 5 Lai, Yu, Zhou y Wand).

Sin lugar a dudas la evaluación de riesgo de crédito es un importante campo en la gerencia de riesgo financiero. Especialmente para instituciones aseguradoras de crédito, tales como bancos comerciales y minoristas, la capacidad de discriminar los buenos clientes de los malos clientes es crucial. Por ende la necesidad de modelos confiables que predigan precisamente es imperativa. Por ello es que se están usando tanto técnicas estadísticas clásicas como técnicas de Inteligencia artificial con esta finalidad.

I.2 Definición del problema

Encontramos que en la base de datos histórica provista por el profesor Hoffman para la evaluación crediticia de clientes, al realizar la aplicación del modelo de estadístico paramétrico como es la Regresión Logística en sus 4 indicadores de capacidad predictiva, sería posible poder obtener mejores indicadores de capacidad predictiva al aplicar la técnica no paramétrica de Maquina de Vectores de Soporte, al ser la última más robusta en escenarios más complejos.

La importancia de contar con modelos de evaluación de riesgo de crédito confiables que nos proporcionen predicciones de no pago (default) con una alta precisión yde forma oportuna ha llevado al planteamiento de muchos modelos, incluyendo técnicas tradicionales, tales como análisis de discriminación y análisis logit, y emergentes técnicas de inteligencia artificial, tales como redes neuronales artificiales y Máquinas de Vectores de Soporte (SVM) las cuales fueron ampliamente aplicadas a la tarea de las puntuaciones de crédito obteniéndose algunos resultados interesantes.

Así mismo, el otorgamiento de un crédito a un cliente que no debía otorgarse debe tener una probabilidad mínima que estadísticamente nos representa el Error tipo II y lo mismo al no otorgar un crédito a un cliente que posee una buena capacidad de pago de sus cuotas denominado Error Tipo I.

A pesar de que existenmuchas técnicas de clasificación que pueden ser usadas para la evaluación del crédito, el rendimiento y robustez de la mayoría de estosmétodos necesita mejoraraúnmás. Por lo tanto existen aún algunos inconvenientes en las aproximaciones existentes. Por ejemplo, el modelo de evaluación de crédito basado en técnicas estadísticas usualmente requiere fuertes supuestos acerca de la data, tales como la distribución normal y continuidad de la data. Más aun, ellos generalmente no pueden

tratar eficientemente con relaciones no lineales implícitas entre las características y resultados. En las técnicas de inteligencia artificial, el modelo de redes neuronales sufre frecuentemente problemas de mínimo local y sobreajuste, mientras el modelo de Máquina de Vectores de SoporteSVM, propuesto primeramente por Vapnik, tienen una complejidad computacional larga cuando resuelve problemas de programación de escala cuadrática larga.

I.3 Preguntas de Investigación

- ¿Cuál de los dos modelos predictivos: Máquina de Vectores de Soporteo el modelo de Regresión Logísticamuestra mejores indicadores de capacidad predictiva de clasificación, en el análisis de riesgo de crédito – persona en una base de datos de clientes en una entidad Financiera?
- ¿Qué diferencias metodológicas nos conlleva a definir el Error de pronóstico al emplear los modelos de Regresión Logística y Máquina de Vectores de Soporteen el análisis de riesgo de crédito – persona en una base de datos de clientes en una entidad Financiera?

I.4 Objetivos

I.4.1 Objetivo General

Realizar una comparación descriptiva del performance alcanzado por las técnicas señaladas a continuación: la técnica Máquinas de Vectores de Soporte SVM (por sus siglas en ingles), y la técnica de Regresión Logística,evaluando su capacidad predictiva en el contexto específicode la evaluación de crédito.

I.4.2Objetivos Específicos

- Analizar los indicadores de capacidad predictiva que el modelo de Máquina de Vectores de Soporte (SVM) yel modelo de regresión logística presentan en cuanto

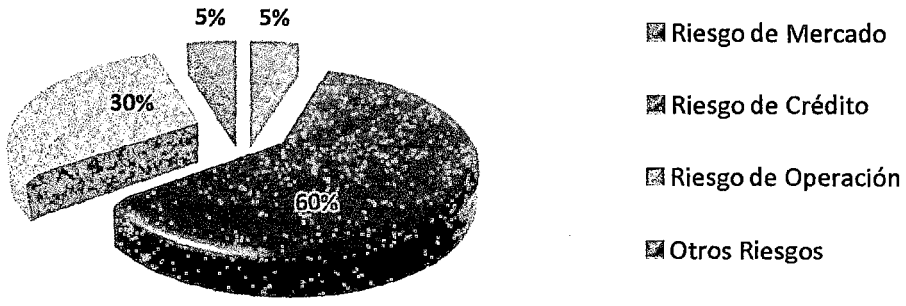
a poder de predicciónen el análisis de riesgo de crédito – persona en una base de datos de clientes en una entidad Financiera.

- Describir el factor o los factores que conllevaron a uno de los dos modelos presentados (una vez identificado el modelo predictivo con mejores indicadores de precisión de pronóstico) a tener un mejor performance en capacidad predictiva en el análisis de riesgo de crédito en el contexto de análisis mencionado.

I.5Justificación

Una vez definido nuestro problema, es muy sencillo justificar proceder con este estudio, en primer lugar es muy importante comprender cuán fundamental es medir el riesgo dentro de una compañía financiera (que brindan crédito) ya sea el Riesgo de crédito, Riesgo de Mercado, Riesgo Operativo o Riesgo de Liquidez, esto sumado a que los cambios en el entorno mundial han vuelto más complejo el seguimiento de operaciones y riesgos en el sistema bancario, dando lugar a retos para la dirección y gerencia de un banco. Estos retos aparecen a la par con las nuevas tendencias actuales como son: Globalización, Desregularización, Consolidación de Instituciones, Nuevos Instrumentos, Nuevas Tecnologías, Comunicación Inmediata, Contabilidad Creativa, etc. Todas estas tendencias son las manipuladoras del contexto actual y las que ocasionan que los riesgos (riesgo de crédito, riesgo de imagen, riesgo político, etc.) que las compañías financieras asumen se incrementen más, es debido a esto que el estudio y manipulación del Riesgo ha venido evolucionando en las últimas décadas, con la experiencia de muchas empresas que cayeron en crisis o simplemente quebraron teniendo millones de pérdidas a causa de su mala o nula manipulación. Dentro de este mar de riesgos que asumen las instituciones podemos fácilmente graficar aquellos riesgos que producen mayores pérdidas a la compañía, observando claramente cuan perjudicial pueden ser las perdidas en relación al Riesgo de Crédito.

Tipos de Riesgos



Fuente: SEBTON *MARKET RISK MODELS Mayo 2003*

A su vez, el análisis se centra en información de créditos en el portfolio provisto por el profesor Dr. Hofmann*. La metodología desarrollada será extendida para análisis crediticios en nuestro país.

En conclusión, debido al contexto mundial las compañías cada vez estánmás en la obligación de mejorar las prácticas de cuantificar los riesgos, presionados por las nuevas tendencias tales como son la aparición de nuevas herramienta, instrumentos y software donde la única ventaja final que marque la diferencia viene a ser el nivel de predicción que tenga cada institución a la hora de medir el Riesgo.

La complejidad intrínseca del Riesgo de crédito, Riesgo de Mercado, Riesgo Operativo o Riesgo de Liquidez en el mercado se torna más complejo aún con los cambios en el entorno mundial. En tal sentido, la incorporación de más variables para predecir situaciones de clasificación del riesgo ha vuelto más complejo el seguimiento de operaciones y riesgos en el sistema bancario los cuales generan superficies de frontera no modelables por parámetros secundarios (o híper-planos).

---

*Fuente de información: Profesor Dr. Hans Hofmann "InstitutfurStatistik un OkonometrieUniversit "en Hamburgo FB Wirtschaftswissenschaften Von-Melle-Park 5 2000 Hamburg 13

I.6 Hipótesis de trabajo

I.6.1 Hipótesis General

La aplicación de las técnicasMáquina de Vectores de Soporte define una mejor regla de discriminación que la aplicación de la Regresión Logística sobre las variables: Duración de Crédito, Propósito del crédito, Tasas de ingreso disponible del cliente, historia crediticia previa, Balance de cuentea de ahorro y estado civil de las variables en el contexto de la evaluación del riesgo crédito en nuestra data de estudio.

I.6.2Hipótesis de trabajo específicas

- La aplicación de Máquina de Vectores de Soporte presenta mejores indicadores de pronostico (Curva ROC y matriz de confusión) que la Regresión Logística sobre los datos de estudio.
- La asignación de importancia (pesos) para las variables, en ambos modelos son similares, siendo la mejora de predictiva justificada por el procedimiento Kernel.

I.7 Delimitación del problema

La data obtenida para realizar el presente estudio ha sido realizada en el contexto de evaluación crediticia de un banco obtenido de la página Web, de acuerdo a las características históricas de los clientes que han realizado el pago de los créditos en las cuotas acordadas y con los clientes que caen en un incumplimiento de los pagos. Este estudio cuantitativo pretende determinar una regla de clasificación entre los segmento de población de cumplimiento e incumplimiento de pago con indicadores de poder predictivo muy precisos.

En este escenario se pretendió emplear luego de una preparación de los datos, la aplicación de los algoritmos de Regresión Logística y Máquinas de Vectores de

Soportecon el objetivo de determinar la mejor regla de clasificación además de una descripción de los distintos procedimientos de ambas técnicas.

II.Marco Teórico

II.1 Teoría del Riesgo

La incertidumbre o riesgo es una parte constante en cualquier empresa de negocios. Los riesgos pueden provenir de diversas fuentes que requieren diferentes datos y modelos para poder evaluarlos, en lo posible medir el nivel de riesgo y gestionarlo.

Además en los negocios se ha definido 4 tipos de riesgos: negocio, crédito, mercado y operacional. En donde el Riesgo de Crédito, es la incertidumbre asociada comportamiento de pago de la contraparte deudora de un contrato crediticio. A su vez no solo está asociado a los incumplimientos de pago, sino también cambios en los grados de riesgo que influencian el valor en el mercado de las negociaciones de deuda, y la posibilidad de incurrir en costos extra para recuperar el dinero.[1]

Entonces con la finalidad de medir el riego crediticio se enmarcan los estudios sobre los datos históricos y el desarrollo de los modelos predictivo. Los modelos predictivos son la mejor aproximación del pronóstico de comportamiento de pago crediticio empleando reciente información histórica.

"Nada es constante, es cambiante, lo más probable es que el futuro aproxime más al pasado reciente que al pasado distante"

Mark Scheir

[1][1]Anderson (pág. 157).

Los tipos de modelos predictivos han sufrido una revaluación industrial y tecnológica en los últimos 200 años. Las etapas de evaluación del riesgo crediticio han pasado por las siguientes fases de evolución:

Juicio puro: Poco estructurado, empleado poca cantidad de data histórica. Se basaba en las evaluaciones subjetivas sin modelo o plantilla de evaluación de solicitudes de préstamo.

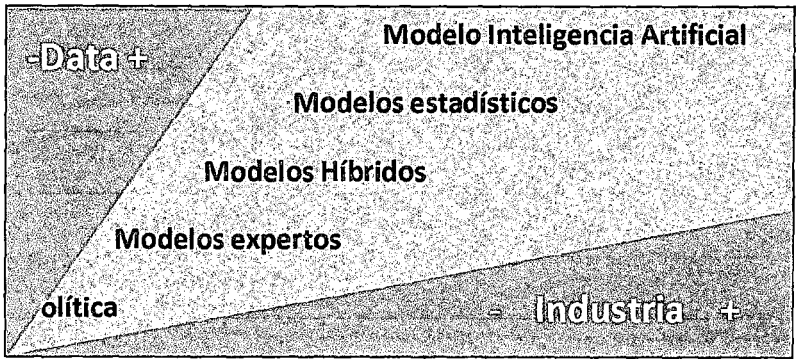
Política: Reglas usadas para limitar la decisión de préstamo. Es usualmente en la experiencia pasada, especialmente donde las perdidas por incumplimiento sean asociados por condiciones fuera de las normales.

Modelos Expertos: Existe pequeños repositorios de datos, y analistas que tienen experiencia suficiente para construir una política o un proceso productivo.

Modelos Híbridos: Disponía de variedad de datos disponible. Una combinación de tipos de modelos es usado, dependiendo las construcciones analíticas que se podrían hacer para evaluar el riesgo. Finalmente el resultado es la integración de diferentes modelos predictivos en un solo modelo.

Modelos Estadísticos: Altamente estructurado, sobre un alto nivel de datos. Mientras que las predicciones son altamente confiables, tienen la desventaja de una dependencia de data, y solamente el mejor y más altamente data estructurada.

Modelos Inteligencia Artificial: Son modelos que requieren data altamente estructurada, se emplea el modelamiento con técnicas matemáticas se le da prioridad el mejor pronóstico y empleado múltiples relaciones complejas de las variables predictivas.



Cuando nos centramos en las técnicas empleadas para los modelos predictivos en la evaluación del riesgo crediticio, (Thomas 2002) nos describe dos grupos de métodos usualmente empleados para la construcción de Scorecards crediticio:

II.1.1 Métodos Estadísticos

Cuando los modelos predictivos de riesgo de crédito fueron desarrollados inicialmente en 1950's y 1960's, los únicos métodos empleados fueron discriminación estadística y métodos de clasificación. Podemos diferenciar este primer grupo de métodos, por emplear distribuciones de probabilidad en el desarrollo.

Inicialmente los métodos eran basados en los métodos discriminantes de Fisher (1936) para un problema general de clasificación. Luego, la aproximación de Fisher como **Análisis Discriminante**, pasó a ser vista como una regresión que no requiera supuestos tan estrictos. El caso más exitoso en ese enfoque fue la **Regresión Logística**, que es el método comúnmente más aplicado.Otro método que se ha desarrollado en los últimos 20 años son los **Arboles de Clasificación o Particionamiento Recursivo**, cuyo procedimiento es una división del total de la muestra según los cortes de las variables predictoras, para poder identificar grupo más homogéneos en nivel de riesgo crediticio. A pesar de que los arboles de clasificación no tienen como resultado final de las

variables una ponderación, la finalidad es la misma identificar aquellos grupos recomendables o no recomendables de otorgar préstamos[2].

II.1.2 Métodos No Estadísticos

Si bien la idea original del desarrollo de los modelos predictivos de riesgo de crédito usando el análisis estadístico de una muestra histórica de clientes que soporte la decisión de las características de futuros clientes admitidos.

El punto de vista no estadístico se enfoca sobre la misma problemática. En los 80's se aplicó por primera vez un enfoque no estadístico al aplicarse **Programación Lineal**(Freed y Glover en 1981), que es una aplicación de procedimiento iterativo que garantiza los resultados con una tasa de error de mala clasificación. Por otro lado en los 70's hubo un enorme investigación de la Inteligencia Artificial cuya función principal era la generación de reglas a partir de grandes volúmenes de información. En los años 80's se desarrolló un método de la Inteligencia Artificial basado en el problema de clasificación las **Redes Neuronales**, que son modelos de proceso de decisión que aprende de un conjunto de casos históricos creando una red de posibles escenarios y una respuesta potencial para cada uno por tal motivo, cuando tiene que generar un pronóstico acerca de un caso un escenario de la Red Neuronal se activa y genera el pronóstico.

Dado que podríamos enfocarnos que el desarrollo de un modelo predictivo de riesgo de crédito es un tipo problema de optimización combinatoria. Esto quiere decir que teniendo un número de parámetros ponderadas las variables según data historita de los clientes designa un score de riesgos crediticio, se desarrolló una serie de algoritmos que se aproximan a la solución de este problema, los llamados **Sistemas Expertos** tales

[2][21]Lyn C. Thomas (pág. 41).

como **Support Vector Machine y el Algoritmo Genético**. Estos tipos de algoritmos se caracterizan principalmente por alcanzar alta precisión de pronósticos[3].

II.1.3 Riesgo Crediticio en Perú

En nuestro país tenemos autores como Edgardo Venero[4], que detallan la aplicabilidad de los métodos Scoring Flat como herramientas simplificada de evaluación crediticia para los sectores comerciales e industriales, también cajas de ahorro y empresas crediticias y considera que es un método por excelencia como herramienta de evaluación de riesgo, mas no detalla acerca de las metodología de la minería de datos en los modelos predictivos de riesgo de crédito ni la aplicación de algoritmos de Inteligencia artificial.

De la documentación revisada encontramos un estudio que desarrollaba parte técnica Metodológica del desarrollo de un Modelo Crédito Scoring para la evaluación de Riegos en Micro financieras peruanas para la Entidad de Desarrollo de la Mediana y Pequeña Empresa (Ed pyme) desarrollado por profesores de las Universidades de Granada - España y Carlos III de Madrid[5] enmarcado en la normativa internacional de Riesgo Financiero Basilea II. La publicación desarrollada por los profesores describe las diferentes técnicas de modelos predictivos para la evaluación de riesgo crediticio, pero centrándose en la aplicaciónde la Técnica Regresión Logística, donde emplea indicadores de la evaluación de sus pronósticos como son la Curva ROC y La Matriz de confusión teniendo como variables finales en su modelo de riesgo las siguientes: zona de residencia, la situación laboral, Ratio de liquidez, número de créditos concedidos con anterioridad, propósito del crédito, garantías y la variación anual de la tasa de cambio.

---

[3] [21]Lyn C. Thomas (pág. 64).
[4] [22] Edgardo Venero (pág. 32).
[5] [P-1] Rayo – Lara - Camino

II.1Técnicas a utilizar

II.1.1 Detección de valores atípicos

Existe el problema de identificar valores anómalos desde hace mucho tiempo como señala Bernoulli (1777). El tratamiento estadístico de los Outliers proviene de problemas de distorsiones de las asociaciones entre variables o casos atípicos encontrados en la recopilación de información a ser analizada[6].

Desde que los procedimientos de minería de datos se basan en patrones (medias de tendencia central, indicadores de asociación) los valores atípicos fácilmente pueden distorsionar el modelo o indicador representativo del conjunto de datos recopilado.

Algunas implicancias de los Outliers:

- El promedio aritmética se ve fuertemente influenciado por valores extremos.
- Las correlaciones, coeficientes de modelos también sufren de sesgo por estos valores anómalos.

Es por ello que una fase crucial en un estudio cuantitativo es la detección de los valores atípicos.

II.1.1.1 Detección de Outliers univariados para las variables cuantitativas no normales

Utilizamos como medida de tendencia central la mediana y como medida de dispersión el rango intercuartílico (la diferencia entre el Q1 y el Q3) por ser indicadores más robustos ante la presencia de valores atípicos. Y guiados por el criterio que la información valiosa de la variable estará contenida alrededor de la mediana en un alcance de 3 rangos intercuartílicos hacia la derecha y 3 rangos intercuartílicos a la

[6] [2] SanjoyKumarSinha (pág. 6).

izquierda, aquellos valores fuera del rango señalado serán etiquetados como potenciales outliers.

II.1.1.2. Detección de Outliers multivariados mediante K-means

Al aplicar el algoritmo de segmentación denominado k-means, tiene la desventaja de ser influencia por los valores atípicos en un procedimiento de netamente de segmentación. Por otro lado en un procedimiento de detección de valores atípicos esto es muy ventajoso[7]. Si empleamos el procedimiento de la siguiente manera, como sugiere Mandouft[8]:

Si se generase 50 cluster, entonces aquellos segmentos con baja frecuencia de concentración de casos y muy separados en distancias entre los centros de los segmentos de los demás cluster, será un potencial grupo de Outliers multivariado.

En la exploración de los datos es común encontrar casos que evaluados como datos examinados de manera univariados no muestren signos de ser atípicos, pero con una perspectiva multivariada como nos lo permite la técnica K-meansson valores atípicos que distorsionen los patrones hallados.

II.1.2Conversión de variables categóricas a variables Dummy

Si bien en la mayoría de estudios se emplea variables explicativas numéricas en la aplicación de modelos predictivos, la intervención de variables cualitativas u ordinales directamente es incorrecta ya sean estas nominales u ordinales.

[7] [10] Montgomery (37 pág.)
[8] [19] Mamdouh (94 pág.)

Matriz de diseño DUMMY

La representación de la matriz de diseño no es única como son los números 1's y 0's; eligiendo un nivel de referencia, sino también existen otros diseño que estarán sujetos a un diferente interpretación y que se aplicará dependiendo de la intención de del estudio y de los efectos como señala Arnold[9].

Efecto Codificado		
Respuestas	**Matriz de diseño**	
	I1	**I2**
Respuesta 1	0	0
Respuesta 2	1	0
Respuesta 3	1	1

Usando este diseño cada coeficiente de la variable dummies, se interpretaría como una medida del cambio de riesgo al pasar de una categoría a la siguiente.

Efecto Codificado		
Respuestas	**Matriz de diseño**	
	I1	**I2**
Respuesta 1	-1	-1
Respuesta 2	1	0
Respuesta 3	0	1

En caso de que una categoría no puede ser usada naturalmente como un nivel de referencia, se emplea un -1. Donde cada coeficiente de la matriz indicadora tiene una interpretación directa como cambio en el riesgo con respecto a la media de las tres respuestas.

[9][20] Tim Arnold pág 2343

II.1.3 Indicador de Poder Predictivo WOE (Weight of Evidence – Peso de la evidencia) e IV (InformationValue – Valor de la información).

La reducción de variables predictivas o selección de las más relevantes es una buena prácticas según Refaat[10]. Por principio de parsimonia se ha establecido que la mejor solución es la más simple. Tratándose de modelos predictivos no hay lugar a duda que el mejor modelo será el que alcance un buen pronóstico empleando la menor cantidad de variables. Refaat, también sugiere que nos enfoquemos en dos puntos:

1. Eliminar las variables independientes que no contribuyen o tienen muy baja contribución en el modelo.

2. Mantener las variables que son buenas o potenciales predictoras de nuestro modelo, debido a que contienen la mayor información en ellas.

El concepto de indicador de Poder Predictivo, requiere necesariamente una variable dependiente y otra independiente. Siendo la medida del indicador el grado de asociación entre las variables. Existen diferentes medidas de poder predictivo, dependiendo de la naturaleza de las variables[11].

Una de las medidas de poder predictivo más empleada en las metodologías de Riesgo de Crédito[12] es el indicador **InformationValue (valor de la información)**y el indicador **WOE (Weight of Evidence – Peso de la evidencia)**, que suponen que la variable dependiente es dicotómica y la variables independiente tiene rangos ya definidos (i=1,... L). Estos indicadores son empleados en el desarrollo, implementación y

[10] [19] Mamdouh (pág. 207)
[11] [19] Mamdouh (pág. 209)
[12] [24] Naem Siddiqi (pág. 83)

seguimiento de los modelos predictivos de riesgo de crédito como establecen los autores Naeem Siddiqi y Raymond Anderson[13].

WOE (Weight of Evidence – Peso de la Evidencia).

Es el indicador de riesgos de relativo de ocurrencia del evento (en nuestro caso el incumplimiento de pago de las cuotas de crédito) en un rango determinado de las variables independiente.

$$W_i = log_e\left(\frac{\left(\frac{P_i}{\sum P_i}\right)}{\left(\frac{N_i}{\sum N_i}\right)}\right)$$

Donde:

Ni: La cantidad de casos de **no ocurrencia** del evento en el rango i-ésimo.

Pi: La cantidad de casos de **ocurrencia** del evento en el rango i-ésimo.

La interpretación del WOE será:

El nivel de riesgo asociado al rango de la variable independiente, según las distribuciones de los resultados.

IV (InformationValue – Valor de la información)

Indicador de poder predictivo (o discriminación)calculado mediante la suma de los WOE en los rangos ponderados por la diferencias proporcionales halladas en los rangos. Esta medida resumen nos indica cuan relevante es la variable independiente para poder pronosticar la variable dependiente.

$$IV = \sum_{i=1}^{L}\left|\left(\left(\frac{P_i}{\sum P_i}\right) - \left(\frac{N_i}{\sum N_i}\right)\right) xW_i\right|$$

[13] [25] Raymon Anderson (pág. 251)

Existen niveles de IV de las variables independientes que según la experiencia tienen niveles de relevantes de poder predictivo.

Según Anderson [14]sugiere como alertas:

Rango de IV	**Interpretación**
Menos de 0.02	No predictivo, poco aporte en poder discriminador.
Entre 0.02 a 0.1	Débil,
Entre 0.1 a 0.3	Medio,
Mayor a 0.3	Fuerte, alto poder discriminador.

II.1.4 Regresión Logística

Dada una población de observación

$$\mathcal{L} = \{(x_i, y_i): i = 1,2 \dots, n\} \qquad (1.1)$$

De manera matricial, se puede expresar como una matriz de dimensiones nX(r+1):

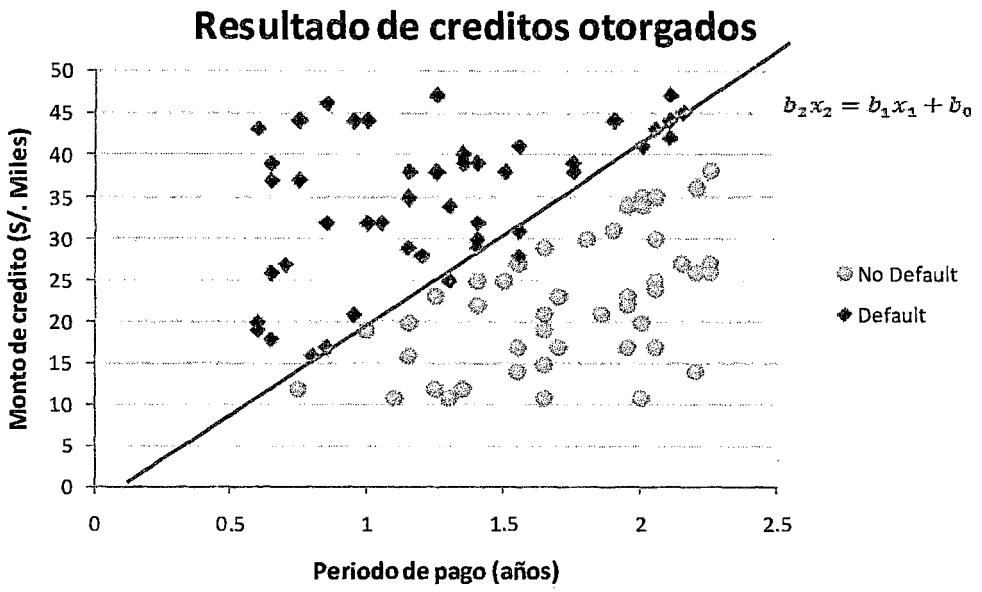
$$(X_{nxr}|Y_{nx1}) = \begin{pmatrix} x_1 & y_1 \\ \cdots & | \cdots \\ x_n & y_n \end{pmatrix} \qquad (1.2)$$

Donde cada observación está representada por el par (x_i, y_i), cada x_i ($x_i \in \Re^r$) representa las"r" características de la i-ésima observación, mientras que la variable dependiente de la i-ésima observación es y_i. En caso de un problema separación entre dos poblaciones (Π_1, Π_2), la variable dependiente es dicotómica ($y_i \in \{1,0\}$). Notación de Izenman[15].

Una representación gráfica del problema de la estimación de una regla de separación lineal de 2 poblaciones se muestra en el **gráfico 2**.

[14][25] Raymon Anderson (pág. 251)
[15][3] JulianIzenman pág. 370

Gráfico 2



En el gráfico anterior, se presenta el caso de dos poblaciones que son separables linealmente por una recta, en función de las dos variables independientes (Monto de crédito y Periodo de pago). De modo que el problema estadístico es estimar la ecuación adecuada que separe ambas poblaciones (recta celeste en el gráfico).

——— (1.3)

La interpretación breve de la expresión anterior es la siguiente: "*Dado que se conoce las características de la observaciónx, se asigna una probabilidad de pertenecer a uno de los dos grupos en base a patrones históricos*".

Se emplea la función sigmoidea Logit, debido a su propiedad de enlace de un dominio en los números reales a un rango acotado entre 1 y 0.Donde consideramos que =1, implica que la probabilidad de la i-ésima pertenezca a una población 1 sea 100% (por ejemplo: población de clientes que incumplen sus pagos).

Las reglas de discriminación estadística y las probabilidades de pertenecerá una población para clasificación binaria son:

$$L(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^t \boldsymbol{x} \qquad (1.4)$$

$L(\boldsymbol{x})$, esta también denominada score de discriminación de poblaciones. Nos encontraríamos ante la necesidad de encontrar una función que resuelva la relación:

$$Probabilidad(Default_i) = F(Score_i)$$

Las funciones que cumplen esta relación $x \in R \xrightarrow{F} y \in [0,1]$, son denominadas funciones sigmoideal **Anexo [1]**,entre ellasuna de las más usadas es la función ***Logit*** *(Logística)*.

