UNIVERSIDAD NACIONAL DE INGENIERIA

FACULTAD DE INGENIERIA ECONOMICA, ESTADISTICA Y CIENCIAS SOCIALES

ESCUELA PROFESIONAL DE INGENIERIA ESTADISTICA

Design of a Statistical System for Predicting the Quantity and Characteristics of Non-Performing Loans Using the Method of Bootstrap with Crossed Validation, and Support Vector Machines

TESIS

Para optar el Título Profesional de Ingeniero Estadístico

Por modalidad de Tesis

Elaborado por:

REYES RAMIREZ GIANCARLOS

Lima – Perú

2016

ÌNDICE

Dedicatoria

Este trabajo es dedicado con mucho esmero y entrega, a mi familia y a todas aquellas personas, que siempre estuvieron brindándome su apoyo de manera incondicional. Es parte del esfuerzo, que significo seguir estos 5 años de lucha constante, en donde se tuvieron que superan diversas dificultades. Y sobre todo para demostrar lo importante que resulta ser la estadística en el mercado financiero.

Agradecimientos

Agradecer en primer lugar a Dios, por darme la oportunidad de haber podido culminar una gran etapa en mi vida, y ser así un gran profesional de mucho éxito.
A mis padres por el esfuerzo y sacrificio, que significó mucho en mí día a día, lo que me motivo a ser cada vez mejor y luchar por mis sueños.

A la Universidad Nacional de Ingeniería por haberme acogido en sus aulas durante estos 5 años de estudios.

A los profesores de la Escuela de Ingeniería Estadística, que se esforzaron día a día por transmitirnos sus enseñanzas y experiencias.

RESUMEN

Este trabajo de investigación, tiene como objetivo fundamental determinar el mejor método de remuestreo, enla construcción del modelo de Maquina de

Vectores de Soporte, para la predicción de la morosidad de los clientes del Banco Banwest, para lo cual se analizaran distintos escenarios que nos permitan analizar el rendimiento del modelo de Maquina de Vectores de Soporte, a través de información histórica del banco Banwest correspondiente al periodo de julio y setiembre del 2016.
Dado que resulta importante para las entidades financieras contar con modelos de predicción en cuanto a la morosidad, se busca plantear una manera opcional de predecir dicha variable. Por ello, mediante esta aplicación, nos permitirá mostrar que el método Validación Cruzada puede ser más potente que el método de Bootstrap, y esto se debe que presenta mayores indicadores de GINI y sensibilidad.

Palabras Claves

Máquina de Vectores de Soporte
ValidaciónCruzada
Remuestreo
Morosidad

ABSTRACT

This research has as its main objective to determine the best method of resampling, construction model vector machine Support for predicting late payments Banwest bank customers, for which we will analyze different scenarios that allow us to analyze the performance of the Model of Support Vectors, through historical information of Banwest bank corresponding to the period of July and September of 2016.
Given that it is important for financial institutions to have predictive models for late payments, it is proposed to propose an optional way of predicting said variable. Therefore, through this application, it will allow us to show that the Cross Validation method can be more powerful than the Bootstrap method, and this is due to the fact that it has higher GINI and sensitivity indicators.

KeysWords

Support Vector Machine
Cross Validation
Bootstrap
Late Payment

CAPÍTULO I

ANTECENDENTES

1.1 Investigaciones

A Comparison of Classification/Regression Trees and Logistic Regression in Failure Models

Este estudio compara el rendimiento predictivo de una metodología no paramétrica, es decir Clasificación / Regresión árboles (CART), en contra de regresión logística tradicional (LR) mediante el empleo de un vasto conjunto de cuentas de pares emparejados de las empresas más pequeñas, conocidas como micro-organismos, del Reino Unido para el período de 1999 a 2008, que incluye las variables financieras, no financieras y macroeconómicas. Nuestros resultados muestran que la CART supera el enfoque estándar en la literatura, LR.

Los factores que determinan la calidad de la cartera crediticia de las entidades microfinancieras de la Amazonía peruana en el periodo 2008-2011.

El presente estudio de investigación es importante y relevante porque contribuirá a que las microfinancieras alcancen niveles de gestión de calidad de cartera que sean iguales o mejores que los de las empresas líderes del sector Edyficar y Mibanco. Las microfinancieras valoran la calidad de su cartera de créditos y si la investigación demuestra que las variables a trabajar en el modelo son relevantes, las empresas microfinancieras enfocarán su estrategia en ellas. Siendo importante la inclusión financiera.

Construcción de modelos de predicción con las máquinas de clasificación y agrupación de vectores soporte-lanzado en la puntuación de crédito.

En los últimos años, los investigadores han aplicado la SVM basada en la predicción de la puntuación de crédito, y los resultados han demostrado que para ser eficaz. En este estudio, se seleccionaron dos conjuntos de datos de crédito del mundo real en la Universidad de California Irvine Machine Learning Repositorio. SVM y un nuevo clasificador, la clasificación de la agrupación-lanzado (CLC), se utilizaron para predecir la exactitud de la puntuación de crédito. Las ventajas de usar CLC son que puede clasificar los datos de manera eficiente y sólo se necesita un parámetro tiene que ser decidido. En esencia, los resultados muestran que la CVX es mejor que la SVM. Por lo tanto, la CVX es una herramienta eficaz para predecir la puntuación de crédito.

CAPÍTULO II

PLANTEAMIENTO DEL PROBLEMA

2.1 Descripción del problema

A raíz de la crisis de la crisis económica y debido al incremento de los índices de morosidad y la cartera vencida, las instituciones financieras se vieron obligadas a redefinir los componentes de su modelo de cobranzas con el propósito de aumentar el nivel de recuperación y reducir los costos.

En el contexto del Sistema financiero en el Perú, según el Reporte de Estabilidad Financiera Mayo 2016 emitido por el BCRP muestra la evolución de la Cartera Morosa por el tipo de créditos y empresas.

CUADRO N° 2.1
RATIO CARTERA MOROSA POR TIPO DE CRÉDITO (%)

	Sistema		**Banca**		**Financieras**		**CM**		**CRAC**	
	Mar. 15	**Mar. 16**	**Mar. 15**	**Mar. 16**	**Mar. 15**	**Mar. 16**	**Mar. 15**	**Mar. 16**	**Mar. 15**	**Mar. 16**
Total	**4.1**	**4.2**	**3.6**	**3.8**	**7.8**	**7.7**	**8.2**	**8.3**	**9.7**	**10.4**
Total Empresas	**4.4**	**4.3**	**3.7**	**3.7**	**9.1**	**8.7**	**9.9**	**9.9**	**10.1**	**10.8**
Corporativos	0.1	0	0.1	0	0	0	0	0	0	0
Grandes Empresas	1.3	1.6	1.3	1.6	16.6	15.9	8.1	2.9	21.6	0
Medianas Empresas	7.4	8.2	7.2	8	6.7	5.8	9.4	12.3	8	9.6
Pequeñas Empresas	11.9	11.4	11.9	11.8	11.3	11.1	11.6	10.9	14.4	14.6
Microempresas	6.7	6	4.8	3.8	6.7	6	7.6	7.8	7.9	9
Total de Hogares	**3.5**	**4**	**3.3**	**3.9**	**6**	**6.6**	**3.1**	**3.6**	**6.2**	**7.2**
Consumo	4.7	5.2	4.