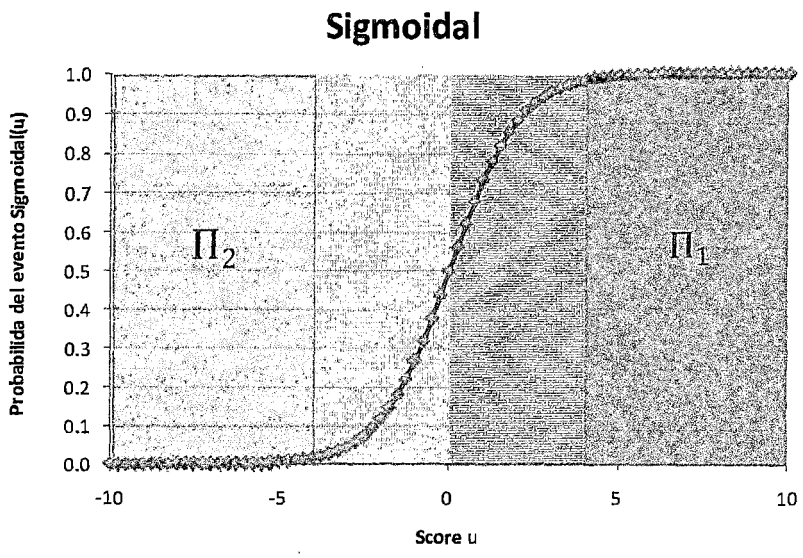
$$p(\Pi_1|x) = \frac{e^{L(x)}}{1+e^{L(x)}} \qquad (1.5)$$

$$p(\Pi_2|x) = \frac{1}{1+e^{L(x)}} \qquad (1.6)$$

La función de enlace que nos permite discriminar es la función sigmoidal logística, tales también pueden ser el caso como las funciones log- log y probit (la función de probabilidad normal inversa)[16].

[16] [3] Julianlzenman pág. 257-261

Gráfico 3



En el gráfico la función sigmoidal de activación es Logit,

——— (1.7)

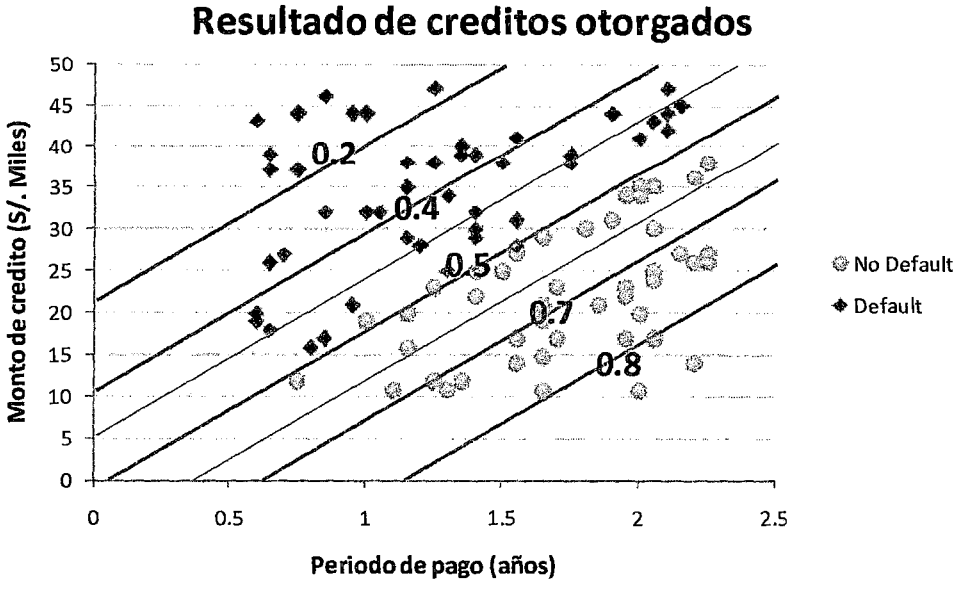
Para valores pequeños del score, la probabilidad de pertenecer a la población 1 será muy pequeña (como puede ser en riesgo de crédito, a la población de clientes que incumplan el pago de sus préstamos).

En el gráfico 2 (solo se tiene 2 variables regresoras), por ejemplo buscamos hallar los valores de los estimados que definan nuestra recta de discriminación entre las poblaciones, como se muestra en la siguiente expresión:

Cuando el Score toma el valor de cero, es decir la probabilidad de las observaciones de pertenecer a las 2 poblaciones es la misma, se puede graficar (la recta de discriminación) usando la expresión:

Además, cabe destacar que uno de los atractivos de un modelo de regresión logística es la simpleza de sus predicciones. Los contornos son simples líneas rectas (en mayores dimensiones serían hiperplanos), son líneas de isoprobabilidad de pertenecer a una población detalladas en el gráfico4[17].

Gráfico 4



II.3.1.1Multicolinealidad

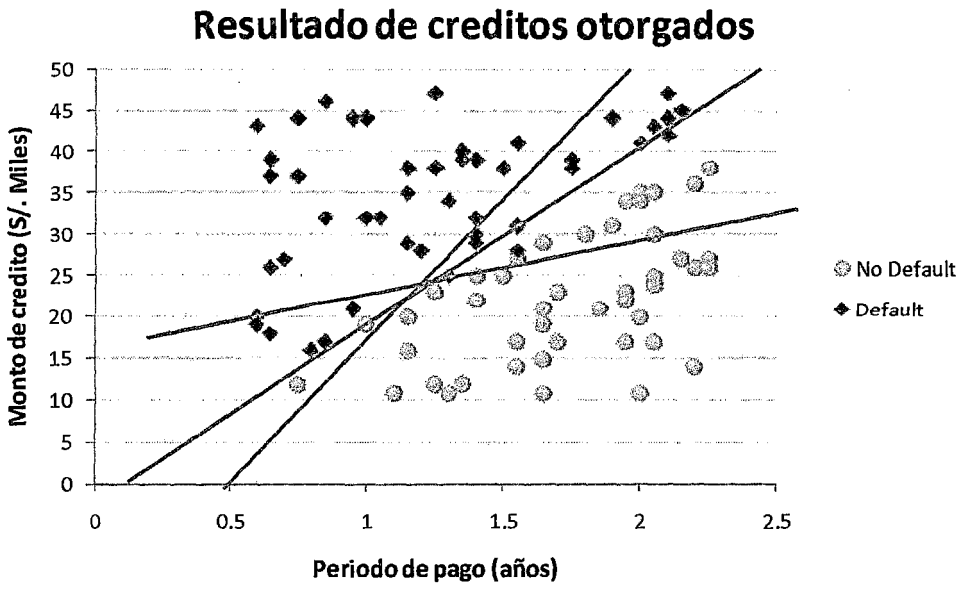
La presencia de Multicolinealidad tiene una gran cantidad de efectos graves sobre los estimadores de coeficientes de regresión y en modelos que impliquen una ecuación lineal. Estos problemas de Multicolinealidad o redundancia de información afecta a la regresión logística con valores grandes de varianza de los estimadores[18].

En el gráfico 5 siguiente exponemos las implicancias de hallar alta Multicolinealidad entre nuestras variables de interés. El efecto provocado es que al determinar muestras aleatorias que sean representativas de la población (m1, m2 y m3) para estimar la recta

[17][10] MAYSA pág. 20-22
[18]Montgomery Análisis de Regresión y Draper Smith

de discriminación de las dos poblaciones obtengamos diferentes ecuaciones (l1, l2 y l3) las cuales son mostradas en el gráfico.

Gráfico N° 5



Partiendo de los supuestos de un modelo de lineal múltiple:

(1.8)

Sigue una distribución aleatoria y

Var ( )

EL impacto de la Multicolinealidad es desarrollado de la siguiente manera, al detallar el error cuadrático de los estimados de la ecuación de recta:

(1.9)

=

A partir de la expresión , al reemplazarlo sobre la expresión desarrollar por el vector y.

$$= \beta^t X^t X (X^t X)^{-1} (X^t X)^{-1} X^t X \beta + \sigma^2 tr[X^t X (X^t X)^{-1} (X^t X)^{-1}] - \beta^t \beta$$

$$= \beta^t \beta + \sigma^2 tr[(X^t X)^{-1}]$$

Por lo tanto:

$$E[(\beta - b)^t (\beta - b)] = \sigma^2 tr[(X^t X)^{-1}] \qquad (1.10)$$

De la última expresión se pone en evidencia que la consistencia de los estimadores del modelo (las desviaciones estándar de los coeficientes), es notorio que dependerá de la carga de información cuadrática por variable contenida en la matriz de datos $X^t X$.

SI asumimos que $X^t X$ tienen r diferentes valores característicos $(\lambda_1, \lambda_2, \dots, \lambda_r)$ y los correspondientes vectores característicos normalizados $(v_1, v_2, \dots, v_r)$, se puede escribir de la siguiente manera:

$$V^t (X^t X) V = diag(\lambda_1, \lambda_2, \dots, \lambda_r) \qquad (1.11)$$

Entonces: $tr[V^t (X^t X) V] = tr[V V^t (X^t X)] = tr(X^t X) = \sum_{i=1}^{r} \lambda_i$ (1.12)

Si consideramos que la expresión que representa la consistencia de los coeficientes estimados es la matriz $(X^t X)^{-1}$, con los valores característicos $(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_r})$

$$E[(\beta - b)^t (\beta - b)] = \sigma^2 tr[(X^t X)^{-1}] \qquad (1.13)$$

$$E[(\beta - b)^t (\beta - b)] = \sigma^2 \sum_{i=1}^{r} \frac{1}{\lambda_i} \qquad (1.14)$$

Es evidente ahora que valores muy pequeños del valor característico, tal como $\lambda_i = 0.0001$ será sobre estimada por $1000\sigma^2$ como señala*XinYanXiao*[19].

En base al señalado, se desarrollan 2 criterios detección efecto de Multicolinealidad:

1) Los números condición:

[19][12] XinYaoXiaoGangXu pág. 81-87

Una comparación de los valores característicos hallados en la matriz de variables centrada analizados (matriz de variables regresoras cuantitativas), se realiza una comparación de 2 a dos entre las variables:

$$\eta_j = \sqrt{\frac{\lambda_{máx}}{\lambda_j}} \qquad (1.15)$$

Este criterio fue establecido por Webster, Gunst y Mason [1974], como Proporción de Descomposición de varianza para los datos. Que es evaluado en la siguiente matriz:

Descomposición de varianza para regresores centrados			
Valor característico	**Número de condición**	X1	X2
λ_1	$\sqrt{\frac{\lambda_{máx}}{\lambda_1}}$		
λ_2	$\sqrt{\frac{\lambda_{máx}}{\lambda_2}}$		

Criterio establecido es aquellos números condición mayores que 30, y proporciones de descomposición de varianza mayor que 0.5.

Un método de diagnóstico de Multicolinealidad es sugerido por Belsey con el procedimiento siguiente:

1.- Grandes valores de número de condición indican dependencia lineal entre variables. Observaremos los índices de condición que tengan valores superiores a 30.

2.-Grandes proporciones en la fila de la matriz mostrada, nos indican los candidatos de dependencia. Aquellos valores con proporciones mayores a 0.5, serán variables muy

asociadas linealmente y a la variable de la fila revisada. Es más la proporción de varianza mayor en la fila será la variable más dependiente[20].

2) Valor de Inflación de varianza:

El factor de inflación de varianza es una medida que puede ser empleada para cuantificar la Multicolinealidad. El i-ésimo factor de inflación de varianza es la versión escalada (variables centradas) del coeficiente de correlación múltiple entre la variable independiente i-ésima y el resto de variables independientes.

$$VIF_i = \frac{1}{1-R_i^2} (1.16)$$

Un factor de inflación de varianza cuyo valor supere el número 10, podría ser un indicador potencial de problemas de multicolinealidad.*XinYao*[21].

II.3.1.2Evaluación de potenciales conjuntos de variables predictivas mediante el criterio de información de Akaike(1973)

Puede ser ventajoso para una selección de variables por dos principales motivos de su uso frecuente. El primero es debido a su bondad de ajuste balanceado y a la penalidad para la complejidad de un modelo. Entenderemos como el AIC más pequeño para el mejor modelo.

Es una medida de bondad de ajuste, tal que está comprendida de factores: el primero que es el valor negativo de la verosimilitud asociada con el modelo ajustado y el segundo es K una medida de complejidad del modelo (donde K=r+2).

$$AIC = 2[-\log(L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4, .., \hat{\beta}_r, \hat{\sigma}^2 \mid Y)) + K] (1.17)$$

[20][10] Douglas Montgomery pág. 303-307
[21][12] XinYaoXiaoGangXu pág. 85-87

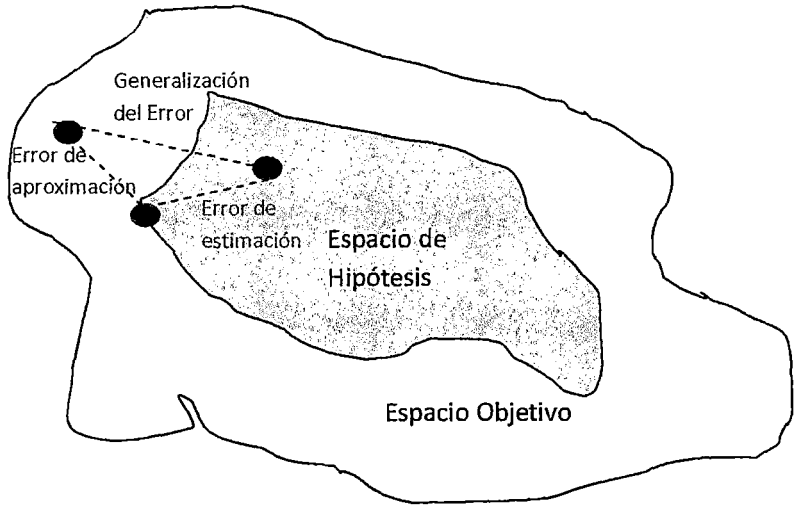
Los betas y sigma son los valores estimados de nuestro modelo.

II.1.5Máquinas de Vectores de Soporte

El fundamento de la técnica de Máquina de Vectores de Soporte está construida sobre las bases de la Teoría de Aprendizaje Estadístico (StatisticalLearningTheory) Vapnik 1998.

El objetivo en el modelamiento es elegir un modelo desde un espacio de hipótesis (supuestos), aquel modelo que sea más cercano a una función en el espacio objetivo. Luego que es posible determinar 2 tipos de errores.

Gráfico N° 6



1.-Error de aproximación: Es una consecuencia del espacio de hipótesis establecidas no cubra el espacio objetivo. Una elección pobre de un modelo resulta en un error de aproximación grande.

2.-Error de estimación: Es un error debido al procedimiento de aprendizaje empleado, cuyos resultados son derivados de una selección no-óptima de la técnica desde el espacio de hipótesis.Steve Gunn[22].

En esta sección, describiremos el concepto básico de SVM (Support Vector Machine)Frente a un problema de clasificación binaria. Estos conceptos fueron propuestos por Kecman(2001). La forma en que se aborda y se soluciona esta problemática es mediante una aproximación matemática de regla de clasificación. Cuya aproximación no se basa en supuestos de distribuciones de probabilidad, sino más bien en replicación de diferentes potenciales reglas y la búsqueda de la regla que disminuya la tasa de error de clasificación entre poblaciones.

Dado un conjunto de aprendizaje por pares (x_i, y_i), que forman la data de aprendizaje:

$$L = \{(x_i, y_i) : i = 1,2,....,n\},\ i = 1,2,...,n \text{ Donde } x_i \varepsilon R^n \text{ y } \ y_i \varepsilon \{+1,-1\} \qquad (2.1)$$

El problema de clasificación binaria es emplear L, para construir una función $f: R^r \xrightarrow{f} R$, tal que:

$$C(x) = sign(f(x)) \qquad (2.2)$$

$$C(x) = \begin{cases} 1 & si \quad f(x) \geq 0 \\ -1 & si \quad f(x) < 0 \end{cases}$$

La función de separación f, clasifica a cada observación x dentro de una de las 2 poblaciones Π_+ o Π_-, dependiendo del valor de f(x) si C(x) es +1 (si $f(x) \geq 0$) o -1 (si $f(x) < 0$). El objetivo es identificar la función o regla de clasificación f(x), que asigne a todos los positivos a la población Π_+ y todo negativo a Π_-. En la práctica, es sabido que no es posible alcanzar 100% de asignaciones o predicciones correctas. *SteveGunn*

[22] [13] Steve R. Gunn pág. 2

Partimos de tres supuestos importantes, para el desarrollo teórico de la técnica SVM desde un enfoque más sencillo hacia uno más complejo:

I.- Las poblaciones son linealmente separables.

II.- Las dos poblaciones se pueden separar mediante una regla de clasificación sin errores de clasificación.

III.- Existe un supuestos sobre los parámetros, que es empleado por simplicidad.

$$min_i|\beta_0 + X_i\beta| = 1 \qquad (2.3)$$

Este supuesto en otras palabras se refiere a: "La norma del vector de coeficientes (de la recta de clasificación) será igual a la inversa de la distancia,del punto más cercano de la recta de clasificación al conjunto dedatos".*Steve Gunn*[23]. Como muestra el gráfico Nro. 7, las rectas son denominadas hiperplanosCanónicos.

Luego definimos la ecuación de la recta de clasificación como:

$$\{x: f(x) = \beta_0 + \beta^t x = 0\} \qquad (2.4)$$

Donde:

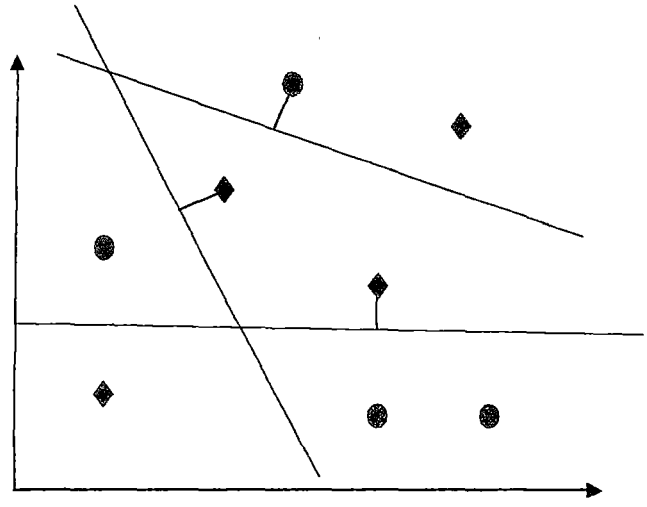
- β es el vector de coeficientes de la ecuación de la recta

- $\|\beta\|$ es la norma Euclidiana del vector de coeficientes

- β_0, es el sesgo o umbral de la recta

[23][13] Steve R. Gunn pág. 6

Gráfico N° 7



En base al supuesto III y ecuación (2.4), definimos las siguientes expresiones:

$$\beta_0 + \beta^t x_i \geq +1, \quad Si \ y_i = +1 \qquad (2.5)$$

$$\beta_0 + \beta^t x_i \leq -1, \quad Si \ y_i = -1 \qquad (2.6)$$

Y mediante una combinación de ambas expresiones, se llega a:

$$y_i(\beta_0 + \beta^t x_i) \geq +1, \quad i = 1,2,3,\dots,n \qquad (2.7)$$

La cantidad $y_i(\beta_0 + \beta^t x_i)$, es denominado margen de (x_i, y_i) con respecto al hiperplano (2.4).

Además desde que la distancia de una observación al hiperplano se define como:

$$d(x_i, f(x)) = \frac{|\beta_0 + \beta^t x_i|}{\|\beta\|} \qquad (2.8)$$

De (2.7) y (2.8):

$$Min\ d(x_i, f(x)) = \frac{1}{\|\beta\|} (2.9)$$

Considerando una distancia "1/**h**" de la observación x_i al hiperplano L. La distancia **"1/h"**, estaría restringido a:

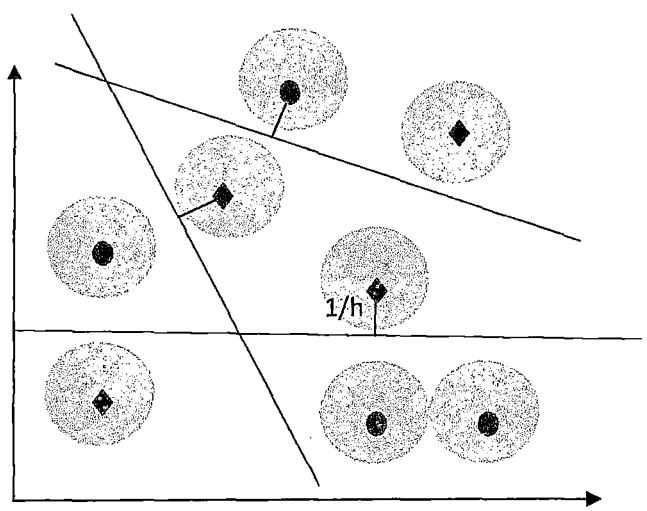
$$\|\beta\| > h$$

En consecuencia esto delimitaría una región de separabilidad mínima entre las observaciones y el hiperplano. Esto es denominado **restricción de hiperplanos canónicos** y es gráficamente representado de la siguiente forma:

$$d(x_i, f(x)) \geq \frac{1}{h} (2.10)$$

El hiperplano no puede estar más cerca que 1/h

Gráfico N° 8



Es posible también obtener a partir de buscar las igualdades dentro de las ecuaciones (2.5) y (2.6), de la manera siguiente:

$$H_{+1}: (\beta_0 - 1) + \beta^t x = 0 (2.11)$$

Es denominado hiperplano limítrofe de la población Π_+.

$$H_{-1}: (\beta_0 + 1) + \beta^t x = 0 (2.12)$$

Es denominado hiperplano limítrofe de la población Π_-.

En el conjunto de datos de aprendizaje definido inicialmente **L** (2.1), encontraremos observaciones (x_i, y_i), cumplan las igualdades (2.10) y (2.11), estos serán denominados Vectores de Soporteuobservaciones críticas de clasificación. Estas observaciones sin típicamente un pequeño porcentaje del total de observaciones de la muestra.

Si definimos 2 observaciones que cumplan las igualdades señaladas, dentro de ambos hiperplanos:

$$\beta_0 + \beta^t x_{-1} = -1 (2.13)$$

$$\beta_0 + \beta^t x_{+1} = +1 (2.14)$$

La diferencia de ambas ecuaciones resulta:

$$\beta^t x_{+1} - \beta^t x_{-1} = 2 (2.15)$$

Y la suma de ambas ecuaciones resulta:

$$\beta_0 = -\frac{1}{2}\{\beta^t x_{+1} - \beta^t x_{-1}\} (2.16)$$

II.1.5.1 Caso de poblaciones separables linealmente

Las distancias de las observaciones señaladas a la recta de clasificación o hiperplano f(x) son:

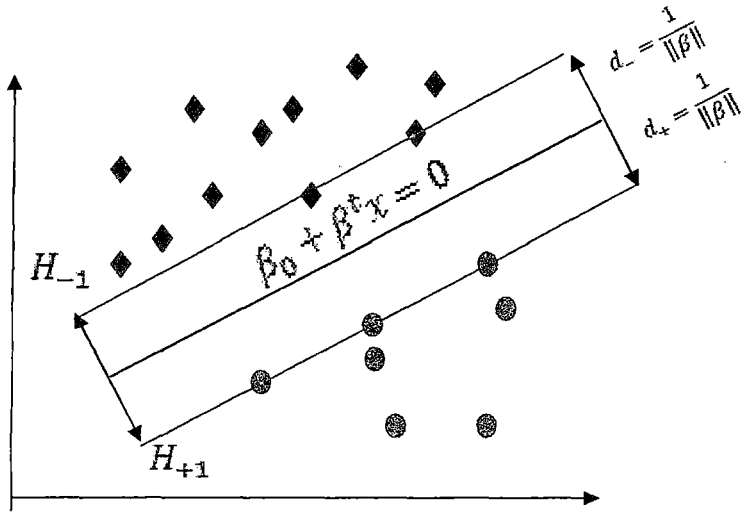
$$d_- = \frac{|\beta_0 + \beta^t x_{-1}|}{\|\beta\|} = \frac{1}{\|\beta\|} (2.17)$$

$$d_+ = \frac{|\beta_0 + \beta^t x_{+1}|}{\|\beta\|} = \frac{1}{\|\beta\|} (2.18)$$

La distancia mínima de separación entre las poblaciones Π_+ y Π_-, es denominada margen de separación.

$$d = d_+ + d_- = \frac{2}{\|\beta\|} (2.19)$$

Gráfico N° 9



La expresión (2.9) enmarca la problemática de las reglas de clasificación de la Máquina de Vectores de Soporte. El problema es encontrar el hiperplano de separación óptimo; denominado hiperplano que maximiza el margen de separación entre las poblaciones —, restringida a la condición (2.7). Por ello de manera equivalente planteamos el siguiente problema de optimización No-Lineal. Por ende el problema de discriminación se puede redefinir como un problema de minimización de la función objetivo.

- (2.20)
(2.21)

Es evidente que nos encontramos frente a un problema de minimización cuadrática restricto. El hiperplano de clasificación óptimo es llamado fuerte o solución marginal. Para hallar la solución deseada aplicaremos el teorema de Karush-Khunt-Tuker (**anexo 4**), que requiere la aplicación de multiplicadores Lagrangianos y verificar los supuesto de suficientes y necesarios del teorema.

El siguiente paso es minimizar la función primal de **F**, con respecto a las variables β y β_0, y luego maximizar el resultado mínimo de F con respecto a las variables dualesα.

$$F_p(\beta_0, \beta, \alpha) = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i\{y_i(\beta_0 + \beta^t x_i) - 1\} \quad (2.22)$$

Donde:$\alpha = (\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_n)^t \geq 0$ (2.23)

La ecuación (2.23) son los Lagrangianos no negativos.

Luego los supuestos suficientes y necesarios del procedimiento de solución de optimización Krush-Kunt-Tukerson los siguientes:

$$\frac{\delta F_p(\beta_0,\beta,\alpha)}{\delta \beta_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0 (2.24)$$

$$\frac{\delta F_p(\beta_0,\beta,\alpha)}{\delta \beta} = \beta - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \quad (2.25)$$

$$y_i(\beta_0 + \beta^t x_i) - 1 \geq 0 \quad (2.26)$$

$$\alpha_i \geq 0 \quad (2.27)$$

$$\alpha_i\{y_i(\beta_0 + \beta^t x_i) - 1\} = 0 \quad (2.28)$$

Resolviendo las ecuaciones (2.24) y (2.25), se obtienen las siguientes expresiones.

$$\sum_{i=1}^{n} \alpha_i y_i = 0 (2.29)$$

$$\beta^* = \sum_{i=1}^{n} \alpha_i y_i x_i \quad (2.30)$$

Substituyendo (2.29) y (2.30), en la ecuación primal funcional, de manera que minimize el valor la posible de los coeficientes (2.22).

$$F_p(\beta_0, \beta, \alpha) = F_D(\alpha) = \frac{1}{2}\|\beta^*\|^2 - \sum_{i=1}^{n} \alpha_i\{y_i(\beta_0^* + \beta^{*t} x_i) - 1\}$$

$$F_D(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j (x_i{}^T x_j) - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j\, y_i y_j (x_i{}^T x_j) + \sum_{i=1}^{n} \alpha_i$$

$= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j (x_i{}^T x_j)$ (2.31)

Lo siguiente es encontrar los valores de los multiplicadores Lagrangianos α para maximizar la función dual sujeta a las restricciones mencionadas, expresado matricialmente:

$$Maximizar\ F_D(\alpha) = \mathbf{1}_n^t \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^t \boldsymbol{H} \boldsymbol{\alpha} \quad (2.32)$$

$$Restricto\ a: \boldsymbol{\alpha} \geq \mathbf{0}, \boldsymbol{\alpha}^t \boldsymbol{y} = 0 \quad (2.33)$$

Donde:

$\boldsymbol{y} = (y_1, \ldots, y_n)^ty\boldsymbol{H} = (H_{ij})$es una matriz cuadrada de grado "n" con$H_{ij} = y_i y_j (x_i{}^t x_j)$.

Ahora, si suponemos $\hat{\alpha}$ sea el vector de solución de Lagrangianos y por ende el problema de optimización:

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \, x_i \quad (2.34)$$

De la restricción (2.33) $\alpha_i > 0$, y el supuesto antes mencionado (2.28) $\alpha_i \{y_i(\beta_0 + \beta^t x_i) - 1\} = 0$

Tendría $y_i(\beta_0 + \beta^t x_i) - 1 = 0$ para cumplir la expresión (2.28), y como solo los Vectores de Soporte cumplen esta la igualdad $y_i(\beta_0 + \beta^t x_i) = 1$, $\hat{\beta}$ sería una función lineal de los Vectores de Soporte$\{x_i, i \in sv\}$, donde sv es el subconjunto de índices que identifica a los Vectores de Soporte de la recta de clasificación.

$$\hat{\beta} = \sum_{i \in sv}^{n} \hat{\alpha}_i y_i \, x_i \quad (2.35)$$

En este, también se puede interpretar que los Vectores de Soporte contienen toda la información necesaria para determinar el hiperplano óptimo de clasificación. *Alan Julian.*[24]

Así también, a pesar que el sesgo β_0 no fue determinado explícitamente por la solución del problema de optimización, es posible estimarlo realizando los reemplazos de (2.35) en (2.28).

$$\hat{\beta}_0 = \frac{1}{|sv|}\sum_{i\in sv}^{n}\left(\frac{1-y_i x_i^t \hat{\beta}}{y_i}\right) \qquad (2.36)$$

$|sv|$: Número de Vectores de Soporte hallados.

De modo que podemos obtener un hiperplano de clasificación:

$$\widehat{f(x)} = \hat{\beta}_0 + x^t\hat{\beta} (2.37)$$

$$= \hat{\beta}_0 + \sum_{i\in sv}^{n}\hat{\alpha}_i y_i (x_i^t x_i) (2.38)$$

Evidentemente, en la estimación del hiperplano de clasificación son relevantes los Vectores de Soporte; mientras que aquellos vectores (observaciones o casos) que no lo sean no juegas un rol determinante en determinar la regla de clasificación. Entendamos la regla de clasificación lo siguiente:

$$C(x) = sign\{\widehat{f(x)}\} (2.39)$$

II.1.5.2 Caso de poblaciones no-separables linealmente

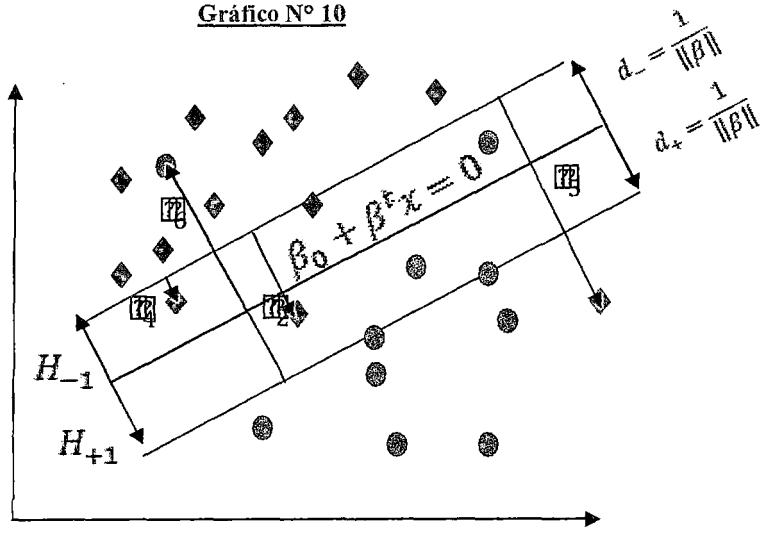
En aplicaciones reales, es poco probable que encontremos una separación clara y bien definida entre ambas poblaciones, siendo lo más probable una superposición entre ambas poblaciones. Como es lógico se podría esperar que existan casos atípicos de una población que posean características más asociadas a la segunda población. Entonces al

[24][3] Alan Julian pág. 374

momento de construir una regla de clasificación esperamos algunos casos de mala clasificación de observaciones en cada clase señalada. Una causa de la mala clasificación podría deberse alto nivel de ruido entre ambas clases de poblaciones.

Existen también los casos No-separables linealmente, esto sucede cuando las poblaciones son separables pero no linealmente. Teniendo como antecedente la existencia de errores de mala clasificación debemos considerar un modelo de optimización flexible ante estos casos, el siguiente paso es penalizar aquellos errores iniciales de mala clasificación.En esta problemática, podemos reformulando el problema de optimización de una manera más flexible deseada, estableciendo el concepto de **variables de holgura** no negativas, para cada observación de la data de aprendizaje estudiada ( ). Estas variables de holgura representaran el error de mala clasificación de las variables entre las 2 poblaciones como muestra el gráfico 10.

Gráfico N° 10



Es decir:

(2.40)

Dada las variables de variables de holgura modificamos la restricción (2.26) a la siguiente expresión:

$$y_i(\beta_0 + \beta^t x_i) + \varepsilon_i \geq 1 \quad (2.41)$$

Aquellas observaciones que tienen $\varepsilon_i = 0$, mantendrán la restricción (2.25).

Además, conocemos según la Cibergrafía [c-5] Matriz Algebra, señala dos tipos de escala de medición de la distancia entre observaciones. Para proseguir con el desarrollo de la técnica emplearemos la llamada **1-norma.**

Se denominada un problema de optimización de margen-suavizado bajo la 1-norma, el proceso de hallar aquellos β_0, β y ε. Minimizando la expresión:

$$Minimizar\ \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi \quad (2.42)$$

$$Restricto\ a: \xi_i \geq 0,\ y_i(\beta_0 + \beta^t x_i) \geq +1 - \xi_i,\ \ i = 1,2,3,\dots,n \quad (2.43)$$

Donde C>0 es un parámetro de regularización, que cumple el rol de penalizar la cantidad de la carga de los errores admitidos en el problema de optimización. C es también denominado sintonizador constante que controla el tamaño de la holgura de las variables y equilibra la función de optimización.

Obtenemos así una función primal del problema de optimización, $F_p = F_p(\beta_0, \beta, \xi, \alpha, \eta)$

$$F_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\xi_i + C\sum_{i=1}^{n}\alpha_i\{y_i(\beta_0 + \beta^t x_i) - (1 - \xi_i)\} - \sum_{i=1}^{n}\eta_i\xi_i \quad (2.44)$$

Con los Lagrangianos

Realizando el desarrollo de la diferenciación de la función primal Fp respecto a:

$$\frac{\delta F_p}{\delta \beta_0} = -\sum_{i=1}^{n}\alpha_i y_i, (2.45)$$

$\frac{\delta F_p}{\delta \beta} = \beta - \sum_{i=1}^{n} \alpha_i y_i x_i$,(2.46)

$\frac{\delta F_p}{\delta \beta} = C - \alpha_i - \eta_i$(2.47)

Estableciendo que las derivadas halladas sean iguales a cero conseguimos:

$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad \beta^* = \sum_{i=1}^{n} \alpha_i y_i x_i, \ \alpha_i = C - \eta_i$ (2.48)

Substituyendo las (2.44) en expresiones encontradas en la función dual,

$F_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i{}^t x_j)$(2.49)

Además de la restricción $C - \alpha_i - \eta_i = 0$ y $\eta_i \geq 0$, luego tendríamos $0 \leq \alpha_i \leq C$, se deslinden las condiciones necesarias de Krush-Kuhn-Tucker, y por ende es posible plantear el problema de dual de maximización en notación matricial. Encontrando $\boldsymbol{\alpha}$

$$Maximizar\ F_D(\boldsymbol{\alpha}) = \mathbf{1}_n^t \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^t \boldsymbol{H} \boldsymbol{\alpha} \quad (2.50)$$

$$Restricto\ a: \boldsymbol{\alpha}^t \boldsymbol{y} = 0, \mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}n \quad (2.51)$$

La diferencia única diferencia entre las últimas expresiones y (2.32) (2.33), es que en la últimas los coeficientes del Lagrangiano $\alpha_i, i = 1,2,..,n$ con cada uno limitados por C; este límite superior restringe la influencia de cada observación en determinar la solución óptima. Este tipo de restricción es referida como una *restricción de caja*, debido a que $\boldsymbol{\alpha}$es restricta por una caja de lado C en el borde positivo. Desde la restricción (2.51) vemos que la región de factibilidad para la solución para este problema de optimización convexa es la intersección del hiperplano $\boldsymbol{\alpha}^t \boldsymbol{y} = 0$con la caja restricción $\mathbf{0} \leq \boldsymbol{\alpha} \leq C\mathbf{1}n$. S$C = \infty$, el problema se reduciría a el caso de márgenes – difícil de separar.

Si resuelve el problema de optimización, luego:

$$\hat{\beta} = \sum_{i \in sv}^{n} \hat{\alpha}_i y_i \, x_i \quad (2.52)$$

Siendo para tal caso los pesos del vector óptimo, aquel en el cual el conjunto de Vectores de Soporte contiene estas observaciones en la data de aprendizaje inicial.

II.1.5.3 Validación cruzada del error de pronóstico

Entre los diferentes métodos aceptables para estimar el error de predicción (error del modelo), es la denominada validación cruzada (cross-validation en inglés).

Si suponemos que D es una muestra aleatoria que sigue una distribuciónde probabilidad conjunta de $(\boldsymbol{X}, Y)$ en $(r+1)$ dimensiones o variables. Si n=2m, podemos aleatoriamente dividir en 2 subconjuntos la data, tratando un conjunto de datos de entrenamiento como L y la otra data T como testeo. Donde $D = L \cup T$ y $0 = L \cap T$.

Si establecemos que $T = \{(X_i', Y_i'), i = 1,2,..,m\}$. Una estimación del error de estimación PE_Res establecida de la manera siguiente.

$$\widehat{PE} = \frac{1}{m}\sum_{i=1}^{m}(Y_i' - \hat{\mu}(X_i'))^2 \qquad (2.53)$$

Donde mu, representa una estimación del valor Yi, tomemos como ejemplo la estimación mediante el método de regresión de mínimos cuadrados ordinarios

$$\hat{\mu}(X_i') = \widehat{\beta_0} + X_i'^T\widehat{\beta_{ols}}.$$

El conjunto de datos de aprendizaje y el de testeo son luego intercambiados y el resultado de los estimados de PE_Res promediados como la estimación final del error pronosticado con el modelo elegido.

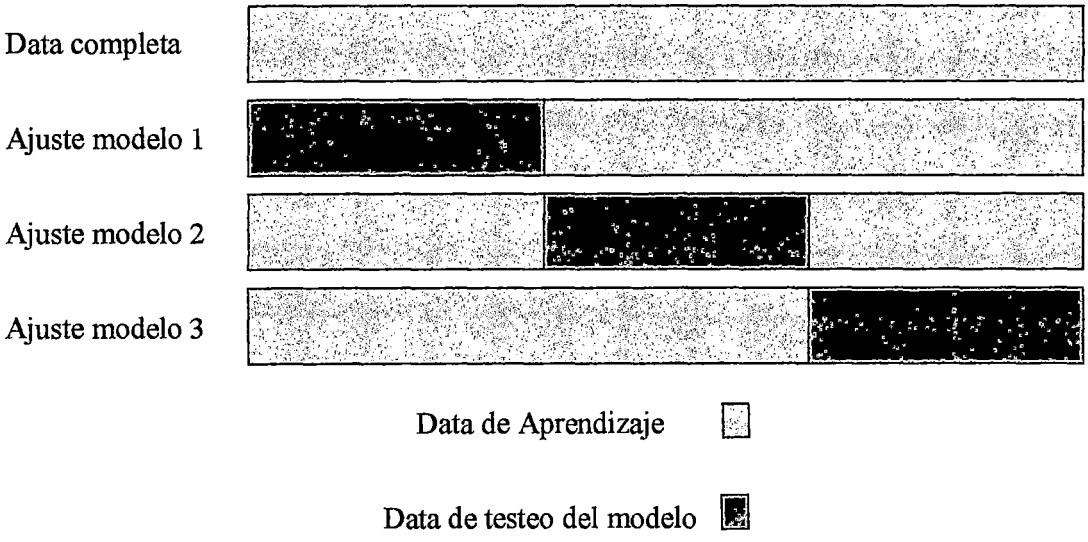
Para generalizar el procedimiento señalado, asumiendo que la n=Vm done V es un número entero pequeño entre 5 o 10. Dividiremos de manera aleatoria el conjunto de datos iniciales en subconjuntos disjuntos $T_v, v = 1,2,..,V$, por lo tanto tendremos V diferentes escenarios de muestra testeando el mismo modelo, obtendremos así

intencionalmente un sinceramiento del error de pronóstico. Que nos servirá en el proceso de modelamiento como garantía de la extensión o generalización de nuestros resultados experimentales.

Observamos en el gráfico N°11 el procedimiento de cross-validation el procedimiento en data es además aleatoriamente particionada en data de entrenamiento e independientes datas de prueba vía k-veces validación cruzada.

Gráfico N° 11

Cross-validation 3 fold, muestras de testeo disjuntas



Cada uno de los k subconjuntos actúa como un conjunto de prueba para el modelo entrenado con los restantes k-1 subconjuntos.La ventaja de la validación cruzada es que el impacto de la dependencia de la data de aprendizaje es minimizado y la confiabilidad de los resultados puede ser mejorado [25](Salzberg 1997).

[25][3] Julianlzenman pág. 121-122

II.1.5.4 Máquina de Vectores de Soporte No lineal

Debemos saber que existen escenarios de poblaciones, en los cuales la clasificación lineal no sería apropiada. Es posible extender las construcción del SVM no lineal, sobre una transformación (producto interno) de la data $\langle x_i, x_j \rangle = x_i^T x_j, i, j = 1,2, \ldots, n$.

Si suponemos que realizamos la transformación de cada observación, $x_i \in \Re^r$ sobre la data de aprendizaje empleando un mapeo no lineal $\Phi : \Re^r \to H$, donde H es un espacio de características N_H – dimensional. Asumiremos que H sea el espacio de Hilbert de la valores reales de la función sobre $\Re$ con producto interno $\langle . \rangle$ y normal $\| . \|$.

$$\boldsymbol{\phi}(x_i) = (\phi_1(x_i), \phi_2(x_i), \ldots, \phi_{N_H}(x_i))^T \in H, i = 1,2,..,n$$

La muestra transformada es luego $\{\boldsymbol{\phi}(x_i), y_i\}$, donde $y_i \in \{-1, +1\}$ identificando ambas clases. Si sustituimos $\boldsymbol{\phi}(x_i)$ por x_i en el desarrollo de la técnica SVM lineal, podríamos abordar el problema de optimización solamente empleando el producto interno $\langle \boldsymbol{\phi}(x_i), \boldsymbol{\phi}(x_j) \rangle = \boldsymbol{\phi}(x_i)^T \boldsymbol{\phi}(x_j)$. La dificultad en el uso de la transformación no lineal de este modo es el cálculo tales como el producto interno en el espacio de Hilbert. A su vez la propiedad de representar la información de cada observación de otra dimensión es clave para estimar una regla de clasificación de no lineal que disminuya el grado de error de clasificación. De la expresión de optimización dual (2.49) notamos que la información de las observaciones tiene una única intervención como un producto escalar de ellos.

$$F_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i{}^t x_j)$$

El procedimiento siguiente sería realizar una reemplazo en el producto escalar por una función que brinde la misma información pero pueda ser extendida a una espacio de

información en más dimensiones que es el rol que cumple la función kernel. Por lo tanto podemos señalar la siguiente función de optimización dual con reemplazamiento kernel:

$$F_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j (\boldsymbol{\phi}(x_i)^T \boldsymbol{\phi}(x_j)) \qquad (2.54)$$

O su equivalente expresión:

$$F_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j k(x_i{}^t x_j) \qquad (2.55)$$

En la ecuación dual de LaGrange al remplazamos el producto interno por la función Kernel, entonces podemos señalar los mapas no lineales del SVM de la muestra de entrenamiento, es decir en un espacio de características con una dimensión mucho más alta que la dimensión de la data original vía la función de mapeo $\boldsymbol{\phi}$. A partir de la identificación de este espacio de características transformado, la estimación de los parámetros seleccionados puede mejorar en la precisión del modelo Máquina de Vectores de Soporte. Eligiendo la función Kernel de base radial RBF, existen dos parámetros secundario o hiperparámetros: C y δ. Estos parámetros secundarios son determinados de manera experimental al realizar el proceso de modelamiento de la data de aprendizaje. Tendremos en mente que durante el modelamiento de SVM Kernel el parámetro secundario C representa la ponderación para la flexibilidad y generalización del modelo. Por otra parte δ es parte de la función kernel y su rol es de escalar la información de las observaciones sobre el espacio de la transformación kernel.Una alternativa para encontrar los mejores C y δ cuando empleamos la función RBF kernel es el de usar el conjunto de combinaciones de manera sistemática (Thegridsearchapproach, Hsu, Chang y Lin 2003).

II.1.5.5Función Kernel

A diferencia de los modelos lineales paramétricos que emplean la data aprendizaje solo en la fase inicial de modelamiento para estimar sus parámetros y realizar pronósticos, sin embargo existen otra clase de técnicas denominadas de Reconocimiento de Patrones en las cuales algunos un subconjunto de los datos de aprendizaje son almacenados y empleados en la realización de pronósticos son métodos como el modelo de densidad de probabilidad de Parzen y el modelo de Vecinos Cercanos.

Varios modelos lineales paramétricos pueden ser descritos en el contexto de un problema de optimización en una representación dual del problema, en el cual los pronósticos se basan en una combinación lineal de una función kernel en las observaciones de la data de aprendizaje. Así también, para modelos no lineales se basan sobre un Mapa del Espacio de Características no lineales de la transformación $\phi(x)$.

La función kernel es dada por:

$$K(x, \grave{x}) = \phi(x)^T \phi(x) (2.56)$$

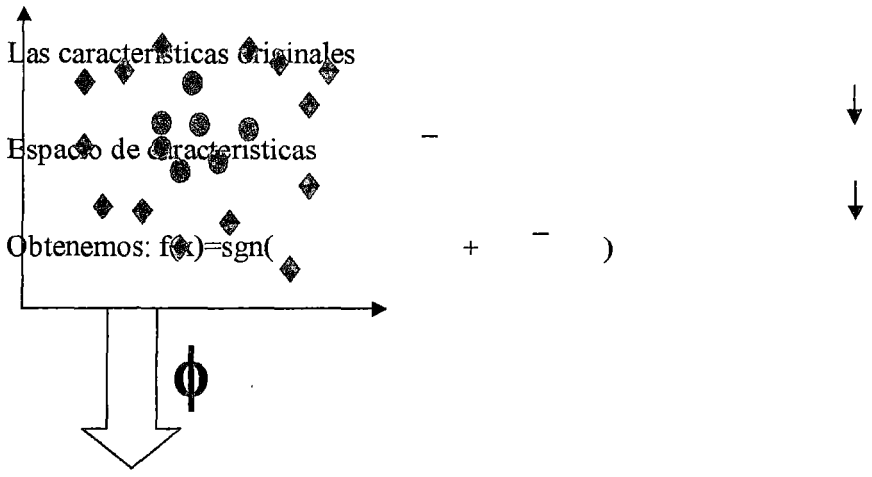
El concepto de kernel como producto interno del espacio de características, permite construir funciones interesantes complementarias a varios algoritmos. La idea general es que, si hemos formulado un algoritmo que recibe como único input el productos escalar de las observaciones tengamos la potestad de elegir una función kernel adecuada. Un ejemplo de aplicación esta substitución en los Componentes Principales No Lineales, otro ejemplo de la aplicar la substitución kernel es sobre la técnica de vecinos cercanos aplicando la función kernel discriminante.

Una representación gráfica de la funcionalidad de la substitución kernel se muestra en el gráfico Nro. 12:

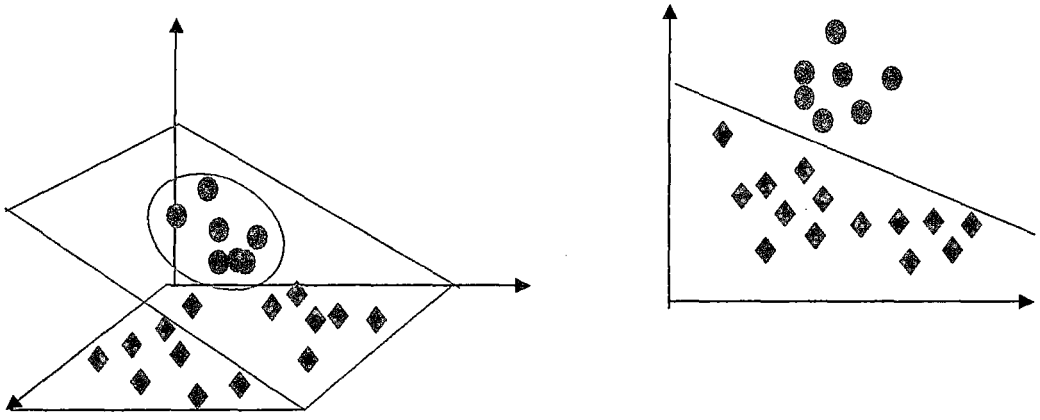
Sea la funciónuin ϕ parte de la función Kernel que mapea el espacio de características sigente: $\phi=R^d \rightarrow H$

Gráfico N° 12

Información de observaciones originales



Espacio de caracteristicas transformada



Considerando la expresión de óptimización (2.54):

Realizando la transformación kernel K, sobre el producto interno para el caso de poblaciones no separables linealmente.

$$F_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{\phi}(x_i)^T \boldsymbol{\phi}(x_j))$$

Y desde que $K(x, \dot{x}) = \phi(x)^T \phi(x)$no es necesario especificar la función explicita $\boldsymbol{\phi}$.

II.2 Operacionalización de las variables

Haciendo uso de la data acerca de operaciones de crédito, se busca evaluar el desempeño de la metodología SVM comparando los índices de diagnóstico que son derivados de la **matriz de confusión** que se mostrara a continuación.

Es imprescindible conocer detalladamente la exactitud de las distintas pruebas diagnósticas, es decir, su capacidad para clasificar correctamente a los clientes, empresas en categorías o estados en relación con el riesgo (típicamente dos: estar o no estar en default, respuesta positiva o negativa a los pagos del crédito otorgado).
La capacidad del modelo para representar confiablemente el sistema real, se relaciona esencialmente con la precisión.

No existe un modelo clasificador mejor que otro de manera general; para cada problema nuevo es necesario determinar con cuál se pueden obtener mejores resultados, y es por esto que han surgido varias medidas para evaluar la clasificación y comparar los modelos empleados para un problema determinado. Las medidas más conocidas para evaluar la clasificación están basadas en la matriz de confusión que se obtiene cuando se prueba el clasificador en un conjunto de datos que no intervienen en el entrenamiento.

Una vez obtenido el modelo predictivo de la probabilidad de default mediante el modelo usado en la metodología Scoring o en la metodología Rating, se procede a someterlo a una prueba de eficiencia.

II.2.1 Definición deMatriz de confusión

También es llamada tabla de contingencia. Representa la clasificación de las instancias clasificadas correcta o incorrectamente con respecto a los verdaderos valores y los valores pronosticados del modelo empleado. El número de instancias clasificadas correctamente es la suma de los números en la diagonal de la matriz; los demás están clasificados incorrectamente.

A partir de una matriz de confusión se deducen los índices relativos a la exactitud de la clasificación.

PRUEBA DIAGNOSTICA				
MATRIZ DE CONFUSIÓN		REAL		TOTAL
		BUENO	MALO	
PRONÓSTICO	BUENO	Verdadero Positivo (VP)	Falso Positivo (FP)	**VP + FP**
	MALO	Falso Negativo (FN)	Verdadero Negativo (VN)	**FN + VN**
TOTAL		**VP + FN**	**VN + FP**	**N**

II.2.2Indicadores de la matriz de confusión

Generalmente, la exactitud diagnóstica se expresa como sensibilidad y especificidad diagnósticas. Cuando se utiliza una prueba dicotómica (una cuyos resultados se puedan interpretar directamente como positivos o negativos).

Sensibilidad

Es la probabilidad de que una medida clasifique correctamente a un cliente o empresa que no está en default (cliente bueno). La sensibilidad es la capacidad del test para detectar o identificar a losclientes malos.

Es decir es la proporción de sujetos que presentan la característica estudiada y son clasificados correctamente por la prueba. Razón por la que también es denominada fracción de verdaderos positivos (FVP)

$$\textbf{Sensibilidad} = \frac{VP}{VP + FN}$$

Especificidad

Representa la probabilidad de que una medida clasifique correctamente a un cliente malo. Es decir, la proporción de personas que no tienen la característica estudiada y son clasificados correctamente por dicha prueba.

Es igual al resultado de restar a uno la fracción de falsos negativos (FFN).

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

$$= 1 - FFP \text{ (Fracción de falsos positivos)}$$

Un tratamiento estadístico correcto de cantidades como las calculadas por el método descrito en la tabla exigiría incluir medidas de su precisión como estimadores, y, mejor aún, utilizarlas para construir intervalos de confianza para los verdaderos valores de sensibilidad y especificidad.

Precisión

Proporción de elementos que realmente tienen clase x de entre todos los elementos que se han clasificado dentro de la clase x. En la matriz de confusión es el elemento diagonal dividido por la suma de la columna en la que estamos.

$$\textbf{Precisión} = \frac{VP}{VP + FP}$$

Otros indicadores

$$\text{Exactitud} = \frac{VP + VN}{N}$$

$$\textbf{Ratio Misclasificacion o Error medio} = \frac{FN + FP}{N}$$

$$FFN = \frac{FN}{VP + FN} = \textbf{Fracción Falsos Negativos}$$

$$FFP = \frac{FP}{VN + FP} = \textbf{Fracción Falsos Positivos}$$

Otra forma de evaluar el rendimiento de un clasificador es por las curvas ROC.

II.2.3 La Curva Roc (Receiver Operating Characteristic)

Este procedimiento es un método útil para evaluar la realización de esquemas de clasificación en los que exista una variable con dos categorías por las que se clasifiquen los sujetos.

Como consecuencia de aplicar un método de predicción (modelo) para determinar el score o rating de la probabilidad de default se producirán falsos positivos (se predice el éxito, pero realmente no lo obtuvieron), y falsos negativos (se predice como no éxito cuando realmente si lo obtuvieron), además de aciertos en uno y otro sentido.

Las curvas ROC empíricas se construyen a partir de los valores que se obtienen para la sensibilidad y la especificidad usando los distintos valores de P_o (puntos de corte) que se definan. En el eje de las abscisas se sitúa la probabilidad de un falso positivo (complemento de la especificidad); en el eje de las ordenadas, la probabilidad de declarar a un verdadero positivo (sensibilidad). Esto se hace para cada punto de corte que se escoja de 0 a 1.

Si se eligen m+1 puntos de corte en ese intervalo. Sean $P_0, P_1, \ldots, P_m$. Los puntos puede ser cualesquiera con la restricción: $P_0=0$; $P_m=1$.

Y ordenados de menor a mayor. Si m=10, entonces los valores serían: $P_0=0$; $P_1=0.1$; $P_2=0.2;\ldots$; $P_{10}=1$. Para cada punto se tiene una configuración tabular y por lo tanto una estimación para la sensibilidad y de la especificidad. Llamemos A_1 y B_i, respectivamente a las estimaciones que corresponden al punto P_i, puesto que $P_0=0$ y $P_m=1$, se tendrá:

$A_0 = 0$; $B_0 = 1$

$A_m = 1$; $B_m = 0$

La curva ROC se construye ubicando en el plano los m+1 puntos $(1- B_i, A_i)$. Cuanto más alejada del eje de abscisas esté la curva que se genera uniendo estos puntos, más

eficiente resulta la función para los efectos de la predicción. Precisamente el área bajo la curva ROC da una medida de la capacidad predictora global de la función, ya que,cuanto mayor sea esa área (más próxima sea al máximo 1), mayor capacidad predictiva tendrá la función. Asimismo, el área bajo la curva ROC es una vía para comparar diferentes funciones predictivas. Si una de ellas está por encima de la otra a lo largo del intervalo (0,1), no quedaría dudas de que el primer modelo es más eficiente que el segundo. Si ambas curvas s e interceptan, entonces debemos definir un indicador que permitirá decidir.

Este indicador gobal de la eficiencia de la prueba (el área bajo la curva ROC empírica) se mueve necesariamente entre las cotas 0 y 1, y viene dado por la fórmula:

$$A = \frac{1}{2}\sum_{i=0}^{m-1}(B_i - B_{i+1})(A_i + A_{i+1})$$

A fin de probar que el modelo empleado en la metodología de Scoring o la Metodología de Rating es eficiente se requiere usar ésta función para predecir las condiciones que deseamos usar colocando como **Y=1**.

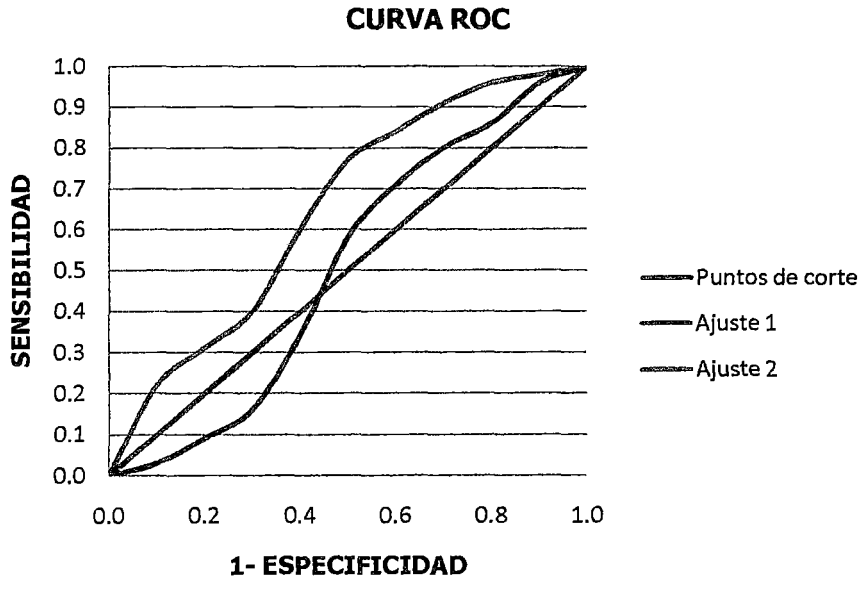
Al llevar a cabo dos ajustes (usando un conjunto de variables explicativas) se requieren evaluar cuál de ellas es más eficiente. Suponiendo que se tomaron en cuenta los m=11 puntos de corte sugeridos. Cada punto dará a lugar a una pareja de estimaciones A y B. Veamos la tabla:

Puntos de Corte	Ajuste 1		Ajuste 2	
	Sensibilidad	Especificidad	Sensibilidad	Especificidad
0.0	0.00	1.00	0.00	1.00
0.1	0.03	0.94	0.22	0.85
0.2	0.09	0.90	0.31	0.83
0.3	0.16	0.85	0.40	0.82
0.4	0.34	0.70	0.60	0.80
0.5	0.58	0.69	0.77	0.68

0.6	0.71	0.66	0.84	0.49
0.7	0.80	0.60	0.91	0.35
0.8	0.86	0.59	0.96	0.21
0.9	0.96	0.32	0.98	0.06
1.0	1.00	0.00	1.00	1.00

Si al realizar los cálculos correspondientes de A, y los valores por ejemplo serían 0.68 y 0.73 para el Ajuste 1 y el Ajuste 2 respectivamente, se deduciría que el segundo ajuste tendría mayor capacidad predictiva que el primero.

Gráfico N° 13



La figura muestra dos curvas ROC, teniendo mayor capacidad el Ajuste 2

Dentro de las consideraciones debemos tener en cuenta lo siguiente:

Los Datos:

Las variables de contraste son cuantitativas. Las variables de contraste suelen estar constituidas por probabilidades, resultantes de un análisis discriminante o de una regresión logística, o bien compuestas por puntuaciones atribuidas en una escala

arbitraria que indican el "grado de convicción" que tiene un evaluador de que el sujeto pueda pertenecer a una u otra categoría. La variable de estado puede ser de cualquier tipo e indicar la categoría real a la que pertenece un sujeto. El valor de la variable de estado indica la categoría que se debe considerar positiva.

Los Supuestos

Se considera que los números ascendentes de la escala del evaluador representan la creciente convicción de que el sujeto pertenece a una categoría. Por el contrario, los números descendentes representan la creciente convicción de que el sujeto pertenece a la otra categoría. El decisor deberá elegir qué dirección es positiva. También se considera que se conoce la categoría real a la que pertenece el sujeto.

Área bajo la curva

La mayor exactitud diagnóstica de una prueba se traduce en un desplazamiento "hacia arriba y a la izquierda" de la curva ROC. Esto sugiere que el área bajo la curva ROC se puede emplear como un índice conveniente de la exactitud global de la prueba: la exactitud máxima correspondería a un valor de 1 y la mínima a uno de 0.5 (si fuera menor de 0.5 debería invertirse el criterio de positividad de la prueba).

En términos probabilísticos, si X_B y X_M son las dos variables aleatorias que representan los valores de la prueba en las poblaciones PAGO E IMPAGO, respectivamente, puede probarse que el área bajo la curva de la "verdadera" curva ROC (intuitivamente, aquella que obtendríamos si el tamaño de la muestra fuera infinito y la escala de medida continua) es precisamente $\theta = Pr(X_B > X_M)$, es decir, la probabilidad de que, si se eligen al azar un cliente bueno y otro malo, sea mayor el valor de la prueba en aquél que en éste.

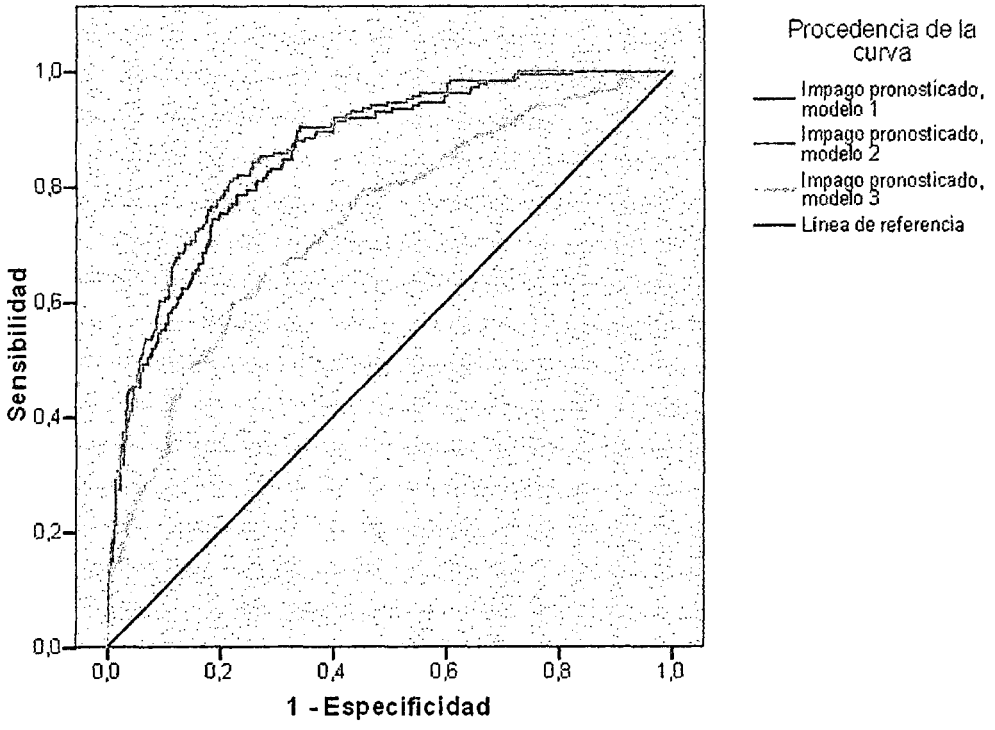
El área bajo la curva permite clasificar el orden como sigue:

CLASIFICACIÓN (MODELO)	RANKING
REGULAR	0.50 – 0.75
BUENO	0.75 – 0.92
MUY BUENO	0.92 – 0.97
EXCELENTE	0.97 – 1.00

Si se quisiera comparar la efectividad de tres modelos, se tendría el resultado de las curvas ROC, determinando el valor del área bajo la curva.

Gráfico N° 14

CURVA ROC



Área bajo la curva

Variables resultado de contraste	Área
Impago pronosticado, modelo 1	0.856
Impago pronosticado, modelo 2	0.870
Impago pronosticado, modelo 3	0.735

Para efectos de los datos obtenidos se elegiría el segundo modelo el cual el pronóstico de los datos es mejor comparado a los demás modelos usados.

III.Metodología de Investigación

III.1Tipo de estudio

Nuestro estudio es de tipo exploratorio e inferencial en el contexto de la evaluación del Riesgo de Crédito en nuestra data de estudio, debido a que el objetivo de nuestra investigación es conocer el riesgo relativo asociado a cada variable (capacidad predictiva por variable) y luego es comprobar la capacidad predictiva de los modelos en cuanto al cumplimiento de pago de los créditos aprobados por cada modelo predictivo.

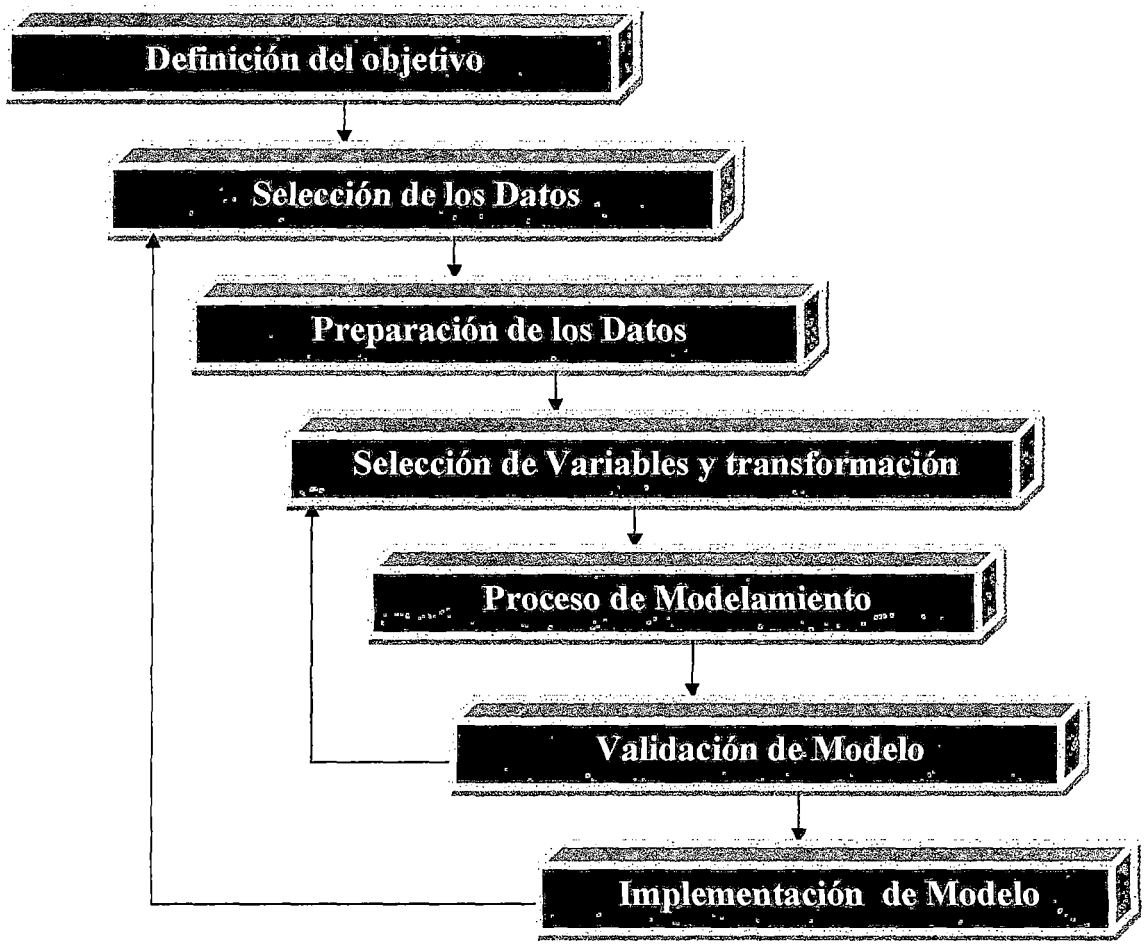
Como tema principal en parte del desarrollo del estudio detallamos las diferencias metodológicas de emplear un modelo paramétrico de Regresión Logística y un modelo no paramétrico como es la Máquina de Vectores de Soporte.

Empleamos la metodología que sugiere cualquier análisis del Descubrimiento del conocimiento en base de Datos (KnowledgeDiscovering in Databases) que implica a su vez una metodología empleada por SAS (StatisticalAnalysisSystem), la metodología **SEMMA(Sample** – Muestra, **Explore** – Exploración, **Modify** – Modificación, **Model**– Modelamiento, **Assess** – Validación)por sus siglas en inglés, que es mundialmente

conocida para el desarrollo de Proyectos de minería de datos y el cual detallaremos a continuación.

Gráfico N° 15

Pasos para un proyecto exitoso de modelamiento[26]



[26] [23] OlviaParrRud pág. 5

III.1.1 Definición de objetivo de modelamiento

El desarrollo y conocimiento de técnicas de modelos predictivos se han extendido en las industrias, en el mando gerencial, en los tomadores de decisiones. Pero lo que sigue siendo una fase clave de éxito del desarrollo de un proyecto de minería de datos es la definición del variable objetivo, específicamente en el ámbito de riesgo crédito se han desarrollado algunas técnicas para identificar casos que presenten un comportamiento de incumplimiento de pago bien marcado como lo es la técnica de Curva de Maduración de los crédito como sugiere Siddiqi[27].

III.1.2 Selección de Muestra / Sample

Fase de identificación de la data inicial que sea confiable eintegrable de las diferentes fuentes de datos para ser analizada. En este contexto lidiamos con enormes volúmenes de datos, y nuestra tarea en esta fase es encontrar aquella fracción del volumen total de datos que contenga información suficiente que respalde un modelo predictivo, de un tamaño suficiente para que los resultados obtenidos sean generalizables y puedan servir para la toma de decisiones.

Consideremos que el procedimiento de muestreo será guiado por un perfil objetivo (identificado por unavariable objetivo). El muestreo es de suma importancia para el estudio debido a que en el modelamiento estadístico identificará los patrones sobre la muestra seleccionada.

Entonces para la definición de la muestra consideremos en un proyecto de Minería de Datos consideremos los siguientes 2 conceptos de:

Probabilidad A priori: Es el porcentaje de clientes que han incurrido en incumplimiento en la cartera muestra de clientes elegidas para el modelamiento.

[27] [24] Naeemsiddiqi pág. 34

Probabilidad A posteriori: Es la probabilidad obtenida de la conjunción de la probabilidad a priori de la muestra y la probabilidad estima del modelo predictivo.

III.1.3 Exploración/Explore

Fase de exploración estadística de la data, de descubrimiento visual, descubriendo relaciones y tendencias tanto esperadas como inesperada. A la vez el conocimiento de ambos tipos de relaciones nos permite una clara comprensión de la data así como el inicial Brainstorming de las potenciales variables que podrían conformar nuestro modelo predictivo.

III.1.4 Modificación y transformación/Modify

Fase de alteración de la data original, con la finalidad de enriquecer la información que se puede obtener variables transformación como pueden ser estandarizaciones, aplicación de funciones matemáticas o por otro lado la creación de variables analíticas como pueden ser promedio trimestrales, tendencias, desviaciones estándar variables indicadoras basados en experiencia del negocio.

III.1.5 Modelamiento/Model

Una vez validados los supuestos del modelo que asegure que los resultados pueden ser generalizables. La fase de la revisión de modelos estadísticos, se orienta a la búsqueda de aquellas variables cuya combinación consiga un pronóstico confiable.

III.1.6 Evaluación de los pronósticos/Asses

Fase de revisión de la confiabilidad de los resultados logrados, esto puede ser con realizando pruebas con datos recientes y manteniendo los indicadores de poder predictivo de la fase de modelamiento.[28]

IV.Desarrollo de estudio

IV.1 Ámbito de desarrollo de los Modelos Predictivos

Con el objetivo de reducir la tasa de morosidad de la cartera de clientes, evitando las perdidas por otorgar préstamos a clientes con perfiles riesgosos, dotar de una herramienta de rápida evaluación de los postulantes a un crédito. Además de tener un mayor conocimiento de las variables que determinen un comportamiento crediticio adecuado, se da inicio como una medida de gestión del riesgo a un proyecto de **creditscoring.**

IV.2Variable objetivo de modelo predictivo

En nuestro caso en particular la variable objetivo ya está definida en la data de estudio, es la variable incumplimiento de pago (variable que en nuestra población solo toma 2 valores):

1: El crédito cae en incumplimiento de pago

0: El crédito es pagado en las cuotas acordadas

[28][18] Randall Matignonpág. ix

IV.3Población de estudio

Con la data del profesor Hoffman[29], se procede a realizar análisis en el estudio de la base de datos histórica de los clientes que tuvieron un préstamo. En la cual a grandes rasgos detalla variables como: variables financieras personales, variables del préstamo, historial del cliente e información demográfica, además de una variable que nos indica si el cliente cayó finalmente en el incumplimiento del pago o cumplimiento del pago del préstamo.La variable de evaluación es el incumplimiento de pago de los créditos otorgados a personas naturales, el incumplimiento de pago superior a 30 días de la cuota o cuotas acordadas y la base de datos de análisis consta de 1000 clientes registrados.

Las bases de datos de comportamiento de pago en Perú son manejados confidencialmente por las empresas financieras, por ello en el presente estudio empleamos una fuente de datos secundaria, obtenida de un estudio de modelización de riesgo de crédito realizado en Alemania y aplicamos una metodología estadística que puede ser fácilmente extendida a casos de nuestra realidad.

[29]Profesor Dr. Hans Hofmann Institut f "urStatistik und" OkonometrieUniversit "en Hamburgo

FB Wirtschaftswissenschaften Von-Melle-Park 5 2000 Hamburg 13

IV.4 Diseño muestral

Para que las técnicas estadísticas Regresión Logística o algoritmos de Máquinas de aprendizaje puedan detectar perfiles incumplimiento de pago. Se ha propuesto aplicar un balance de la muestra como sugiere *Naeem Siddiqi*[30].

"Hay varias maneras de dividir la muestra de desarrollo del modelo scoring. Normalmente se mantiene el 70% o 80% de los casos de aprendizaje o como desarrollo del modelo. Mientras 30% u 20% restante es mantenido independientemente para una prueba de validación del modelo predictivo scoring."

Naeem Siddiqi

En nuestro estudio realizaremos una división aleatoria y conveniente de la data del 70% y 30%, de esta manera de los 1000 registros, 700fueron divididos en una data de aprendizaje y la muestra restante en datos de validación. La proporción de ambos conjunto de datos será de 7 a 3, empleando así 700 registros usados para el entrenamiento del modelo, mientras 300 serán usados para la fase de validación de los pronósticos.

IV.5Construcción de la matriz de datos

A continuación se hará la descripción de las características recopiladas, con esta información crearemos nuestra matriz de modelamiento:

- **Información crediticia del cliente en la empresa.**

[30][24] Naeem Siddiqi pág.63

Número y tipo de cuentas que tiene el cliente, saldo en sus cuentas de ahorro, otras deudas y garantías; además del propósito, monto y cuotas del crédito que será evaluado en el estudio de riesgo de crédito.

- **Nivel socio económico de la persona que ha realizado el préstamo.**

  Fuentes de ingreso de la persona, tipo de trabajo, situación laboral, antigüedad en el trabajo y las propiedades que posee (vehículos y vivienda).

- **Información demográfica de la persona que ha realizado la solicitud del crédito.**

  Estado de civil de la persona, edad, género y número de personas dependientes.

Adicionalmente la base de datos de análisis cuenta con 1000 casos para ser evaluados, se considera que el evento objetivo es definido por la variable incumplimiento depago, la cual toma dos valores posibles de la variable TARGET son:

1: El crédito cae en incumplimiento de pago

0: El crédito es pagado en las cuotas acordadas

Con ambas técnicas de minería de datos, Regresión Logística y Máquinas de Vectores de Soporte se pretende determinar una regla de clasificación entre la población que recae en el incumplimiento de pago y el la población realiza el pago del crédito en las cuotas acordadas alcanzando con la mayor precisión.

	DESCRIPCIÓN	NOMBRE DE VARIABLE	TIPO
Historial crediticio	Estado de cuenta	CHK_ACCT	Cualitativa
	Historial crediticio	HISTORIA_CRED	Cualitativa
	Balance promedio en cuenta de ahorro	BALANC_ACCT	Cualitativa
	Tiempo en meses	DURACION_CRED	Numérica
	Monto de crédito	MONTO_CRED	Numérica
	Nro. de créditos existentes en el banco	NUM_CREDITS	Numérica
	Socio de crédito	CO-APPLICANT	Binaria
	Persona Aval	GUARANTIA	Binaria
	Tiene otro plan de crédito	OTROS_PLAN	Binaria
	Propósito del crédito Auto nuevo	PROP_AUT_NUEVO	Binaria
	Propósito del crédito Auto usado	PROP_AUT_USADO	Binaria
	Propósito del crédito Muebles	PROP_MUEBLES	Binaria
	Propósito del crédito Radio/Tv	PROP_RADIO_TV	Binaria
	Propósito del crédito Educación	PROP_EDUCACION	Binaria
	Propósito del crédito Refinanciamiento	PROP_REFINAC	Binaria
Demográfica	Edad	EDAD	Numérica
	Nro. Dependientes	NUM_DEPENDENTS	Numérica
	Divorciado	SIT_CIVIL_DIV	Binaria
	Soltero	SIT_CIVIL_SOLT	Binaria
	Casado o conviviente	SIT_CIVIL_CASADO	Binaria
Nivel socioeconómico	Antigüedad en su empleo	ANTIG_EMPLEO	Cualitativa
	Antigüedad en su residencia actual	ANTIG_RES_ACTUAL	Cualitativa
	Nivel educativo	NIVEL_EDUCATIVO	Cualitativa
	Porcentaje de ingreso disponible	TASA_INGR_DISPONIBLE	Numérica
	Vehículo propio	VEHICULO_PROP	Binaria
	Empresa propia	EMPRESA_PROPIA	Binaria
	Estado real de la propiedad	ESTADO_PROPIEDAD	Binaria
	No es propietario de residencia	NO_PROPIETARIO_RESID	Binaria
	Vivienda alquilada	RESID_ALQUILADA	Binaria
	Propietario de su residencia	RESID_PROPIETARIO	Binaria
Objetivo	Target	TARGET	Binaria

IV.6Esquema Experimental

El procedimiento experimental es esquematizado en los diagramas 1 y 2, según las técnicas empleadas. Para el desarrollo de una técnica paramétrica como es la Regresión Logística y una técnica computacional como lo es Máquina de Vectores de Soporte, podemos señalar las siguientes observaciones:

Primero, la revisión de supuestos es un paso imprescindible, detectar problemas de multicolinealidad de las variables, detección de casos atípicos o influyentes que puedan sesgar los patrones detectados por los modelos distorsionando el perfil de riesgo crediticio por casos atípicos.

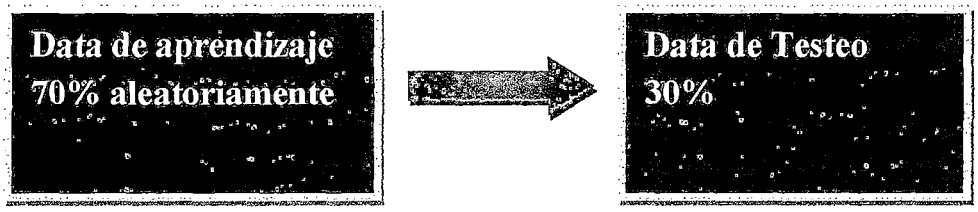
Segundo, una revisión descriptiva de las variables contra la variable target (incumplimiento de pago) es un paso crucial en ambos diagramas, pues es la fase de compresión de los datos. Podemos denominar este paso como el soporte analítico de los modelos predictivos, pues nos brindan una explicación del comportamiento de las variables predictivas previa a los resultados de los modelos.

Tercero, ambos diagrama de modelamiento emplean el mismo particionamiento de la data con la finalidad de poder generalizar el patrón que detectan en los datos y sean evaluados los pronósticos de ambas técnicas sobre el mismo contexto.

Finalmente, para poder evaluar los resultados predictivos separamos una data de TEST, sobre los cuales serán medidos indicadores de pronóstico como la Curva ROC y matriz de confusión.Aquella técnica que tenga un mayor poder de predicción para el contexto de la evaluación del riesgo de crédito mostrará mejores indicadores.

Diagrama de modelamiento de la técnica Regresión Logística

Partición de la data según las fases



- ***Validación de supuestos***
  - Se identifica la existencia de problemas de Multicolinealidad.
- ***Exploración de variables***
  - Se realiza una evaluación descriptiva de las variables.
- ***Selección de variables***
  - Se realiza una selección de un subconjunto de variables que sean significativa para nuestro modelo empleando el criterio de AIC con la regresión logística.
- ***Partición de la muestra***
  - Se realiza un particionamiento del 80% de datos para realizar la regresión Logística y el 20% como muestra de validación para evitar el sobre ajuste del modelo y que los resultados sean generalizables.

- ***Revisión de curva ROC***
  - Verificación de los indicadores de sensibilidad y especificidad del poder predictivo del modelo sobre una base de datos de Test

Diagrama 1

Diagrama de modelamiento de la técnica Máquina de Vectores de Soporte

Partición de la data según las fases

- ***Validación de supuestos***
  - Se identifica la existencia de problemas de Multicolinealidad.
- ***Exploración de variables***
- ***Selección de variables***
  - Se realiza una selección de un subconjunto de variables que sean significativa de nuestro modelo empleando el criterio de AIC con la regresión logística.
- ***Validación del modelo***
  - Aplicación del método Cross Validation 10 veces, participamos la data con la finalidad de poder generalizar los resultados encontrados.
- ***Aplicación del Modelo Máquina de Vectores de Soporte***
  - Se realizara median una selección parámetros secundarios C y gama, que definen el modelo SVM elegido.
- ***Revisión de curva ROC***
  - Verificación de los indicadores de sensibilidad y especificidad depoder predictivo del modelo sobre una base de datos de Test.

Diagrama 2

V. Procesamiento de la data

V.1 Pre procesamiento de los dato

Las fases de validación de los datos, detección de valores atípicos, descripción de las variables predictivas y particionar la data en muestra de modelamiento y test son fases que ambas técnicas tiene en común. Por ellos designamos las siguientes secciones a realizar las fases de pre procesamiento de un procediendo de modelamiento predictivo (o de Minería de Datos).

V.1.1. Validación de los datos

El paso preliminar a todo estudio cuantitativo es la validación de las variables, en nuestro estudio la dataanálisis no presenta valores faltantes de ningún tipo como el caso de valores faltantes en las variables regresoras, valores faltantes en la variable de respuesta o valores faltantes en ambas. Por tal motivo no procederemos a realizar la imputación de valores.

Lo siguiente es una revisión descriptiva de las variables con la finalidad de detectar valores incoherentes de las variables.

Variables cuantitativas:

NOMBRE DE VARIABLE	Mínimo	Máximo	Promedio	DesvStd
DURACION_CRED (meses)	4	72	21	12
EDAD (años)	19	75	36	11
MONTO_CRED (marcos)	250	18,424	3,271	2821

Variables cualitativas y ordinales:

CHK_ACCT (marcos)
Estado de cuenta
0 : < 0 DM
1: 0 < ...< 200 DM
2 : => 200 DM
3: No tiene cuenta

HISTORIA_CRED
Historial crediticio
0: no tiene crédito
1: Todos los créditos en este banco pagados debidamente
2: Créditos existentes pagados debidamente hasta ahora
3: Retraso en el pago en el pasado
4: cuenta critica

BALANC_ACCT (marcos)
Balance promedio en cuenta de ahorro
0 : < 100 DM
1 : 100<= ... < 500 DM
2 : 500<= ... < 1000 DM
3 : =>1000 DM
4 : No tiene cuenta de ahorro

ANTIG_EMPLEO
Antigüedad en su empleo
0 : Desempleado
1: < 1 año
2 : 1 <= ... < 4 años
3 : 4 <=... < 7 años
4 : >= 7 años

ANTIG_RES_ACTUAL
Antigüedad en su residencia actual
1: <= 1 años
2:<...<=2 años
3:<...<=3 años
4:>= 4 años

NIVEL_EDUCATIVO
Nivel educativo
0 : Secundario Incompleta
1 : Secundaria completa
2 : Universidad Incompleta
3 : Universidad Completa

Variables Binarias:

En los casos de todas las variables binarias realizaremos la interpretación asignamos dos valores posibles:

1: La persona tiene la característica (SI)

0: La persona no tiene la característica (NO)

Las variables están representadas en ceros y unos o también denominadas variables dummies.

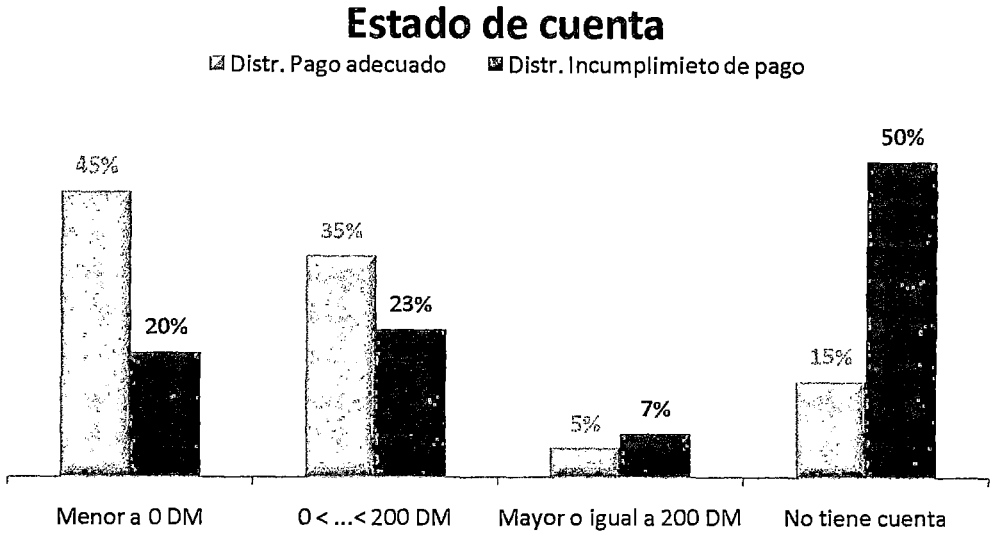
V.1.2. Conversión de variables cualitativas (nominales y ordinales) a Dummy

La conversión a Dummy busca estructurar una visión de la información clara sobre la matriz de datos y que el algoritmo predictivo identifique estas variables del tipo categórico y no identifique estas variables categóricas como variables cuantitativas empleando el nivel de referencia de la categoría.

Definiremos el nivel de referencia para las cualitativas, según descriptivamente este represente como el menor nivel de Riesgo Crediticioen la data de estudio completa para las variables: Estado de cuenta, Historial crediticio, Balance promedio en cuenta de ahorro, Antigüedad en su empleo, Antigüedad en su residencia actual y Nivel educativo.

En la data total se cuenta con 700 casos de incumplimiento de Pago y 300 con un pago adecuado de sus cuotas, emplearemos las distribuciones entre ambas poblaciones para identificar a los niveles de referencia para las variables cualitativas.

Variable: Estado de cuenta Gráfico N° 16



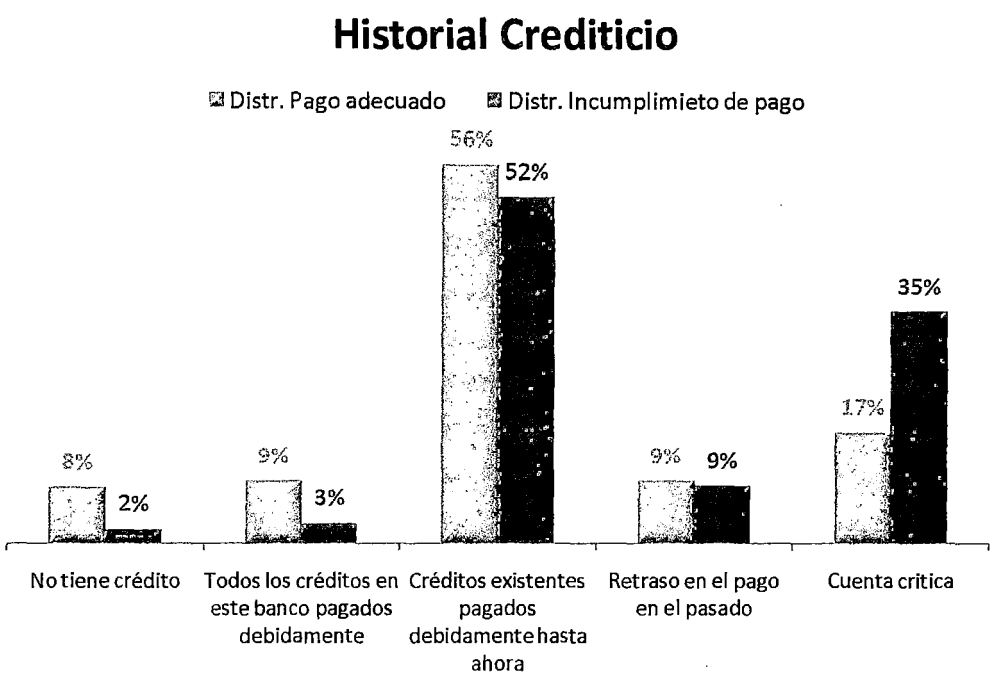
Para este caso el nivel de referencia es ***Menor a 0DM***, por baja concentración de incumplimiento de pago respecto a la distribución del pago adecuado del crédito.

Una vez conocido los niveles de referencia, procederemos a realizar la conversión de las variable *Menor a 0DM* en la matriz de datos como se señala y se mencionó en la sección II.1.2 (pág. 18).

Efecto Codificado			
CHK_ACCT	**Matriz de diseño**		
		Est_Cta	
Respuestas	**Est_Cta_1**	**_2**	**Est_Cta_3**
0:Menor a 0DM	0	0	0
1:0<...<200 DM	1	0	0
2:Mayor o igual a 200DM	0	1	0
3:No tiene cuenta	0	0	1

De manera similar se realizó el procedimiento para cada una de las variables cualitativas, considerando como nivel de referencia la categoría de menor riesgo relativo.

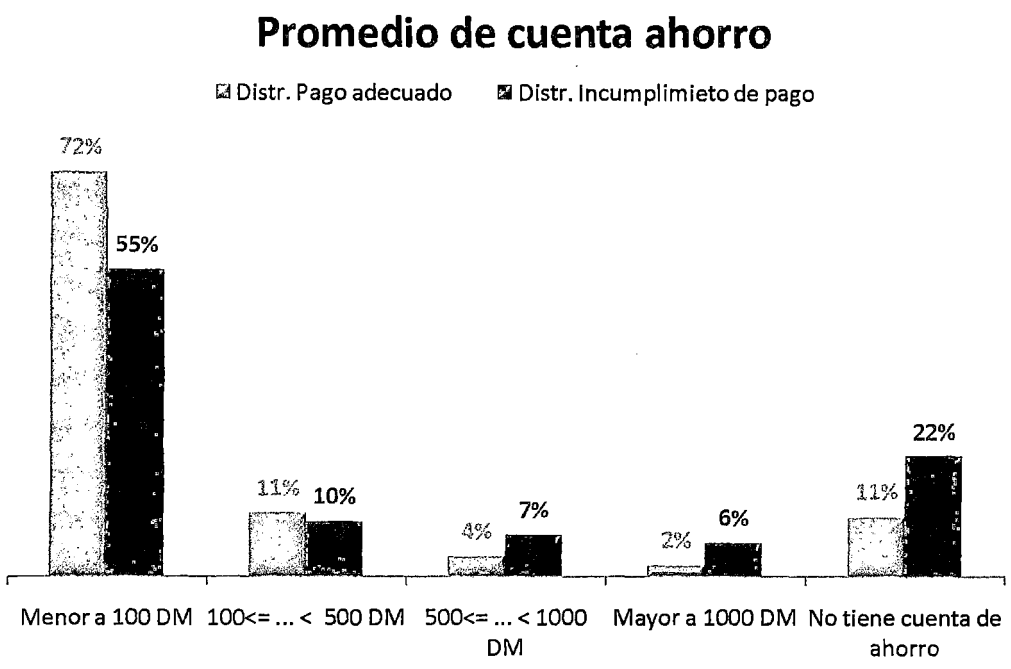
Variable: Historial crediticio Gráfico N° 17



Para este caso el nivel de referencia es ***No tiene crédito***, por mostrar menor nivel de riesgo relativo de incumplimiento de pago.

Efecto Codificado				
HISTORIA_CRED	**Matriz de diseño**			
Respuestas	**Hist_cred_1**	**Hist_cred_2**	**Hist_cred_3**	**Hist_cred_4**
0:No tiene crédito	0	0	0	0
1:Crédito solicitados pagados	1	0	0	0
2:Créditos vigentes pagados	0	1	0	0
3:Retrasó en pagos	0	0	1	0
4:Cuenta crítica	0	0	0	1

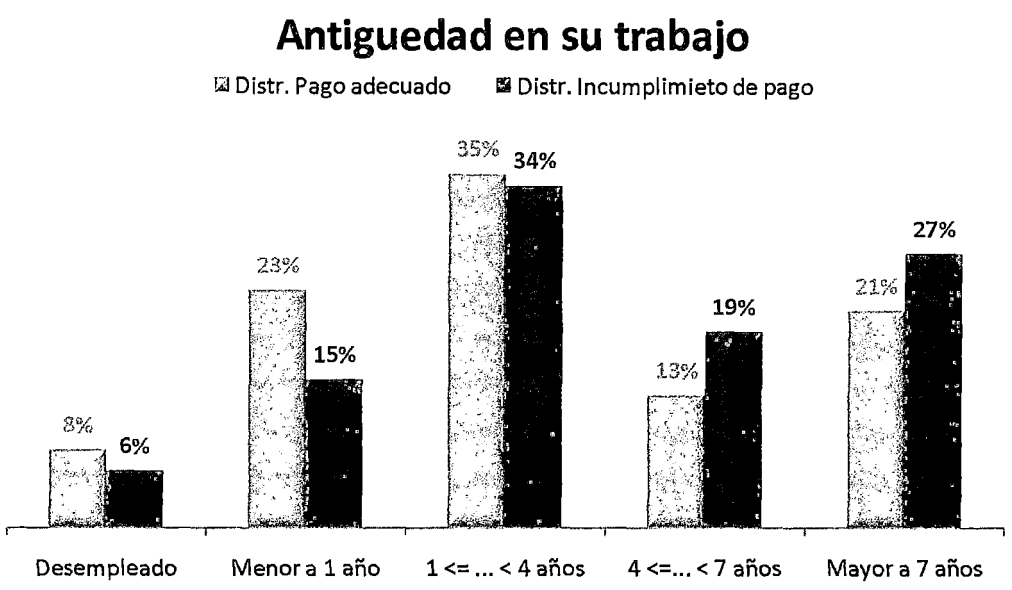
Variable: Balance promedio en cuenta de ahorro Gráfico N° 18



Para este caso el nivel de referencia es ***Menor a 100 DM***, por mostrar menor nivel de riesgo relativo de incumplimiento de pago.

Efecto Codificado				
BALANC_ACCT	**Matriz de diseño**			
Respuestas	**Pr_ct_aho_1**	**Pr_ct_aho_2**	**Pr_ct_aho_3**	**Pr_ct_aho_4**
0:Menor a 100DM	0	0	0	0
1:100<=..<500 DM	1	0	0	0
2:500<=...<=1000 DM	0	1	0	0
3:Mayor a 1000 DM	0	0	1	0
4:No tiene cuenta	0	0	0	1

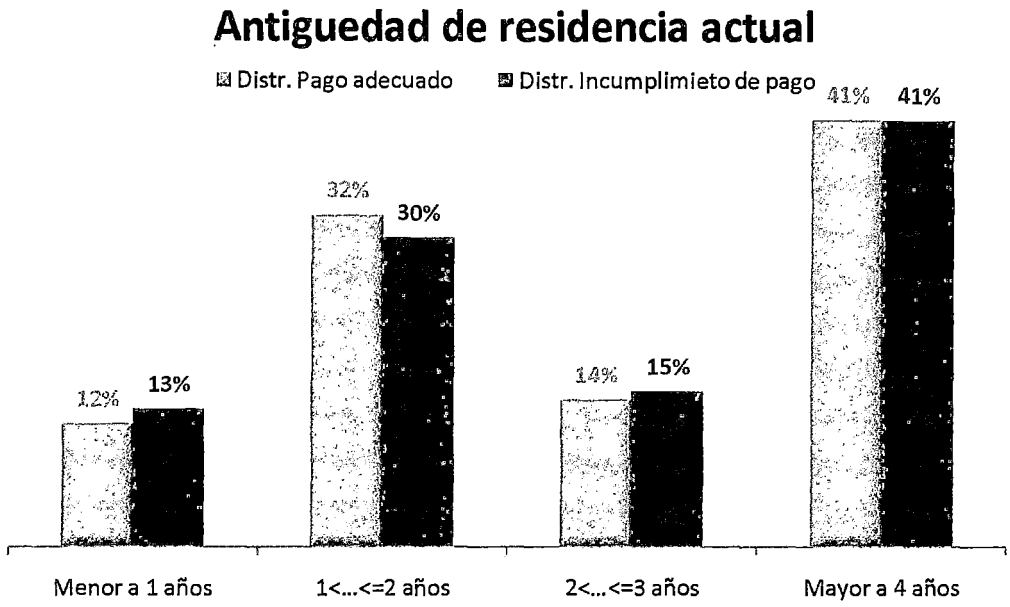
Variable: Antigüedad en su empleo Gráfico N° 19



Para este caso el nivel de referencia es ***Menor a 1 año***, por mostrar menor nivel de riesgo relativo de incumplimiento de pago.

Efecto Codificado				
ANTIG_EMPLEO	**Matriz de diseño**			
Respuestas	**Ant_trab_1**	**Ant_trab _2**	**Ant_trab _3**	**Ant_trab _4**
1:Menor a 1 año	0	0	0	0
2:1<=..<4 años	1	0	0	0
3:4<=... <=7 años	0	1	0	0
4:Mayor a 7 años	0	0	1	0
0:Desempleado	0	0	0	1

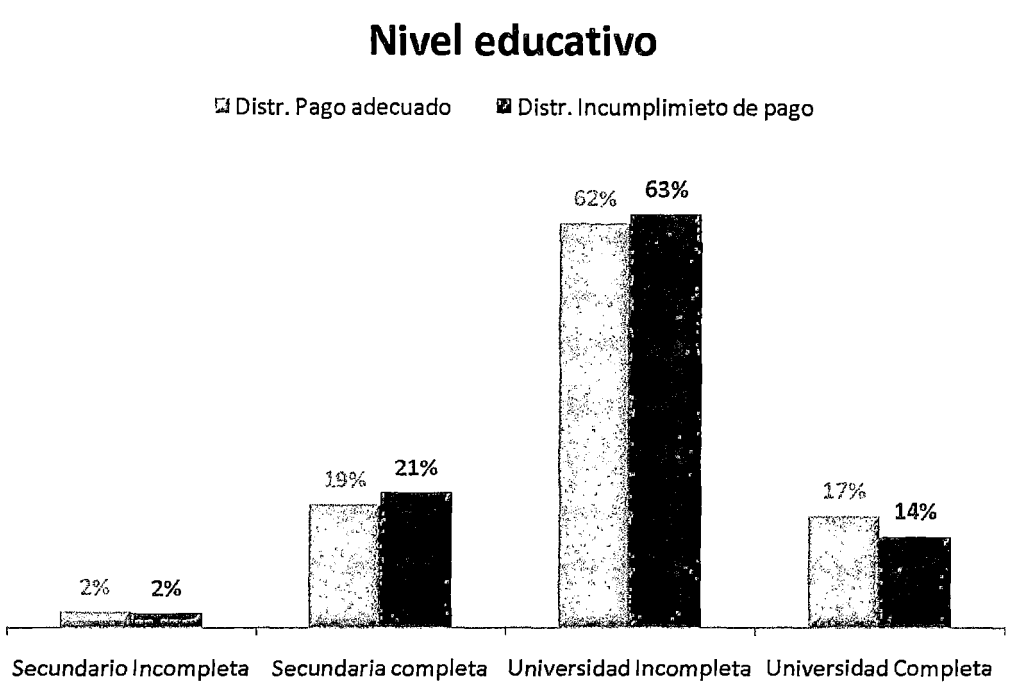
Variable: Antigüedad en su residencia actual Gráfico N° 20



Según la data histórica, para este caso el nivel de referencia es ***Entre 1 a 2 años***, por mostrar menor nivel de riesgo relativo de incumplimiento de pago.

Efecto Codificado			
ANTIG_RES_ACTUAL	**Matriz de diseño**		
		Ant_res	
Respuestas	**Ant_res_1**	**2**	**Ant_res_3**
2:1<..<=2 años	0	0	0
3:1<..<=3 años	1	0	0
4:Mayor a 4 años	0	1	0
1:Menor a 1 año	0	0	1

Variable: Nivel educativo Gráfico N° 21



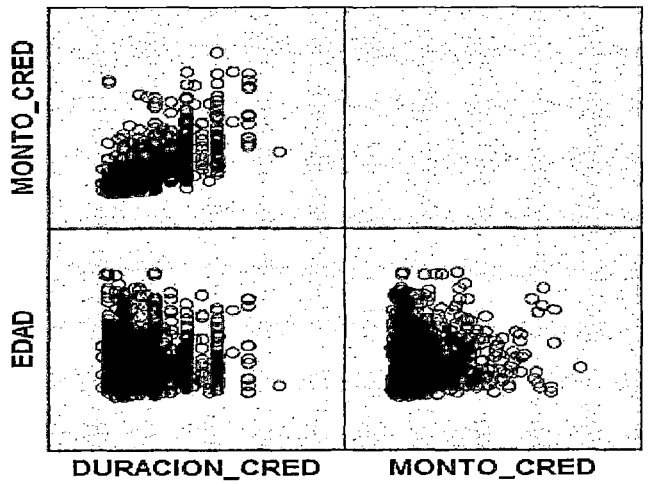
Para este caso el nivel de referencia es ***Universidad completa***, por mostrar menor nivel de riesgo relativo de incumplimiento de pago.

Efecto Codificado			
NIVEL_EDUCATIVO	**Matriz de diseño**		
Respuestas	**Niv_Ed_1**	**Niv_Ed_2**	**Niv_Ed_3**
3:Universidad completa	0	0	0
0:Secundaria incompleta	1	0	0
1:Secundaria completa	0	1	0
2:Universidad Incompleta	0	0	1

V.1.3. Revisión de los valores atípicos

Ante un procedimiento de detección de valores atípicos debemos considerar no ser muy estrictos ni exhaustivos en la detección. Tengamos como premisa principal en el desarrollo de esta tarea, velar por mantener las relaciones entre variables que serán la estructura principal de un modelo predictivo. Dado que una eliminación de gran magnitud o sin revisar el impacto en los coeficientes de correlación de las variables cuantitativas puede conducir a pérdida de información crítica.

Gráfico N° 15



		DURACION_CRED	MONTO_CRED
MONTO_CRED	Correlación de Pearson	**.625**	
	Sig. (bilateral)	*.000*	
EDAD	Correlación de Pearson	**-.036**	**.033**
	Sig. (bilateral)	*.254*	*.301*

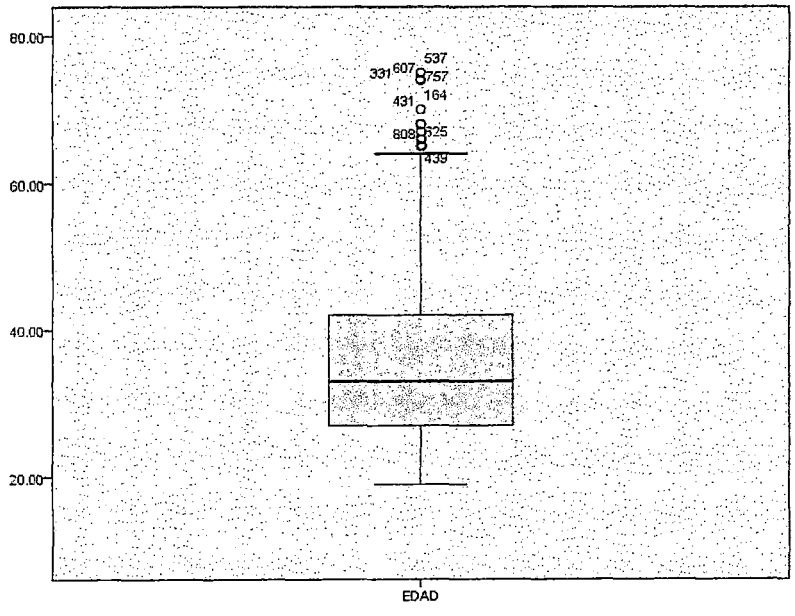
Realizaremos un procedimiento consensado de técnicas de detección de valores atípicos, con la finalidad de identificar a los valores atípicos por comportamiento univariados como por comportamiento multivariado.

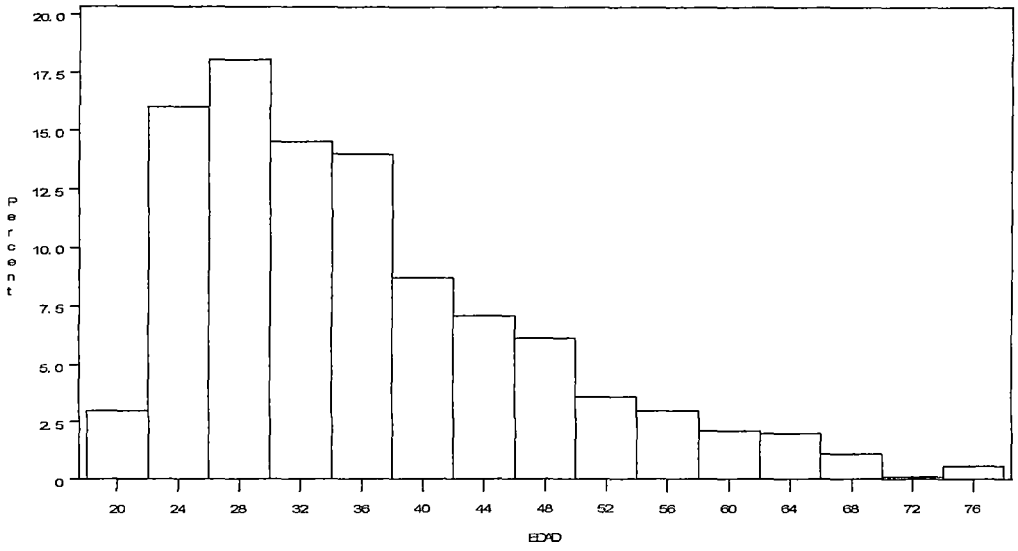
Ya que no es necesario suponer que las variables monto de crédito, edad y duración del crédito tengan una distribución normal. Aplicaremos la detección de casos ubicados fuera del rango intercuartílico, considerando que la información valiosa se encuentra dentro del rango. En la data de estudio tenemos solo tres variables cuantitativas que pasaremos a evaluar las variables son: Edad, Monto de crédito y Número de meses del crédito.

Variable: Edad

Gráfico N° 16

Promedio	35.5	Std. Dst.	11.4
Mediana	33.0	Varianza	129.4
Asimetría	1.0	C.V.	32.0
Kurtosis	0.6	Rango	56.0





Al aplicar la detección de valores outliers potenciales fuera del rango intercuartílico hemos encontrado la siguiente lista de valores de edades:

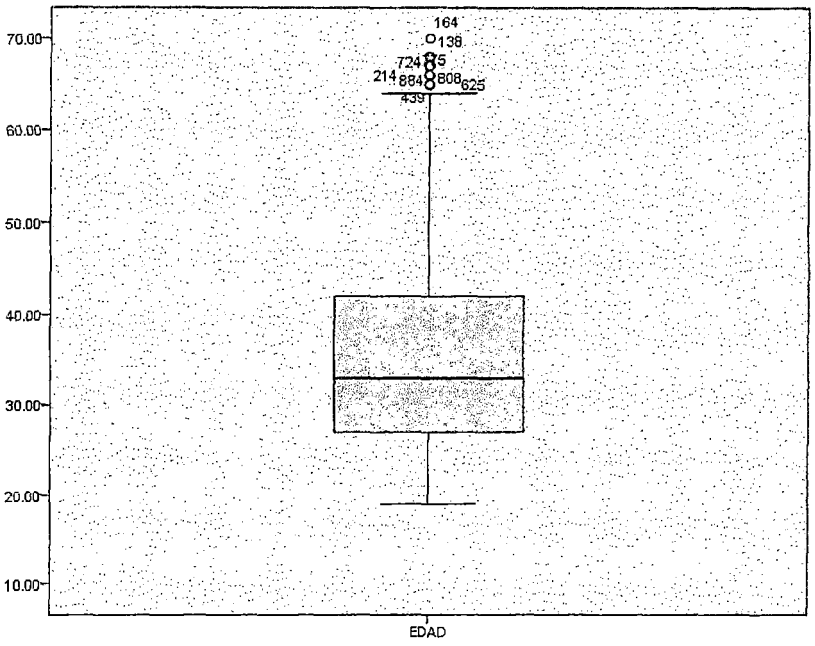
OBS	EDAD
187	74
331	75
431	74
537	75
607	74
757	74

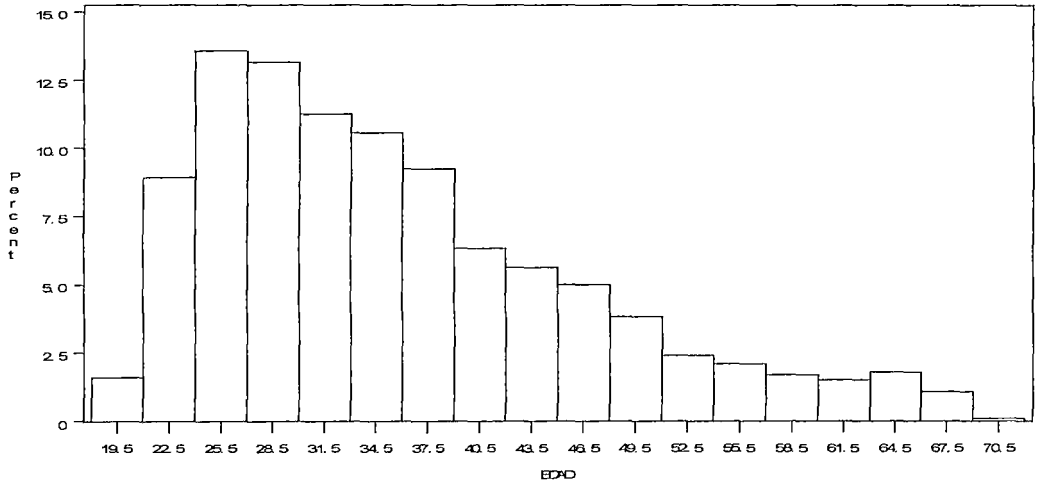
0.6% de valores atípicos potenciales.

Al retirar estos valores encontramos una mayor centralización de los valores de la variable edad.

Gráfico N° 17

Promedio	35.3	Std. Dst.	11.0
Mediana	33.0	Varianza	121.0
Asimetría	0.9	C.V.	31.2
Kurtosis	0.3	Rango	51.0





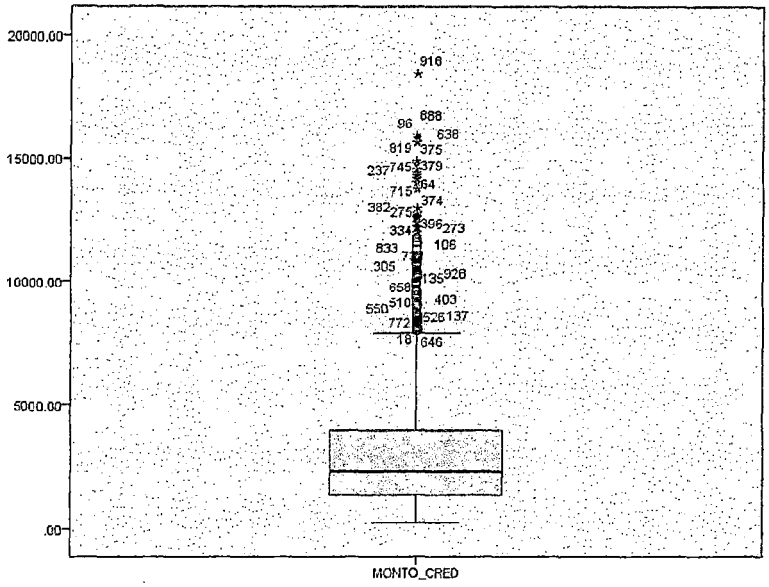
Encontramos una reducción en el coeficiente de asimetría y en el coeficiente de variación.

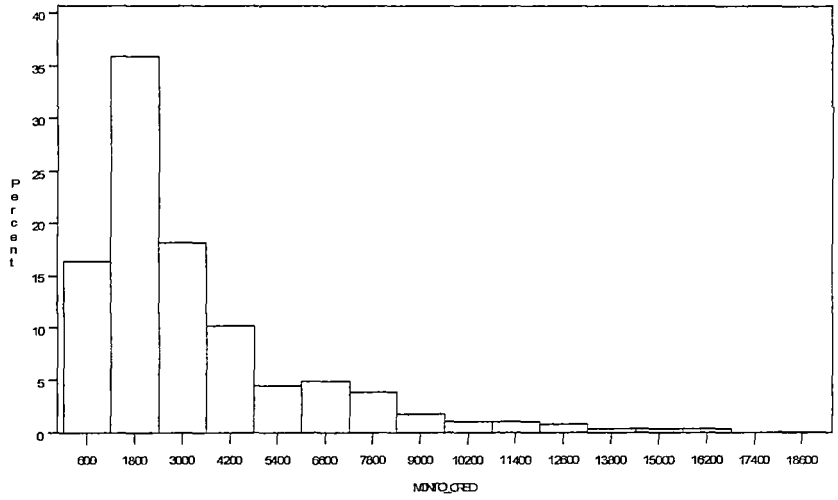
Variable Monto de crédito

Podemos notar que hay un alto nivel de variabilidad por el indicador de C.V. y un sesgo hacia los montos de crédito señalado por coeficiente de Asimetría.

Gráfico N° 18

Promedio	3271.3	Std. Dst.	2823.0
Mediana	2319.5	Varianza	7967843.5
Asimetría	1.9	C.V.	86.3
Kurtosis	4.3	Rango	18174.0





Al aplicar la detección de valores outliers potenciales fuera del rango intercuartílico hemos encontrado la siguiente lista de montos de créditos prestados:

OBS	MONTO_CRED
19	12579
58	9566
64	14421
79	9436
88	12612
96	15945
106	11938
135	10144
181	9572
206	10623
227	10961
237	14555
273	12169
275	11998
286	10722
292	9398
296	9960
305	10127
334	11590
374	13756
375	14782
379	14318
382	12976
396	11760
432	11328
451	11054

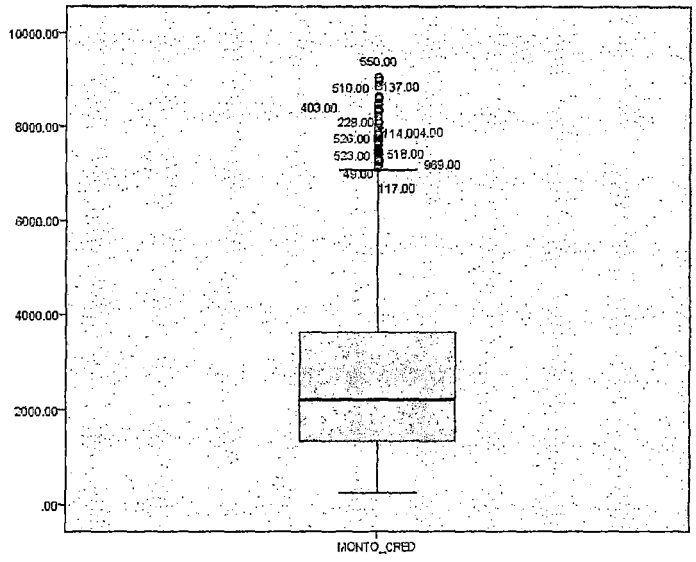
OBS	MONTO_CRED
564	12389
616	12204
617	9157
638	15653
658	10222
673	10366
685	9857
715	14027
737	11560
745	14179
764	12680
806	9271
809	9283
813	9629
819	15857
833	11816
855	10875
882	9277
888	15672
903	10477
916	18424
918	14896
922	12749
928	10297
954	10974

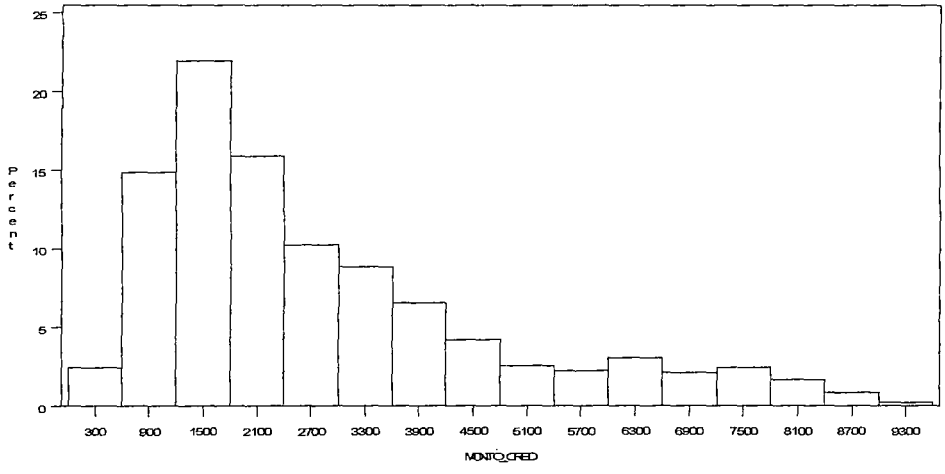
5.1% de valores atípicos potenciales.

Luego de retirar los potenciales valores atípicos detectados por el rango intercuartílico podemos ver el efecto que tiene sobre los principales indicadores, de tendencia central y dispersión (coef. Asimetría, promedio y C.V.).

Gráfico N° 19

Promedio	2804.2	Std. Dst.	1964.0
Mediana	2212.0	Varianza	3857321.0
Asimetría	1.2	C.V.	70.0
Kurtosis	0.8	Rango	8805.0



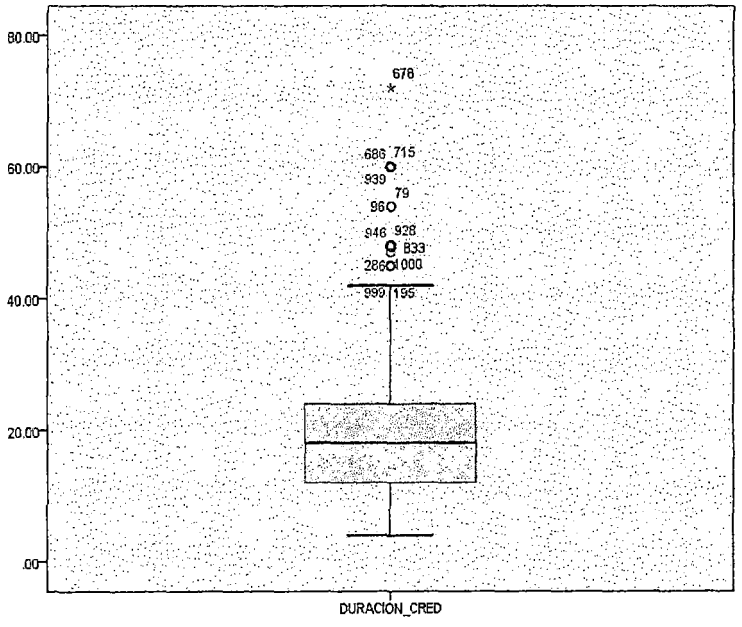


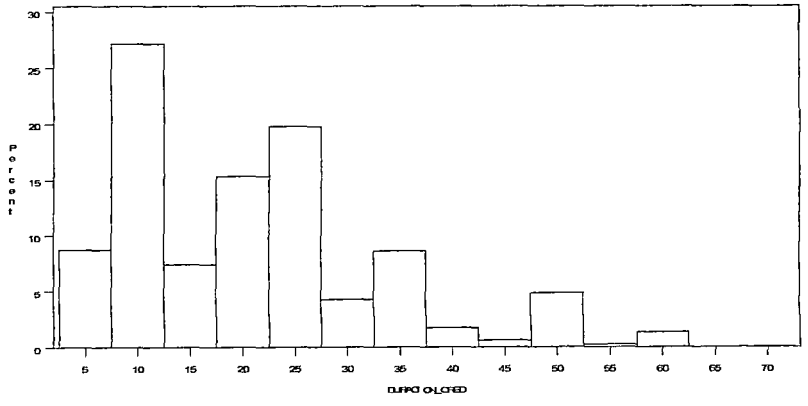
Notamos la persistencia de valores atípicos después de haber retira un primer grupo, pero un segundo retiro implicaría en una mayor pérdida de información, por ello con la finalidad de ser un revisión exploratoria de solo la variable monto de crédito y no haber realizados las asociaciones con las variables restantes, nos limitaremos a etiquetar solo el primer grupo de casos potencialmente atípicos del monto de crédito prestado.

Variable Número de meses del crédito

Gráfico N° 20

Promedio	20.9	Std. Dst.	12.1
Mediana	20.9	Varianza	145.4
Asimetría	1.1	C.V.	57.7
Kurtosis	0.9	Rango	68.0





Al aplicar la detección de valores outliers potenciales fuera del rango intercuartílico hemos encontrado el caso siguiente de meses de crédito:

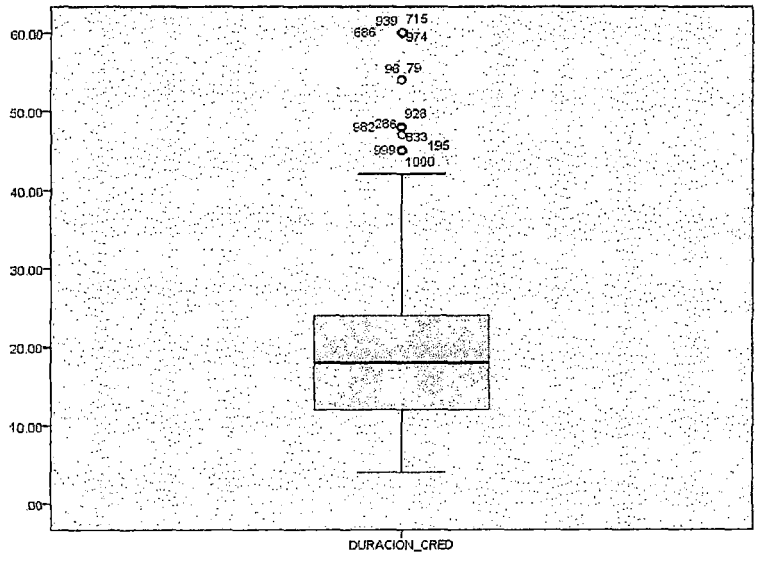
OBS	DURACION_CRED
678	72

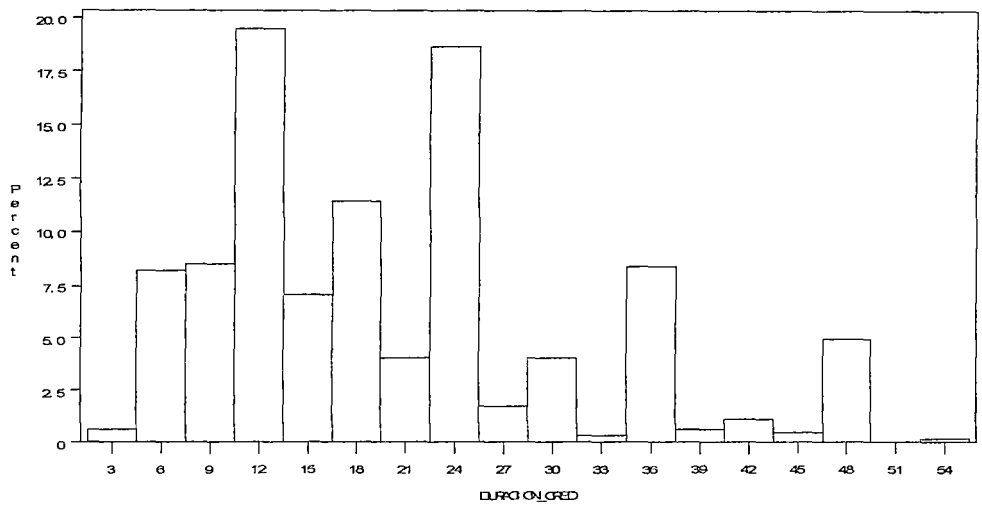
0.1% de valores atípicos potenciales.

Al ser el caso de mayor valor comparado con el resto de casos notamos la reducción en la asimetría de la distribución de los datos.

Gráfico N° 21

Promedio	20.3	Std. Dst.	11.2
Mediana	18.0	Varianza	124.3
Asimetría	0.9	C.V.	54.8
Kurtosis	0.2	Rango	50.0





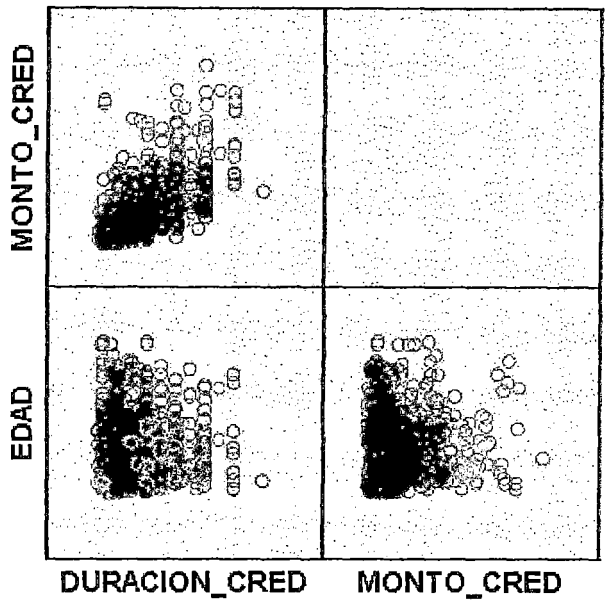
Detección de atípicos multivariados

Variables: Edad – Monto de Crédito – Número de meses del crédito

Por la sensibilidad a la detección de outlier que presenta el método de k-means, es muy utilizado en la detección de outlier multivariados. El procedimiento es solicitar un gran número de cluster (por ejemplo 50 cluster) y aquellos cluster con pocas observaciones y con mayor distancia entre los demás cluster serán los potenciales outliersmultivariados.

Los 50 clústeres fueron generados en el anexo VIII.6.1, notaremos que elegimos que el punto de cortes para denominar al clúster como un grupo de outliers multivariado fu que tenga una frecuencia de menos de 0.4% de los casos, este punto de corte resulto en la detección de 4.6% (46 casos de 1000) de casos potencialmente Outliers multivariados y se grafica en la figura siguiente (donde las esferas de color verde representan los casos de outlier multivariados):

Gráfico N° 22



		DURACION_CRED	MONTO_CRED
MONTO_CRED	Correlación de Pearson	**.599**	
	Sig. (bilateral)	*.000*	
EDAD	Correlación de Pearson	**-.045**	**.024**
	Sig. (bilateral)	*.166*	*.451*

Podemos notar una reducción en las asaciones entre las variables, pero hemos mantenido o mejorado la significación de las asaciones entre las Edad – Duración de crédito y Monto de crédito – Duración de crédito.

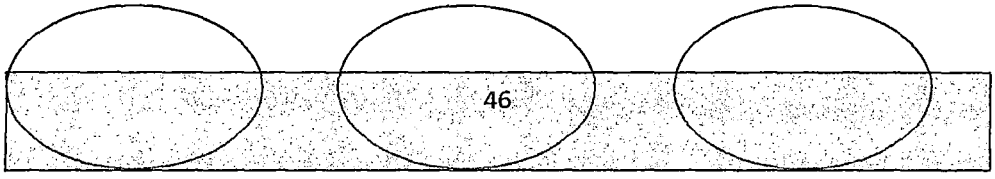
Determinación y eliminación de los valores atípicos

A continuación realizaremos un procedimiento de integración de los diferentes casos potenciales valores atípicos de las variables, buscando cuidar las relaciones entra las variables. Por ellos nos plateamos escenarios de eliminaciones de casos y observamos los indicadores delos coeficiente de correlación de Pearson entre las variables, aquel

escenario que mantenga o fortalezca las relaciones entre las variables será el escenario elegido. Recordemos que lo principal es fortalecer la estructura de las relaciones entre las variables pues estas soportarán los patrones que hallarán los modelos predictivos, al margen de solo reducciones individuales de asimetría de las variables.

Diagrama de outliers detectados:

Variable Edad(6) **Variable Monto de Crédito (51)** **Variable Duración de crédito (1)**



Outliers Multivariados (46)

Encontramos que las 46 variables conglomeran casos de los 3 tipos ya detectados por procedimientos individuales de las variables, por ellos con esta evidencia procedemos a retirar de la muestra principal estos 46 casos como Outliers.

V.1.3. Revisión de la Multicolinealidad de las variables

Procesamos los índices de condición y sus respectivas proporciones de varianza para las variables cuantitativas, debido a que la multicolinealidad es señalada como la combinación lineal de un conjunto de variables que puedan representar a alguna de ellas. Sería incongruente afirmar que apliquemos este enfoque a variables cualitativas ya que las variables cualitativas representan la presencia o ausencia de una característica y una combinación lineal de ellas tendría sentido, tal como afirmar que el efecto del nivel de educación de un cliente es una combinación ponderada de los efectos de saber tiempo viviendo en su residencia y el efecto de su nivel educativo. Entonces, para teste la multicolinealidad de nuestra data seguimos el procedimiento.

Aplicar el indicador de Factor de inflación de Varianza (mencionado en la sección II.3.1.1.), notamos que ningún VIF sobrepasa el a 10 (Anexo VIII.6.2), por ende podemos concluir que no existe problemas de multicolinealidad significativa. Solo las variables créditos debidamente pagados (Hist_cred_2 VIF=8.64) y cuenta crediticia critica (Hist_cred_4 VIF=7.22) son las variables que tiene un ligero problema de colinealidad con las demás variables regresoras como se muestra en el gráfico 23.

Las variables presentadas en el gráfico son las siguientes : Créditos vigentes pagados (Hist_cred_2), Cuenta crítica (Hist_cred_4), Residencia actual del cliente propia (Resid_Propietario), Residencia actual del cliente alquilada (Resid_Alquilada), Propósito del crédito comprar TV y equipo de sonido (Prop_Radio_TV), Propósito del crédito comprar auto nuevo (Prop_Aut_nuevo), Propósito del crédito comprar muebles (Prop_muebles), Retraso en pagos (Hist_cred_3), nivel educativo del cliente Secundaria completa (Niv_Ed_2), Propósito del crédito comprar auto usado (Prop_Aut_usado), nivel educativo del cliente Universidad Incompleta (Niv_Ed_3), No es propietario de su residencia actual (No_propietario_resid), Propósito del crédito

refinanciamiento de deuda (Prop_refinac), Créditos solicitados pagados (Hist_cred_1) y Antigüedad en el trabajo mayor a 7 años (Ant_trab_3) .

Gráfico N° 23

Variable	DF	Estimado	Error Std.	T Value	Pr > |t|	Tolerancia	VIF
Hist_cred_2	1	0.20	0.07	2.7	0.008	0.116	8.64
Hist_cred_4	1	0.31	0.07	4.1	<.0001	0.139	7.22
RESID_PROPIETARIO	1	-0.07	0.07	-1.0	0.325	0.154	6.48
RESID_ALQUILADA	1	-0.15	0.07	-2.0	0.052	0.192	5.21
PROP_RADIO_TV	1	-0.01	0.06	-0.2	0.869	0.202	4.94
PROP_AUT_NUEVO	1	-0.14	0.06	-2.3	0.024	0.226	4.42
PROP_MUEBLES	1	-0.02	0.07	-0.4	0.726	0.245	4.08
Hist_cred_3	1	0.21	0.08	2.6	0.009	0.308	3.25
Niv_Ed_2	1	0.00	0.06	0.0	0.987	0.326	3.07
PROP_AUT_USADO	1	0.09	0.07	1.3	0.200	0.337	2.97
Niv_Ed_3	1	-0.01	0.05	-0.3	0.763	0.341	2.93
NO_PROPIETARIO_RESID	1	-0.09	0.06	-1.5	0.141	0.355	2.82
PROP_REFINAC	1	-0.01	0.07	-0.2	0.837	0.360	2.77
Hist_cred_1	1	0.02	0.09	0.2	0.856	0.394	2.54
Ant_trab_3	1	0.07	0.05	1.6	0.113	0.401	2.50

V.1.4. Análisis descriptivo de las variables predictivas

El análisis descriptivo de las variables de manera univariados contra la variable target (incumplimiento de pago del crédito solicitado), involucra como señala Mandouft[31] alcanzar un mayor conocimiento de las variables y da soporte analítico de las variables que podrían emplearse en el modelo predictivo, lográndose alcanzar un profundo conocimiento de las variables.

Podemos priorizar la visióndescriptiva de las variables según la relevancia del hallazgo estadístico. Por ello generamos el listado de **InformationValue** para las variables

[31][19] MamdouhReffat (pág. 83)

independiente, y lograr una mayor comprensión de los cruces de variables independientes con la variable target cuando sean más significativas. Consideremos que los cruces fueron realizados con variables dicotómicas, cualitativas y cuantitativas (aplicando quintiles por tratarse de una primera inspección de las variables).

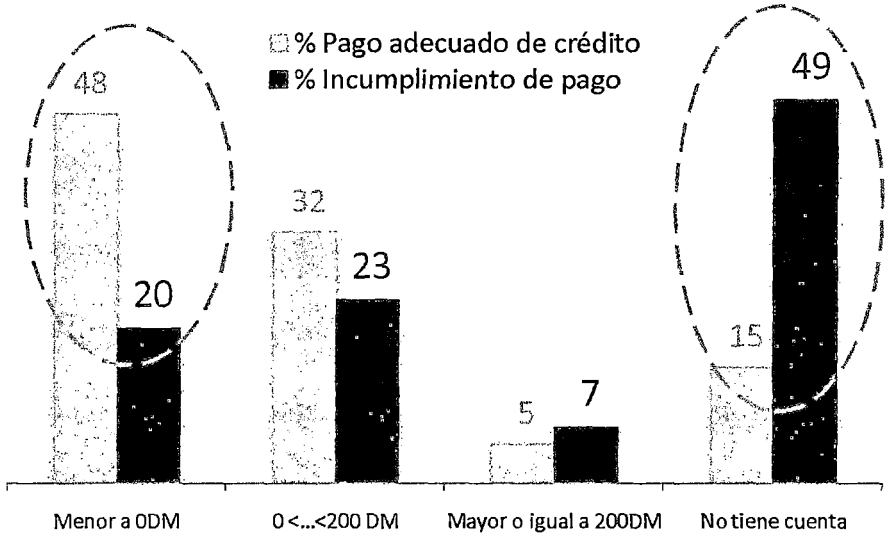
Gráfico N° 24

Descripción	Nombre variable	IV	Poder predictivo
Estado de cuenta	CHK_ACCT	0.69	Fuerte
Historial crediticio	HISTORIA_CRED	0.31	Fuerte
Duración de crédito	DURACION_CRED_B	0.22	Medio
Balance promedio en cuenta de ahorro	BALANC_ACCT	0.2	Medio
Propósito del crédito Auto usado	PROP_AUT_USADO	0.09	Débil
Edad del cliente	EDAD_B	0.09	Débil
Antigüedad en su empleo	ANTIG_EMPLEO	0.09	Débil
Monto de crédito	MONTO_CRED_B	0.08	Débil
Propietario de su residencia	RESID_PROPIETARIO	0.07	Débil
Empresa propia	EMPRESA_PROPIA	0.06	Débil
Tiene otro plan de crédito	OTROS_PLAN	0.05	Débil
Estado real de la propiedad	ESTADO_PROPIEDAD	0.05	Débil
Porcentaje de ingreso disponible	TASA_INGR_DISPONIBLE	0.05	Débil
Propósito del crédito Auto nuevo	PROP_AUT_NUEVO	0.05	Débil
Vivienda alquilada	RESID_ALQUILADA	0.04	Débil
No es propietario de residencia	NO_PROPIETARIO_RESID	0.04	Débil
Propósito del crédito Radio/Tv	PROP_RADIO_TV	0.04	Débil
Estado civil de cliente soltero	SIT_CIVIL_SOLT	0.04	Débil
Propósito del crédito Educación	PROP_EDUCACION	0.03	Débil
Vehículo propio	VEHICULO_PROP	0.02	No predictivo
Nro. de créditos existentes en el banco	NUM_CREDITS	0.02	No predictivo
Socio en la solicitud del crédito	CO_APPLICANT	0.01	No predictivo
Persona Aval	GUARANTIA	0.01	No predictivo
Divorciado	SIT_CIVIL_DIV	0.01	No predictivo
Propósito del crédito Muebles	PROP_MUEBLES	0	No predictivo
Propósito del crédito Refinanciamiento	PROP_REFINAC	0	No predictivo
Nivel educativo	NIVEL_EDUCATIVO	0	No predictivo
Antigüedad en su residencia actual	ANTIG_RES_ACTUAL	0	No predictivo
Casado o conviviente	SIT_CIVIL_CASADO	0	No predictivo
Número de dependientes	NUM_DEPENDENTS	0	No predictivo

Podemos destacar del gráfico 24, que las variables que individualmente tienen mayor relevancia (según los citado en la sección II.1.3) para pronosticas el incumplimiento de pago son Estado en Cuenta (CHK_ACCT) y el Historial Crediticio (HISTORIA_CRED).

Estado de Cuenta

Gráfico N° 25



Del gráfico notamos que hay marcados niveles de riesgo en los estados de cuenta de **Menor a 0 DM** y que **No tiene cuenta**.

Interpretación Estadística:

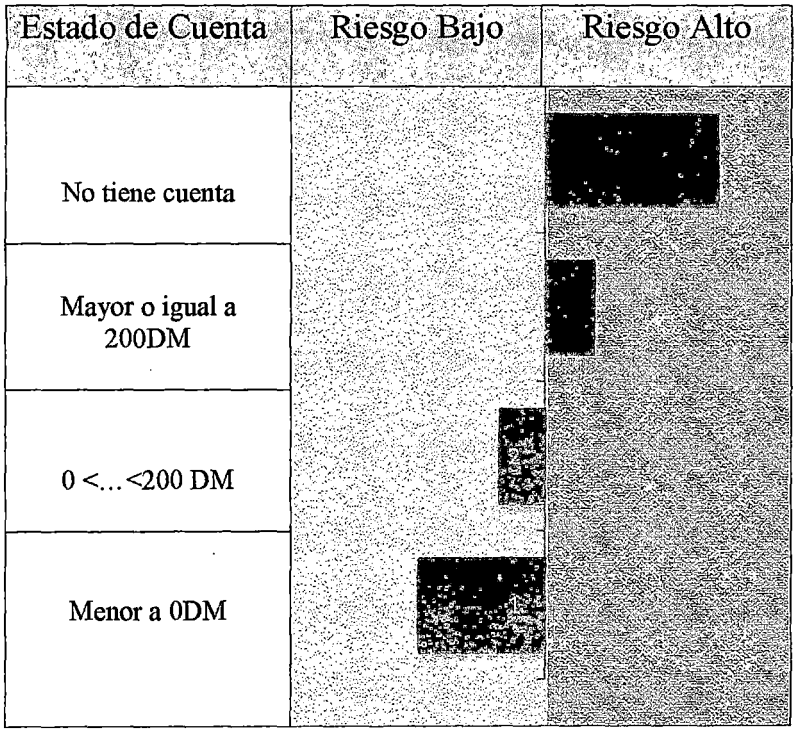
Si elegimos al azar entre un grupo de clientes con un estado de cuenta **Menor a 0 DM**, hay casi el doble de posibilidad (48% / 20%) de elegir a un cliente con un pago adecuado crédito con respecto a elegir a un cliente con caiga en el incumplimiento de pago crediticio.

Interpretación Negocio:

Aquellos clientes que tienen un estado de cuenta **Menor a 0 DM**, debido a su bajo endeudamiento están sujetos, en base a la historia, a un menor nivel de riesgo de incumplimiento de pago.

Adicional a lo mostrado el indicador WOE nos puede dar una visión de la naturaleza del riesgo en los diferentes rangos de las variables así como vemos a continuación:

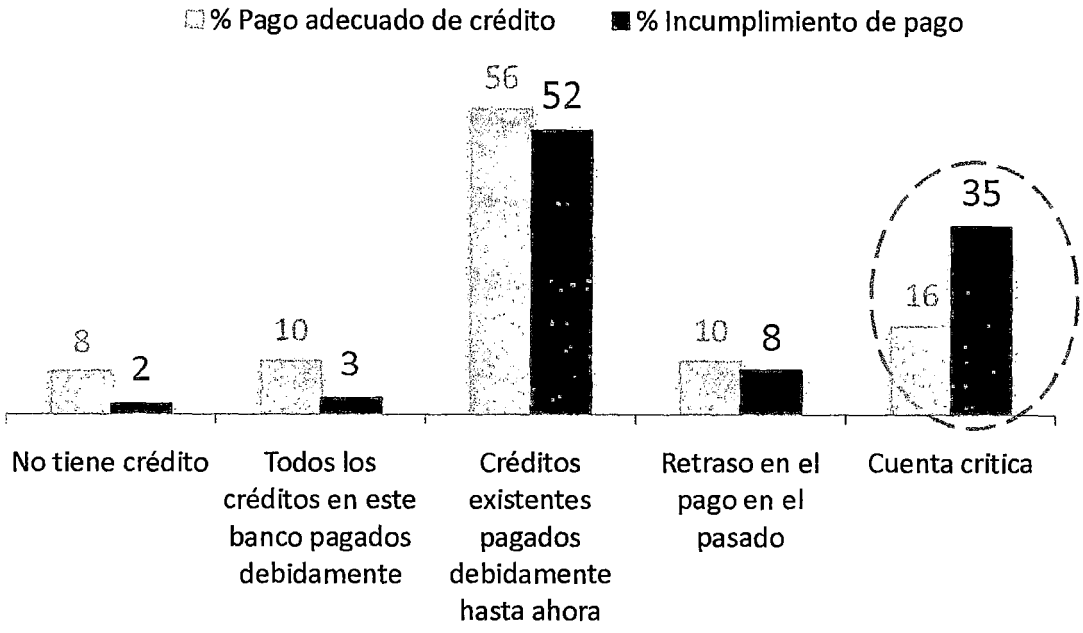
Gráfico N° 26



Es evidente que los clientes más propensos a caer en el incumplimiento de pago sabiendo su estado de cuenta son aquellos **No tienen cuenta** actualmente y que tienen una cuenta o deuda **Mayor o igual a 200 DM**.

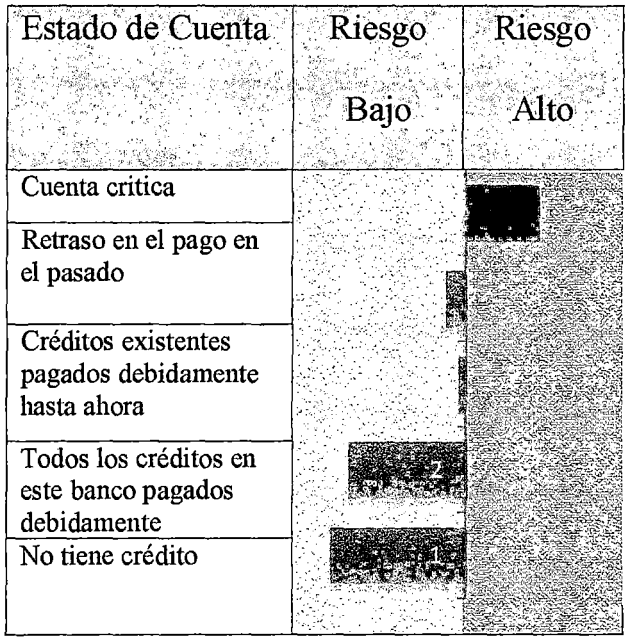
Historial Crediticio

Gráfico N° 27



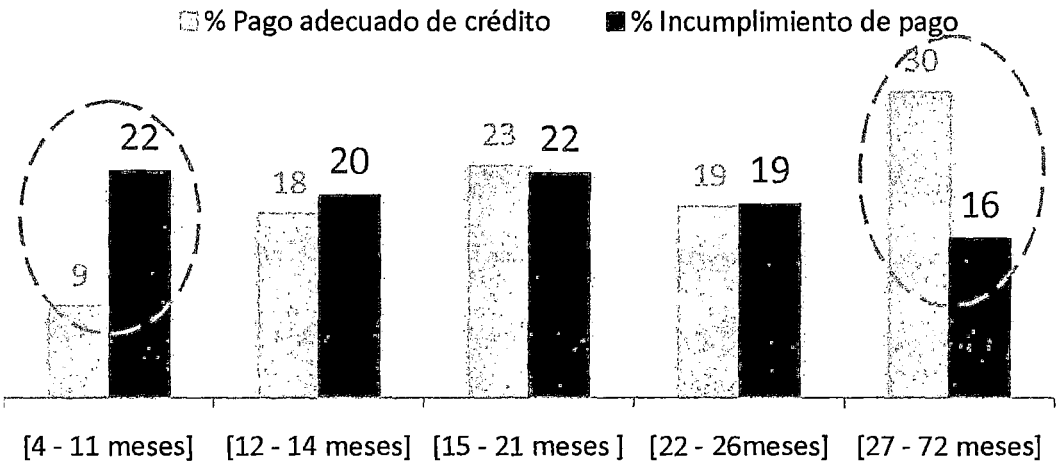
Notamos que hay un marcados nivele de riesgo en los clientes que tenían su historial crediticio una **cuenta crítica,** estos clientestiene una posibilidad de 2 a 1 a caer en incumplimiento de pago.

Gráfico N° 28



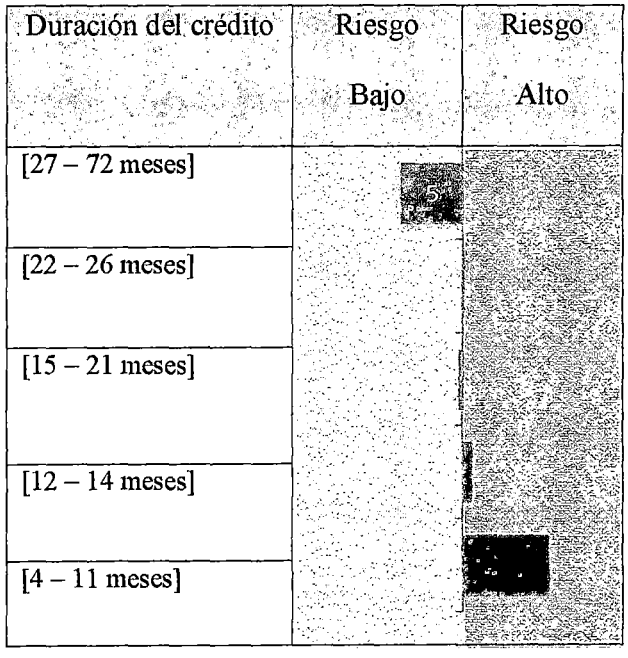
Duración del Crédito

Gráfico N° 29



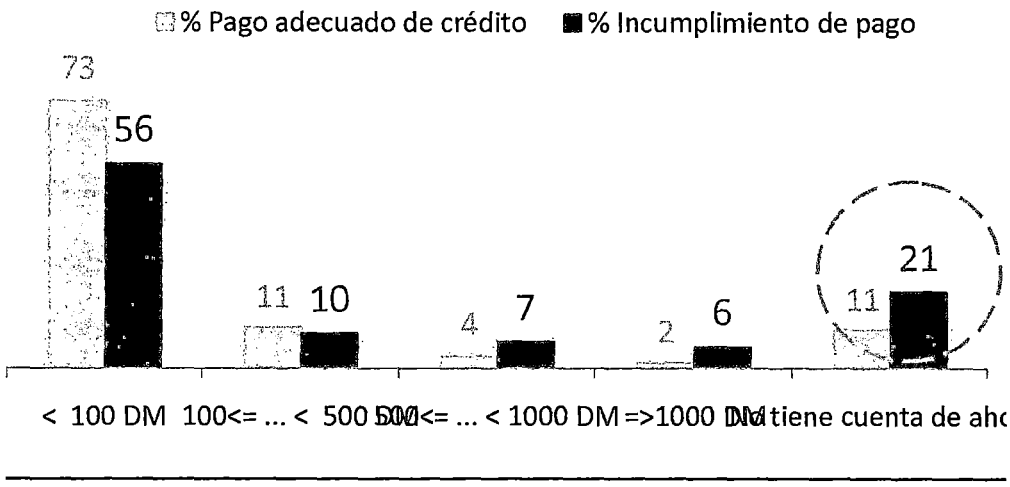
Notamos que hay similares niveles de riesgo entre créditos con duración de 12 a 26 meses.Por otro lado aquellos créditos otorgados con menos de 11 meses presentan el mayor riesgo de incumplimiento de pago y los créditos otorgados con más de 27 meses de pago se registran la propensión más baja de incumplimiento de pago.

Gráfico N° 30



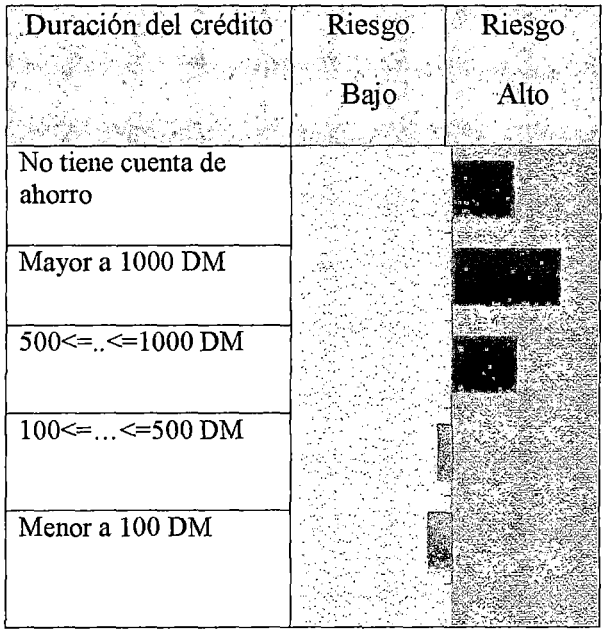
Balance de cuenta de ahorro

Gráfico N° 31



Hay un mayor nivel de riesgo de incumplimiento de pago cuando se tiene un balance de la cuenta de ahorro mayor a 500 DM (siendo pocos estos casos de clientes), adicionalmente la mayor propensión al incumplimiento de pago se concentra en aquellos clientes que **no tienen una cuenta de ahorro**.

Gráfico N° 32



V.1.5. Selección de variables predictivas

Selección de variables sistemática de las variables que puede ser empleando un criterio Stepwise aplicando un criterio de AIC para aplicar el criterio de parsimonia del modelo predictivo. La sección anterior nos fue útil para identificar la información más relevantes(variables), comprenderlas y realizar un primer descarte de las variables que no son predictivas.Del gráfico 23, rescataríamos como variables potenciales para el modelo obviaremos aquellas variables que tuvieron una clasificación de IV no predictivo, quedándonos con 19 variables del grupo inicial de 30 variables.

Las variables con significancia sistemática:

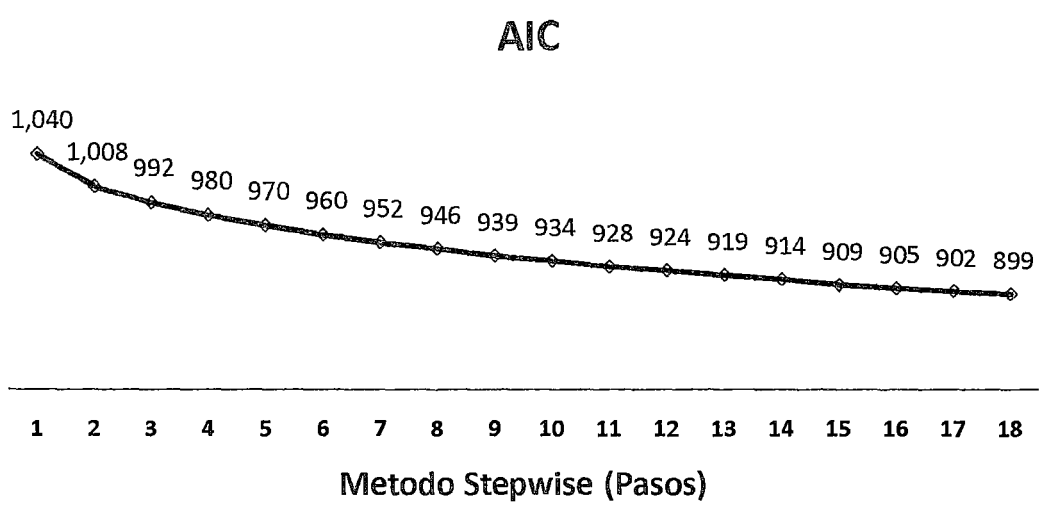
Gráfico Nº 33

Variable entrante	Variable removida	GL	Paso	Score Chi-Square	Chi-Square P-value
Est_Cta_3		1	1	96.8	<.0001
DURACION_CRED		1	2	34.4	<.0001
Hist_cred_4		1	3	17.0	<.0001
PROP_AUT_NUEVO		1	4	14.8	0.0001
Pr_ct_aho_4		1	5	11.5	0.0007
PROP_EDUCACION		1	6	12.4	0.0004
TASA_INGR_DISPONIBLE		1	7	10.0	0.0016
SIT_CIVIL_SOLT		1	8	8.5	0.0035
Hist_cred_2		1	9	9.1	0.0026
Hist_cred_3		1	10	7.1	0.0076
Pr_ct_aho_3		1	11	7.0	0.0083
Ant_trab_2		1	12	5.6	0.0175
PROP_AUT_USADO		1	13	6.0	0.0143
Est_Cta_2		1	14	6.3	0.0122
Est_Cta_1		1	15	7.6	0.0058
EMPRESA_PROPIA		1	16	4.9	0.0274
RESID_ALQUILADA		1	17	4.9	0.0273
OTROS_PLAN		1	18	5.2	0.022

Las variables presentadas en el tabla son las siguientes : No tiene estado de cuenta (Est_Cta_3), Duración del crédito (Duracion_cred), Cuenta crítica (Hist_cred_4), Propósito del crédito comprar auto nuevo (Prop_Aut_nuevo), No tiene cuenta de ahorro (Pr_ct_aho_4), Propósito del crédito Educativo (Prop_Educacion) , Porcentaje de ingreso disponible del cliente después de gastos (Tasa_ingr_disponible), Estado Civil Soltero (Sit_Civil_Sol), Créditos vigentes pagados (Hist_cred_2), Retraso en pagos (Hist_cred_3), Promedio de cuenta de ahorro mayo a 100DM (Pr_ct_aho_3), Antigüedad en el trabajo entre 4 y 7 años (Ant_trab_2),Propósito del crédito comprar auto usado (Prop_Aut_usado), Estado de cuenta Mayor a 200DM (Est_Cta_2), Estado de cuenta entre 0 y menor a 200DM (Est_Cta_1), El cliente tiene empresa propia (Empresa_propia), Lugar de residencia del cliente alquilada (Resid_alquilada) y Otro plan de cuenta (Otros_plan).

Las variables de modelamiento se redujeron a 11 a 12 variables significativas: Estado en cuenta de ahorro, duración del crédito, historial crediticia, propósito de crédito, tasa de ingreso disponible, situación civil soltero, porcentaje en cuenta de ahorro, antigüedad en el trabajo, flag de empresa propia y lugar de residencia. Notemos como el ingreso de las variables en los pasos del método Stepwise para el modelo predictivo hace que caiga el AIC continuamente.

Gráfico N° 34

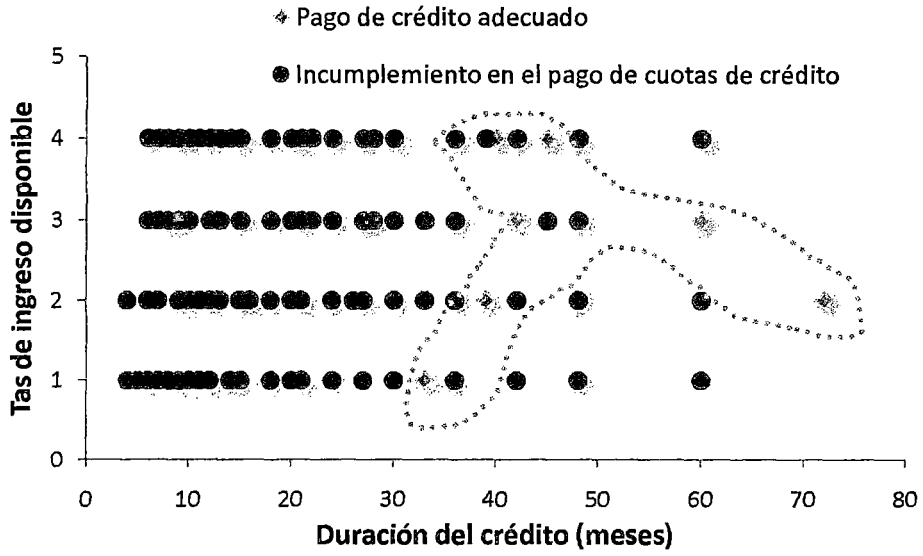


V.1.6. Evaluación de la aplicación de técnica lineal o no lineal

Aplicando el algoritmo de regresión logística solo conseguimos identificar dos clases de poblaciones linealmente separables. Es decir que es rígida la identificación del nivel de riesgo según los tramos o regiones del score de riesgo de crédito, en contraste la técnica Máquinas de Vectores de Soporte No Lineal sugiere modelo más robusto ante esta problemática (en la sección II.1.5.2).

Como herramienta de diagnóstico del problema de separabilidad no lineal de poblaciones, planteamos una revisión descriptiva de las 2 variables independientes continuas.Cuando no es posible definir una recta de clasificación entre dos poblaciones que alcance los requerimientos de precisión del pronóstico será motivo de sugerir un modelo no linealmente separable.

Gráfico N° 35



En el gráfico 35, debido a que no hay una recta clara de clasificación entre los casos de incumplimiento de pago y pago de crédito se propone usar un algoritmo de separabilidad de clases que no sean separables por un hiperplano, tal como el método de Máquinas de Vectores de Soporte Kernel no linealmente separable.

Si buscamos extendemos nuestro diagnostico descriptivo del problema de la separabilidad no lineal para el caso que se cuente con más de 2 variables continuas independientes sugerimos aplicar las 2 primeras componentes principales como herramienta de evaluación gráfica de las características de ambas poblaciones para poder una regla se clasificación.

V.2Modelo de Regresión Logística

1° Estimación de nuestro modelo de Regresión Logística

```
> MLOG<-glm(LOGTRAIN$Default~.,data=LOGTRAIN,family=binomial())
>summary(MLOG)
```

El procedimiento siguiente es estimar la significancia de los parámetros de nuestro modelo Logístico, para todos las variables. Los resultados son:

Variables	Coef. Estimate	Std. Error	z Value	Pr(>|z|)	Significancia
(Intercepto)	0.73	0.52	1.405	0.16	
Est_Cta_3	1.57	0.27	5.897	0.00	***
DURACION_CRED	-0.04	0.01	-4.249	0.00	***
Hist_cred_4	1.61	0.40	4.047	0.00	***
PROP_AUT_NUEVO	-0.91	0.24	-3.861	0.00	***
Pr_ct_aho_4	0.72	0.30	2.417	0.02	*
PROP_EDUCACION	-0.89	0.44	-2.003	0.05	*
TASA_INGR_DISPONIBLE	-0.24	0.10	-2.447	0.01	*
SIT_CIVIL_SOLT	0.38	0.21	1.766	0.08	.
Hist_cred_2	0.99	0.36	2.786	0.01	**
Hist_cred_3	1.06	0.46	2.312	0.02	*
Pr_ct_aho_3	1.10	0.69	1.595	0.11	
Ant_trab_2	0.63	0.30	2.094	0.04	*
PROP_AUT_USADO	1.35	0.56	2.398	0.02	*
Est_Cta_2	0.71	0.39	1.811	0.07	.
Est_Cta_1	0.46	0.25	1.831	0.07	.
EMPRESA_PROPIA	1.26	0.84	1.505	0.07	.
RESID_ALQUILADA	-0.53	0.26	-2.041	0.04	*
OTROS_PLAN	-0.60	0.25	-2.376	0.02	*

Las variables presentadas en el tabla son las siguientes : No tiene estado de cuenta (Est_Cta_3), Duración del crédito (Duracion_cred), Cuenta crítica (Hist_cred_4), Propósito del crédito comprar auto nuevo (Prop_Aut_nuevo), No tiene cuenta de ahorro (Pr_ct_aho_4), Propósito del crédito Educativo (Prop_Educacion), Porcentaje de ingreso disponible del cliente después de gastos (Tasa_ingr_disponible), Estado Civil Soltero (Sit_Civil_Sol), Créditos vigentes pagados (Hist_cred_2), Retraso en pagos (Hist_cred_3), Promedio de cuenta de ahorro mayo a 100DM (Pr_ct_aho_3), Antigüedad en el trabajo entre 4 y 7 años (Ant_trab_2),Propósito del crédito comprar auto usado (Prop_Aut_usado), Estado de cuenta Mayor a 200DM (Est_Cta_2), Estado de cuenta entre 0 y menor a 200DM (Est_Cta_1), El cliente tiene empresa propia

(Empresa_propia), Lugar de residencia del cliente alquilada (Resid_alquilada) y Otro plan de cuenta (Otros_plan).

Códigos de nivel de significancia:

0	'***'
0.001	'**'
0.01	'*'
0.05	'.'
0.1	' '

TABLA II

```
Null deviance: 793.33  on666  degrees of freedom
Residual deviance: 613.38  on648  degrees of freedom
AIC: 651.38
Number of Fisher Scoring iterations: 5
```

Se asocial la diferencia de devianza a la distribución Chi cuadrado con la diferencia de grados de libertad del modelo saturado modelo y modelo estimado como parámetro de la distribución. Y podemos percatarnos que la hipótesis de nula se rechaza aceptando el modelo para un nivel significancia del 5% como se evidencia.

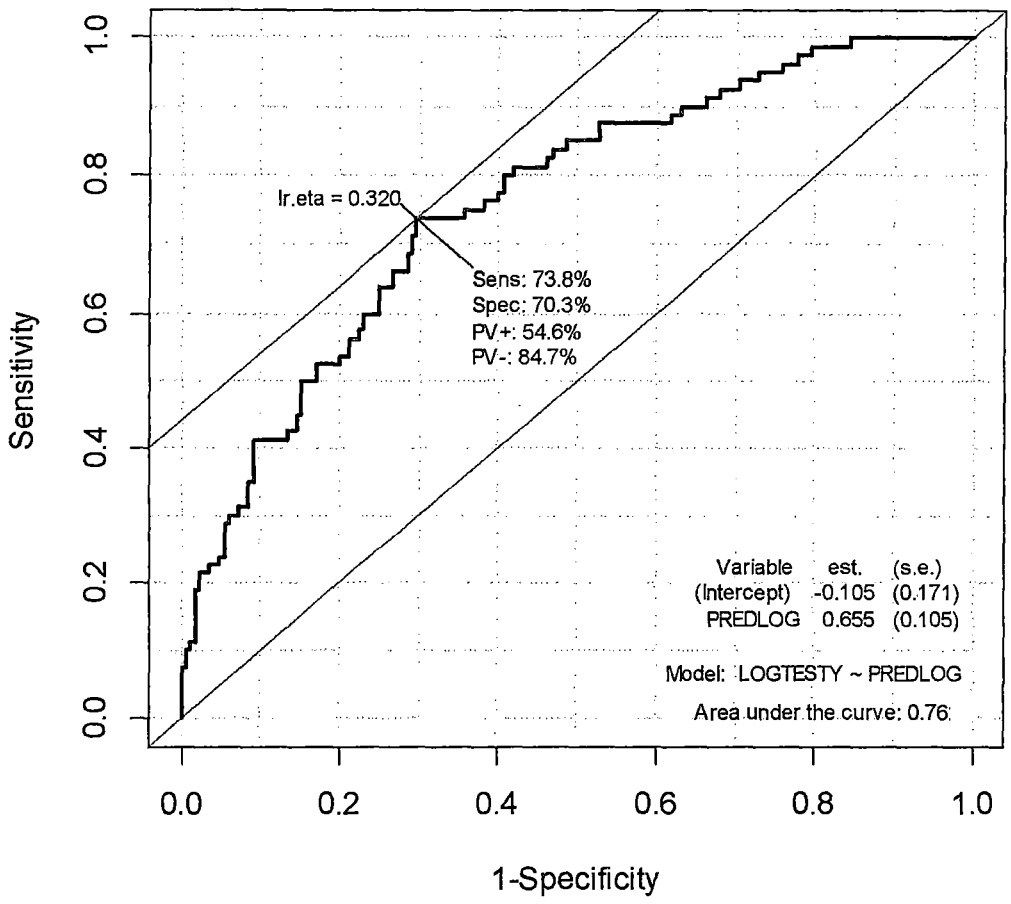
gl	12
D0-D1	179.95

Si bien este modelo se ajusta a nuestra data separada de aprendizaje, el paso siguiente es mejorar nuestra bondad de ajuste mediante una selección de variables que sean significativos para nuestro modelo y no considerar a todas las inicialmente examinadas.

2°-Determinacion de indicadores

El siguiente procedimiento es medir el poder predictivo del modelo de regresión logística hallado, esto se realizara haciendo uso de una muestra de prueba que hemos separado aleatoriamente de nuestra data histórica inicial con la finalidad de encontrar los indicadores de valores predictivos detallados ya antes detallados como son los indicadores de sensibilidad y especificidad.

```
> PREDLOG<-predict(MLOGSTEP,LOGTESTX)
>library(Epi)
>ROC(form=LOGTESTY~PREDLOG,plot="ROC",data=LOGTEST)
```



	Sensibilidad	Especificidad	PV+	PV-
Modelo Logístico	73.80%	70.30%	54.60%	84.70%

V.3Modelo de Máquinas de Vectores de Soporte

Para la correcta descripción del modelo el procedimiento será especificado por pasos:

1° Selección de variables

Para una selección de variable con sustento estadístico, usamos las hasta ahora variables seleccionadas de nuestro modelo de regresión Logística.

La selección de variables mediante la aplicación de una regresión logística, con una revisión previa del supuesto de Multicolinealidad y dar el tratamiento necesario a los casos atípicos. Elegimos los factores que sean significativos en la regresión logística con un nivel de confianza de 1%, tendrán la base estadística suficiente para que ingresen a nuestro modelo predictivo SVM.

2° Determinación de los secundarios

Los modelos de Máquinas de Vectores de Soporte están guiados principalmente por dos secundarios (hiperparámetros) el C costo y el g gamma, los cuales son estimados para un modelo particular, pero para poder validar y generalizar los resultados, directamente con el modelo encontrado no es aceptables. Debido a que por su naturaleza de algoritmo de optimización matemático tiene una rigidez en sus resultados y no podrían ser generalizados para realizar inferencia.

En este contexto el método de validación cruzada nos permitirá hallar los mejores hiperparámetros óptimos que nos permita validar y extender nuestro modelo para hacer inferencia.

En vista a ello nuestro primer paso a seguir será sintonizar (tune) aquellos hiperparámetros óptimos para poder encontrar nuestro modelo SVM (por sus siglas en

ingles).Para esto usamos como conjunto posible (espacio de búsqueda) de valores de hiperparámetros: Costo sea 2^(-3:3) Gamma sea 2^(-3,3)

Para cada par de hiperparámetros, se conducirá una validación cruzada de 10 veces.

La TABLA V nos indica que en las diferentes pruebas realizadas encontramos que los mejores son**C=1 y Gamma=0.125**

```
> SVMTUNE<-tune.svm(Default~ .,data=SVMTRAINSELECT,gamma=2^(-
3:3),cost=2^(-3:3))
>summary(SVMTUNE)
Parameter tuning of 'svm':
- sampling method: 10-fold cross validation
- best parameters:
gamma cost
 0.125    1
- best performance: 0.187861
- Detailed performance results:
gamma cost     error    dispersion
1  0.125       0.125  0.2494859      0.05035040
2  0.250       0.125   0.2532287     0.05210557
3  0.500       0.125  0.2549289      0.05279205
4  1.000       0.125   0.2554047     0.05296285
5  2.000       0.125   0.2555878     0.05298277
6  4.000       0.125   0.2556803     0.05298145
7  8.000       0.125  0.2557646      0.05298487
8  0.125       0.250  0.2343895       0.04534484
9  0.250       0.250  0.2411658      0.04872537
10 0.500       0.250   0.2443624     0.05010065
11 1.000        0.250  0.2452550     0.05044961
12 2.000       0.250  0.2455862       0.05049647
13 4.000       0.250  0.2457494      0.05049640
14 8.000       0.250  0.2459075      0.05050098
15 0.125       0.500  0.2105948       0.03549659
16 0.250       0.500  0.2213230      0.04170158
17 0.500       0.500  0.2269075       0.04443536
18 1.000       0.500  0.2284534       0.04516206
19 2.000       0.500   0.2289727      0.04528466
20 4.000       0.500   0.2292105      0.04529017
21 8.000       0.500  0.2294844      0.04528651
22 0.125       1.000  0.1878610      0.01835935
23 0.250       1.000   0.1983610      0.02734724
24 0.500       1.000   0.2064116      0.03235633
25 1.000       1.000   0.2087272      0.03372912
26 2.000       1.000  0.2093067      0.03393790
27 4.000       1.000  0.2094258       0.03394694
28 8.000       1.000   0.2098050      0.03388069
29 0.125       2.000   0.1887752      0.01224718
30 0.250       2.000  0.1956063      0.01843971
31 0.500       2.000   0.2028421      0.02304820
32 1.000       2.000  0.2049042      0.02382380
33 2.000       2.000   0.2054657      0.02384662
34 4.000       2.000   0.2057129      0.02365180
35 8.000       2.000   0.2062307      0.02335891
36 0.125       4.000   0.1929253      0.01410061
37 0.250       4.000  0.1964792      0.01910133
38 0.500       4.000  0.2026503      0.02261175
39 1.000       4.000  0.2045766      0.02351702
```

3°Modelamiento

Pasaremos a determinar nuestro modelo óptimo con los parámetros hallados:

```
> SVMFINAL<-svm(Default~ .,data=SVMTRAINSELECT,cost=1,gamma=0.125,cross=10)
>summary(SVMFINAL)
Call:
svm(formula = Default ~ ., data = SVMTRAINSELECT, cost = 1, gamma = 0.125,
cross = 10)
Parameters:
  SVM-Type:  eps-regression
 SVM-Kernel:  radial
cost:  1
gamma:  0.125
epsilon:  0.1
Number of Support Vectors:  686
10-fold cross-validation on training data:
Total Mean Squared Error: 0.1853748
Squared Correlation Coefficient: 0.1159999
Mean SquaredErrors:
 0.1758081 0.1677798 0.184328 0.1753073 0.1602967 0.1859789 0.186145 0.2162845
0.2016795 0.1998955
```

Detallamos:

La función kernel elegida es la función de base Radial. Además como ya habíamos establecido: **C=1 Gamma=0.125**

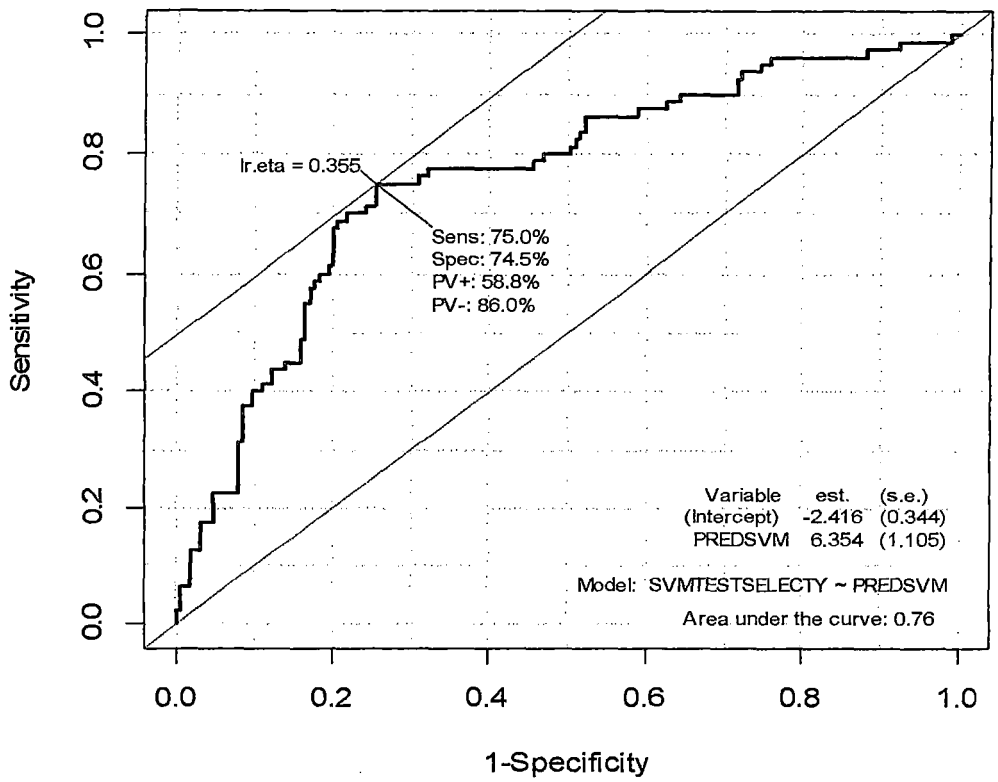
El número de Vectores de Soporte, se puede interpretar como el número de casos que nos definen una determinada clase de comportamiento de pago e incumplimiento de pago.Estos son un total de 686 de un total de 750 registros en nuestra data de aprendizaje.La tasa de mala clasificación para una validación cruzada en nuestra data de aprendizaje es de 0.18537 esto es aproximadamente 18.5% de tasa de mala clasificación promedio con el modelo elegido.

4° Determinación de indicadores

Como último paso de nuestro modelo evaluaremos su capacidad predictiva con una data previamente seleccionada que es un 25% de nuestra data inicial histórica, sobre la cual calcularemos nuestros indicadores de capacidad predictiva.

La metodología empleada del algoritmo de aprendizaje matemático SVM, para poder validar y generalizar sus modelos fue la identificación de parámetros secundarios, la estimación del modelo y la revisión de la tasa de error. LaMáquina de Vectores de Soporte (SVM) está guiado por dos hiperparámetros el C (costo) y gamma (argumento de la función kernel), entonces una vez identificado los hiperparámetros buscamos generalizar los resultados.

Un procedimiento aplicado es unparticionamiento aleatorio de nuestra data de aprendizaje en 10 partes usando siempre una décima parte como data de prueba y nueva decimas como data de aprendizaje. De modo que el último modelo es válido y puede ser extendido para realizar mejores predicciones. Este procedimiento es llamado clasificación cruzada 10 veces (siglas en ingles Cross Validación 10 fold).



	Sensibilidad	Especificidad	PV+	PV-
Modelo Logístico + SVM	75.00%	74.50%	58.80%	86.00%

Consolidado indicadores de rendimiento:

	Sensibilidad	Especificidad	PV+	PV-
Modelo Logístico	73.80%	70.30%	54.60%	84.70%
Modelo SVM	**75.00%**	**74.50%**	**58.80%**	**86.00%**

Encontramos indicadores de poder predictivos superiores a los encontrados con el modelo predictivo de Regresión Logística, estos resultados pueden ser atribuidos a que nos encontramos en un escenario de modelamiento complejo que no es fácilmente representado por una regla linear de clasificación. Una alternativa de modelamiento en este tipo resulta siendo la Máquina de soporte de Vectores este algoritmo trata de representar escenarios de riesgos explicables linealmente en un espacio de datos más complejo.

Adicionalmente, cabe atribuir que el sentido de asociación al riesgo de incumpliendo de pago es similar en ambos modelos predictivos como podemos apreciar:

Variables que mitigan riesgos

Según ambos modelos la propensión al incumplimiento del pago del crédito al estar correlacionados inversamente con el riesgo de crédito (Variables sombreadas en color verde). Las variables son:

Duración del crédito: Si es un crédito con más cuotas se reduce el riesgo al incumpliendo de pago.

Si tiene como **propósito de crédito la compra de un auto nuevo, crédito educativo u otros motivos** los listados tiene una menor propensión relativa al riesgo de crédito.

Tasa de ingreso disponible: Es lógico esperar un buen comportamiento de pago del crédito por aquellos clientes que tiene unas tasas mayores disponibles de ingreso.

Variables	SVM PESOS	SVM LOGIS
Ant_trab_2	304.0	0.63
DURACION_CRED	-2,761.4	-0.04
EMPRESA_PROPIA	438.3	1.26
Est_Cta_1	1,465.7	0.46
Est_Cta_2	1,487.0	0.71
Est_Cta_3	3,404.2	1.57
Hist_cred_2	21,739.4	0.99
Hist_cred_3	12,747.1	1.06
Hist_cred_4	20,613.8	1.61
OTROS_PLAN	-1,060.2	-0.60
Pr_ct_aho_3	778.5	1.10
Pr_ct_aho_4	1,188.9	0.72
PROP_AUT_NUEVO	-1,757.2	-0.91
PROP_AUT_USADO	1,015.4	1.35
PROP_EDUCACION	-822.4	-0.89
RESID_ALQUILADA	-1,695.4	-0.53
SIT_CIVIL_SOLT	670.2	0.38
TASA_INGR_DISPONIBLE	-1,092.7	-0.24

Las variables presentadas en el tabla son las siguientes : Antigüedad en el trabajo entre 4 y 7 años (Ant_trab_2), Duración del crédito (Duracion_cred),El cliente tiene empresa propia (Empresa_propia), Estado de cuenta entre 0 y menor a 200DM (Est_Cta_1), Estado de cuenta Mayor a 200DM (Est_Cta_2), No tiene estado de cuenta (Est_Cta_3), Créditos vigentes pagados (Hist_cred_2), Retraso en pagos (Hist_cred_3), Cuenta crítica (Hist_cred_4), Otro plan de cuenta (Otros_plan), Promedio de cuenta de ahorro mayo a 100DM (Pr_ct_aho_3), No tiene cuenta de ahorro (Pr_ct_aho_4), Propósito del crédito comprar auto nuevo (Prop_Aut_nuevo),Propósito del crédito comprar auto usado (Prop_Aut_usado),Propósito del crédito Educativo (Prop_Educacion), Lugar de residencia del cliente alquilada (Resid_alquilada), Estado Civil Soltero (Sit_Civil_Sol) y Porcentaje de ingreso disponible del cliente después de gastos (Tasa_ingr_disponible).

Variables de riesgos

La variable de **Historial crediticioprevio** es la que marca mayor propensión al incumplimiento de pago. Con mayor incidencia cuando se tiene una cuentas de deuda críticas o diferentes créditos pendientes de pago.

La siguiente variable que determina una alta propensión al incumplimiento de pago es **Balance en la cuenta de ahorro** lamayor propensión se registra cuando no se cuenta con una cuanta de ahorro o un saldo disponible mayor a 1000 DM.

Una variable que también marca un mayor riesgo de incumplimiento de pago es que el cliente tenga un estado **civil de soltero**.

VII. Conclusiones

1. Capacidad predictiva delos modelos.

La capacidad predictiva del modelo de Máquina de Vectores de Soporte (SVM) son superioresa los indicadores del modelo logístico en el análisis de riesgo crediticio para una base de datos de Banca Personal.

En el apartado (1.3) se definió la ecuación del modelo de regresión logística

$$logit\, p(\Pi_1|\boldsymbol{x}) = log_e\left(\frac{p(\Pi_1|x)}{1-p(\Pi_1|x)}\right) = \beta_0 + \boldsymbol{\beta}^t\boldsymbol{x}$$

Y en el apartado (2.37) y (2.38) se definió la ecuación de Máquina de Vectores de Soporte

$$\widehat{f(x)} = \hat{\beta}_0 + x^t\hat{\beta}$$

$$= \hat{\beta}_0 + \sum_{i \in sv}^{n} \hat{\alpha}_i y_i (x_i^t x_i)$$

Siendo la regla de clasificación lo siguiente: $C(x) = sign\{\widehat{f(x)}\}$

Por tanto en el apartado anterior notamos que en cuanto a predicción se trata el modelo de Máquina de Vectores de Soporte cuenta con los mejores indicadores de diagnóstico en sus cuatro variedades: Especificidad, Sensibilidad, predicción positiva y predicción negativa. Siendo la ecuación del Modelo de Máquina de Vectores de Soporte que se elige como modelo final concluyente.

	Sensibilidad	Especificidad	PV+	PV-
Modelo Logístico	73.80%	70.30%	54.60%	84.70%
Modelo SVM	**75.00%**	**74.50%**	**58.80%**	**86.00%**

En el contexto del presente estudio y ámbito de análisis notamos la superioridad de capacidad de pronóstico de la técnica SVM. Además no solamente en detectar clientes que puedan caer en incumplimiento de pago (clientes con un perfil poco deseable para la entidad financiera) sino también clientes que tiene un perfil de buenos pagadores y para la entidad financiera implicara un crecimiento en la cartera de créditos de manera y una buena captación de clientes. Esto último finalmente implicará para la entidad una mejor oportunidad de negocio gracias a la mejora técnica en los indicadores PV+ y PV-.

Adicionalmente si analizamos los resultados desde una perspectiva de gestión del riesgo de la entidad financiera podemos afirmar que los pronósticos a futuras evaluaciones a clientes en cuanto a riesgo crediticio con una regla de clasificación con el SVM tendremos un mejor control de los casos de incumplimiento de pago.

2. Pesos de las Variables

Si bien la asignación de evidencia de la información (pesos de las variables predictivas), en ambos modelos sonsimilares, el factor crítico de mejora es justificada por el procedimiento Kernel.

Evidenciamos que ambos modelos de la propensión de riesgo de crédito de los clientes de la entidad financiera muestra evidencias de información similares, como se mostró en el apartado de variables que mitigan el riesgo crediticio:

Variables	SVM PESOS	LOGIT PESOS
Ant_trab_2	304.0	0.63
DURACION_CRED	-2,761.4	-0.04
EMPRESA_PROPIA	438.3	1.26
Est_Cta_1	1,465.7	0.46
Est_Cta_2	1,487.0	0.71
Est_Cta_3	3,404.2	1.57
Hist_cred_2	21,739.4	0.99
Hist_cred_3	12,747.1	1.06
Hist_cred_4	20,613.8	1.61
OTROS_PLAN	-1,060.2	-0.60
Pr_ct_aho_3	778.5	1.10
Pr_ct_aho_4	1,188.9	0.72
PROP_AUT_NUEVO	-1,757.2	-0.91
PROP_AUT_USADO	1,015.4	1.35
PROP_EDUCACION	-822.4	-0.89
RESID_ALQUILADA	-1,695.4	-0.53
SIT_CIVIL_SOLT	670.2	0.38
TASA_INGR_DISPONIBLE	-1,092.7	-0.24

Se puede interpretar que el riesgo relativo estimado por ambos modelos de las variables son similares, es decir ambos modelos tiene semejante interpretación, y solo la mejora en indicadores de capacidad predictiva es explicable y la mejora en los pronostico será explicada por la aplicación de la función Kernel en el algoritmo de Máquina de Vectores de Soporte.

Precisamente la función kernel elegida es la función de base Radial. Además como ya habíamos establecido: **C=1 Gamma=0.125**

Podemos concluir en propias palabras que los modelo Regresión logística al igual que el modelo SVM cuantifican ponderan el nivel de capacidad de predicción de riesgo de crédito de manera individual y de manera conjunta de las variables. Según nuestros resultados notamos que la función Kernel (una vez modelada) identifica una nivel de información de riesgo crediticio antes no contemplada, que nos brinda la mejora en los indicadores de capacidad predictiva en el contexto de la evaluación de crédito riesgo – persona y para una base de datos de clientes de una entidad financiera.

VIII. Bibliografía

[1] **Raymond Anderson (2007),**The Credit Scoring ToolkitOxford University Press

[2]Sanjoy Kumar Sinha (1997),Sequential Application of Multivariate OutliersDalhousie University Nova Scotia

[3] Alan Julian Izenman (2008), *Modern* MultivariateStatisticalTechniquesRegressionClassification and Main fold Learning. Editorial Springer

[4] Julian Faraway, Generalized Linear, Mixed Effects and Nonparametric Regression ModelsEditorial Chapman & Hall

[5] Kin Keung Lai, Lean Yu,Ligang Zhou, andShouyang Wang CHINA, Credit Risk Evaluation with Least Square Support Vector Machine

[6] Chen Lung Huang Mu Chen Chen CHINA, Credit scoring with a data mining approach based on support vector machines

[7] DebasishBasakSrimanta Pal INDIA, Support Vector Regression

Extending the Linear Model with R

[8] Lyhn Thomas, Consumer Credit Risk Models via Machine Learning Algorithm

[9] Norman Draper, Applied linear Regression

[10] **Douglas Montgomery**, Introducción al análisis de regresión Lineal, Mexico 2004

[11] **Consultora MAYSA Matemática Aplicada y Soluciones Analíticas,** Analytics aplicado a la industria bancaria

[12] **Xin Yan Xiao Gang Su (2008)**, Linear Regression Analysis Theory and Computation

[13] **Steve R. Gunn (1998)**, Support Vector Machine for Classification and Regression *University of Southampton.*

[14] **Gunter Loffle Peter Posch (2007)**, Credit risk modeling using Excel and VBA John Wiley & Sons

[15] **Stephen Nash and Ariela Safer (1996)**, Linear and Nonlinear Programming McGraw-Hill

[16] **Mitacc Máximo y Peche Carlos (1984)**, Calculo III Universidad Nacional Mayor de San Marcos Lima-Perú

[17] **Christopher M. Bishop (2008)**, Pattern Recognition and Machine Learning Springer United kingdom

[18] **Randall Matignon (2008)**, Data Mining Using SAS Enterprise Miner

[19] **MamdouhRefaat (2008)**, Data Preparation for Data Mining Using SAS - Morgan Kaufman

[20] **Tim Arnold y Analytics Solution SAS Division (2004)**, SAS/STAT 9.1 User Guide

[21] Lyn C. Thomas, David B. Edelman y Jonathan N. Crook (2002), Credit Scoring and Its Applications SIAM

[22] Edgardo Venero Orozco (2008), Evaluación del Riesgo de Crédito Lima-Perú

[23] Olivia Parr Rud (2001), Data mining Cookbook Modeling for Marketing, Risk and Customer Relationship Management Jhon Wiley &Sons,Inc.

[24] Naeem Siddiqi (2006), Credit Risk ScoreCards Developing and implementing intelligent credit Scoring Jhon Wiley & Sons, Inc.

[25] Raymon Anderson (2007), The Credit Scoring Toolkit Oxford University Press.

Publicaciones:

[P-1]Salvador Rayo Cantón – Juan Lara Rubio – David Camino Blasco (2010), Un Modelo de Crédito Scoring para instituciones de micro finanzas en el marco de Basilea II

[P-2]Luis Molinero (2001), Asociación de la Sociedad Española de Hipertensión Bioestadística - **www.seh-lelha.org/rlogis1.htm**

Recursos en RED:

[C-1]Curso

http://www.ece.uah.edu/courses/ee448/chapter4.pdf

[C-2] Guide credit scoring in R pdf

http://cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf

[C-3] Support Vector Machine in R pdf

http://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf

[C-4] Guía de R en Machine Learning

http://cran.r-project.org/web/views/MachineLearning.html

[C-5] Package R epid

http://cran.r-project.org/web/packages/Epi/Epi.pdf

[C-6] Package R ROCR

http://rocr.bioinf.mpi-sb.mpg.de/ROCR.pdf

[C-7] Package e1071

http://cran.r-project.org/web/packages/e1071/e1071.pdf

[C-5] Matrix Algebra

http://www.ece.uah.edu/courses/ee448/chapter4.pdf

[C-6] SVM programación R

http://stackoverflow.com/questions/7390173/svm-equations-from-e1071-r-package

http://r.789695.n4.nabble.com/SVM-coefficients-td903591.html

http://www.rcreditscoring.com/binning-continuous-variables-in-r-the-basics/

IX. Anexos

VIII.1 Código de procesamiento

################ R
####################

Los códigos aplicados para este trabajo fueron los siguientes

```
data<-read.table("clipboard",header=T)
d=sort(sample(nrow(data),nrow(data)*0.7))
train<-data[d,]
test<-data[-d,]
train<-subset(train, select=-d)
nrow(train)
nrow(test)
library(e1071)
trainx<- subset(train, select = -Default)
trainy<-train$Default
testx<- subset(test, select = -Default)
testy<-test$Default
MSVM<-svm(trainx,trainy)
predtrainsvm<- predict(MSVM, trainx)
library(Epi)
ROC(form=trainy~predtrainsvm,plot="ROC",data=train)
```

VIII.2 Norma Euclidiana

Si tenemos el vector $\vec{x} = (x_1, x_2, .., x_p) \epsilon R^p$, definimos la norma Euclidiana como real no negativo $\sqrt{{x_1}^2 + {x_2}^2 + .. + {x_p}^2}$ que denotaremos por el símbolo $|\vec{x}|$.

Tenemos así una función $\left|\overrightarrow{\ }\right| : R^p \rightarrow R$ que designamos la norma Euclidiana, la cual asigna a cada vector $\vec{x}\ \epsilon R^p$ un número real $|\vec{x}|$.

VIII.3 Método de estimación máxima verosimilitud para la regresión Logística

El método de máxima verosimilitud (mv) consiste en estimar parámetros de modo tal que la probabilidad de observar *y* sea lo máximo posible $\Rightarrow$ maximizar la Función de Verosimilitud.

Si partimos de un modelo inicial lineal, dado por:

$$Y_i = \beta_1 + \beta_2 X_i + U_i$$

Asumamos que y se distribuye como una normal, con media $\beta_1 + \beta_2 x_i$ y varianza σ^2, es decir:

$$Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$$

Si recordamos de nuestros cursos de estadística, la función de distribución normal de *y* viene dada por

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\tfrac{1}{2}\frac{(Y-\mu)^2}{\sigma^2}\}$$

Donde μ es la media de y.

Para $y_1, y_2, \ldots, y_n$ independientes e idénticamente distribuidas, la función de probabilidad conjunta viene dada por el producto de las funciones de probabilidad marginales

$$f(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} \exp\{-\tfrac{1}{2} \sum \frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{\sigma^2}\}$$

FV (FUNCION DE VEROSIMILITUD)

La cual constituye nuestra función objetivo.

Para ello debemos:

1) simplificar la expresión anterior a través de transformación logarítmica y proceder a derivar

$$\text{MAXIMIZAR(Ln FV)} = -n/2\ \text{Ln}\sigma^2 - n/2\ \text{Ln}\ (2\pi) - \tfrac{1}{2} \sum \frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{\sigma^2}$$

2) igualar derivadas a cero y resolver sistema:

$$\left.\begin{aligned} \frac{\partial Ln\ fv}{\partial \beta_1} &= -1/\sigma^2 \sum (Y_i - \beta_1 - \beta_2 X_i)(-1) = 0 \\ \frac{\partial Ln\ fv}{\partial \beta_2} &= -1/\sigma^2 \sum (Y_i - \beta_1 - \beta_2 X_i)(-X_i) = 0 \end{aligned}\right\} = \hat{\beta}^{MCO}$$

$$\frac{\partial Ln\ fv}{\partial \sigma^2} = -n/2\sigma^2 + 1/2\sigma^4 \sum (Y_i - \beta_1 - \beta_2 X_i)^2 = 0$$

Volvamos ahora el modelo de regresión logística

$$L_i = \text{Ln}\left(\frac{P_i}{1 - P_i}\right) = X_i\beta + U_i = \beta_1 + \beta_2 X_i + U_i$$

CUYA ESTIMACION REQUIERE NO SOLO LOS VALORES DE *X* SINO TAMBIEN LOS DE *L*. La estimación del modelo depende del tipo de datos de que se disponga:

A) DATOS INDIVIDUALES:

En este tipo de datos no puede aplicarse MCO debido a que la variable dependiente carece de sentido:

$$L_i = \text{Ln}\left(\frac{P_i}{1 - P_i}\right) = \begin{cases} \text{Ln}\ (1/0) & \text{SI OCURRE EL EVENTO} \\ \text{Ln}\ (1/0) & \text{SI NO OCURRE EL EVENTO} \end{cases}$$

En este caso se recurre al método de máxima verosimilitud[32] :

De nuevo, para una muestra aleatoria de n observaciones, la probabilidad conjunta $f(y_1, y_2, \ldots y_n)$ viene dada por:

$$f(Y_1, Y_2, \ldots Y_n) = \prod_{i=1}^{n} fi(Yi) = \prod_{1}^{n} P_i^{Yi}(1-Pi)^{1-Yi}$$

FUNCION MAXIMO-VEROSIMIL

Cuyo logaritmo se traduce en:

$$\mathbf{Lnf(Y_1, Y_2, \ldots Y_n)} = \sum_{i=1}^{n} [Y_i Ln(P_i) + (1-Y_i) Ln(1-P_i)] = \sum_{i=1}^{n} [Y_i Ln(P_i) - Y_i Ln(1-P_i) + Ln(1-Pi)]$$

$$= \sum_{i=1}^{n} [Y_i LnP_i/(1-P_i)] + \sum_{1} Ln(1-Pi)$$

$$\mathbf{Lnf(Y_1, Y_2, \ldots Y_n)} = \sum_{i=1}^{n} Y_i(\beta_1 + \beta_2 X_i) + \sum_{i=1}^{n} Ln(1+e^{\beta_1+\beta_2 Xi})$$

Diferenciando la función maximoverosimil con respecto de β se obtiene solución no lineal en parámetros.

B) MINIMOS CUADRADOS CON DATOS AGRUPADOS (OBSERVACIONES REPETIDAS):

Con observaciones repetidas pi puede estimarse a partir de la frecuencia relativa para cada valor de x:

$$\hat{P}_i = n_i / N_i$$

Con n_i número de observaciones para las que Y_i=1 dado un cierto valor de Xi Y N_i el total de observaciones (por ejemplo, cuántas familias de ingreso X* poseen vivienda, con respecto al total).

[32] mv consiste en estimar los parámetros tal que **la probabilidad de observar y dado x sea lo más alta posible (máxima).** Este es generalmente un método para **muestras grandes**, por lo que las propiedades de los estimadores son **asintóticas**.

$$\hat{L}_i = \mathbf{Ln}\left(\frac{\hat{P}_i}{1-\hat{P}_i}\right) = \hat{\beta}_1 + \hat{\beta}_2 \mathbf{X_i}$$

Los residuos del modelo así estimado se distribuyen $U_i \sim N\ [0,\ 1/(N_iP_i(1-P_i)]$ Puede ser estimado por MCO? Note que los residuos son heterocedasticos (su varianza depende de pi), por lo que debe recurrirse a mcp, como se indicara inicialmente.

EVALUANDO EL MODELO:en este tipo de modelos es más importante el signo, significancia y significado de los coeficientes, antes que la bondad de ajuste.

- En estimación mv, siendo que se habla de propiedades asintóticas (muestras grandes), la significancia estadística se prueba a través de la normal estándar (z) en lugar dela tradicional t.

- El coeficiente de determinación r^2 utilizado en mc no tiene sentido aquí, por lo que se recurre a otros criterios, generalmente basados en distribuciones chi-cuadrado.

- **R^2McFadden**= $1 - [\ln(\hat{L}_{mv}) / \ln(\hat{\beta}_1)]$, equivale al cociente del logit no restringido (todas las variables incluidas) y restringido (solo el intercepto es incluido).

- **R^2cuenta**= $\left(\frac{\text{numero prediccion es correctas}}{\text{numero total de observacio nes}}\right)$, para ello se consideran como 1 las probabilidades mayores que 0.5 y como 0 las inferiores a 0.5.

- **SIGNIFICACIÓN CONJUNTA A TRAVÉS DE LA RAZÓN DE VEROSIMILITUD (EQUIVALENTE A LA PRUEBA F):**

Hipótesis planteada: H_o: $\beta_2 = \beta_3 = \ldots = \beta_k = 0$

H_1: al menos uno es distinto de cero

Estadístico de prueba: RV = λ = -2 ln(L) = -2 ln $(\hat{\beta}_1 / \hat{L}_{mv}) \sim X^2_{k-1}$

- **TEST DE HOSMER Y LEMESHOW:** compara frecuencias muestrales observadas con las previstas por el modelo.

  Hipótesis planteada: ho: el modelo ajusta bien

  h_1: mal ajuste del modelo

Estadístico de prueba: $x^2 = \sum \frac{(O_i - n_i p_i)^2}{n_i p_i (1 - p_i)} \sim x^2_{g-1}$

Donde o_i es el número de eventos observados en el grupo *i*; n_i el tamaño del grupo *i*; p_i es la probabilidad estimada de un evento en el grupo *i* y g es el númerode grupos.

VIII.4 Método de Krush-Kuhn-Tucker para la optimización cuadrática

Ver referencia [13] Steve R. Gunn página 159.

VIII.5 Multiplicadores de Lagrange

[14]

El matemático Francés Lagrange ideó un procedimiento para el problema de determinar máximos y mínimos de funciones de varias variables sujetas a condiciones laterales o restricciones. Este método se conoce como Método de los Multiplicadores de Lagrange.

A) Caso de 2 variables:

Supongamos que se desea optimizar una función de dos variables $z = f(x,y)$, llamada función objetivo; cuyas variables no son independientes si no sujetas a una condición lateral, llamada restricción que se expresa como:

$$g(x,y) = 0$$

En este caso utilizamos una nueva variable λ y creamos una nueva función.

$$F(x,y,\lambda) = f(x,y) + \lambda g(x,y)$$

Luego, hallamos los puntos críticos de esta nueva función; es decir resolvemos el sistema de ecuaciones simultáneas.

$$\begin{cases} D_1F(x,y,\lambda) = D_1f(x,y) + \lambda D_1g(x,y) = 0 \\ D_2F(x,y,\lambda) = D_2f(x,y) + \lambda D_2g(x,y) = 0 \\ D_1F(x,y,\lambda) = g(x,y) = 0 \end{cases}$$

De acuerdo a la naturaleza del problema se decide cuáles de los puntos críticos corresponde a un máximo o mínimo relativo.

Observación: Si se impone varias restricciones, el método de Lagrange puede ser extendido usando varios multiplicadores. En partículas si queremos encontrar puntos críticos de la función $z(x,y)$, sujeta a las dos condiciones laterales $g(x,y) = 0$, $h(x,y) = 0$, encontramos los puntos críticos de la función en la nueva función.

$$F(x,y,\lambda_1,\lambda_2) = f(x,y) + \lambda_1 g(x,y) + \lambda_2 h(x,y)$$

Máximo Mitacc([14] Cálculo III pág.112-113).

Para desarrollar satisfactoriamente el tema elegido, se presentarán Progresivamente los contenidos de tópicos básicos previos en la siguiente secuencia:

VIII.6 Resultados de Procesamiento de datos

VIII.6.1 Clústeres en la detección de outlier multivariado según K-means

Los cluster 50 detectado fueron los siguientes:

ClusterIndex	COUNT	PERCENT	_ClusterIndex_	COUNT	PERCENT
1	88	8.8	**26**	10	1
2	4	0.4	**27**	8	0.8
3	3	0.3	**28**	20	2
4	6	0.6	**29**	1	0.1
5	18	1.8	**30**	12	1.2
6	5	0.5	**31**	4	0.4
7	50	5	**32**	2	0.2
8	17	1.7	**33**	10	1
9	5	0.5	**34**	38	3.8
10	3	0.3	**35**	2	0.2
11	83	8.3	**36**	1	0.1
12	19	1.9	**37**	11	1.1
13	15	1.5	**38**	42	4.2
14	34	3.4	**39**	3	0.3
15	1	0.1	**40**	58	5.8
16	52	5.2	**41**	13	1.3
17	84	8.4	**42**	78	7.8
18	3	0.3	**43**	17	1.7
19	4	0.4	**44**	3	0.3
20	10	1	**45**	5	0.5
21	2	0.2	**46**	5	0.5
22	62	6.2	**47**	3	0.3
23	2	0.2	**48**	2	0.2
24	2	0.2	**49**	10	1
25	69	6.9	**50**	1	0.1

Aquellos cluster con pocas observaciones y de distancia más grande serán considerados outlier Multivariados. El punto de corte fue de 0.4% de los valores contenidos en el cluster, estos casos fueron denominados outliers multivariados.

VIII.6.2 Detección multicolinealidad VIF

Variable	DF	Estimado	Error Std	T Value	Pr > |t|	Tolerancia	VIF
Hist_cred_2	1	0.20	0.07	2.7	0.008	0.116	8.64
Hist_cred_4	1	0.31	0.07	4.1	<.0001	0.139	7.22
RESID_PROPIETARIO	1	-0.07	0.07	-1.0	0.325	0.154	6.48
RESID_ALQUILADA	1	-0.15	0.07	-2.0	0.052	0.192	5.21
PROP_RADIO_TV	1	-0.01	0.06	-0.2	0.869	0.202	4.94
PROP_AUT_NUEVO	1	-0.14	0.06	-2.3	0.024	0.226	4.42
PROP_MUEBLES	1	-0.02	0.07	-0.4	0.726	0.245	4.08
Hist_cred_3	1	0.21	0.08	2.6	0.009	0.308	3.25
Niv_Ed_2	1	0.00	0.06	0.0	0.987	0.326	3.07
PROP_AUT_USADO	1	0.09	0.07	1.3	0.200	0.337	2.97
Niv_Ed_3	1	-0.01	0.05	-0.3	0.763	0.341	2.93
NO_PROPIETARIO_RESID	1	-0.09	0.06	-1.5	0.141	0.355	2.82
PROP_REFINAC	1	-0.01	0.07	-0.2	0.837	0.360	2.77
Hist_cred_1	1	0.02	0.09	0.2	0.856	0.394	2.54
Ant_trab_3	1	0.07	0.05	1.6	0.113	0.401	2.50
MONTO_CRED	1	0.00	0.00	0.2	0.865	0.401	2.49
Ant_trab_1	1	0.07	0.04	1.7	0.085	0.450	2.22
DURACION_CRED	1	-0.01	0.00	-4.6	<.0001	0.492	2.03
PROP_EDUCACION	1	-0.18	0.08	-2.2	0.026	0.501	2.00
Ant_res_2	1	0.06	0.03	1.8	0.081	0.552	1.81
Ant_trab_2	1	0.12	0.05	2.7	0.007	0.556	1.80
Est_Cta_3	1	0.28	0.03	8.1	<.0001	0.561	1.78

Variable	DF	Estimado	Error Std.	T Value	Pr > |t|	Tolerancia	VIF
Ant_trab_4	1	-0.01	0.07	-0.1	0.934	0.570	1.75
NUM_CREDITS	1	-0.03	0.03	-1.1	0.287	0.590	1.69
Est_Cta_1	1	0.10	0.04	2.8	0.006	0.611	1.64
SIT_CIVIL_SOLT	1	0.07	0.03	2.1	0.037	0.614	1.63
EDAD	1	0.00	0.00	0.5	0.633	0.670	1.49
TASA_INGR_DISPONIBLE	1	-0.03	0.01	-2.4	0.015	0.689	1.45
Niv_Ed_1	1	0.12	0.10	1.2	0.239	0.690	1.45
Ant_res_1	1	0.05	0.04	1.2	0.224	0.711	1.41
Ant_res_3	1	0.11	0.04	2.5	0.013	0.714	1.40
VEHICULO_PROP	1	0.04	0.03	1.5	0.133	0.766	1.31
ESTADO_PROPIEDAD	1	0.02	0.03	0.8	0.436	0.777	1.29
SIT_CIVIL_CASADO	1	0.03	0.05	0.6	0.569	0.781	1.28
Est_Cta_2	1	0.20	0.06	3.5	0.001	0.801	1.25
Pr_ct_aho_4	1	0.11	0.04	3.0	0.003	0.838	1.19
NUM_DEPENDENTS	1	-0.04	0.04	-1.0	0.329	0.840	1.19
SIT_CIVIL_DIV	1	-0.06	0.06	-1.0	0.324	0.847	1.18
OTROS_PLAN	1	-0.06	0.04	-1.8	0.081	0.866	1.15
Pr_ct_aho_1	1	0.04	0.04	0.9	0.359	0.873	1.15
EMPRESA_PROPIA	1	0.16	0.07	2.3	0.021	0.889	1.13
GUARANTIA	1	0.16	0.06	2.7	0.007	0.890	1.12
Pr_ct_aho_2	1	0.08	0.05	1.6	0.123	0.892	1.12
Pr_ct_aho_3	1	0.15	0.06	2.5	0.013	0.908	1.10
CO_APPLICANT	1	-0.09	0.07	-1.3	0.205	0.936	1.07
Intercept	1	0.59	0.17	3.5	0.001	.	-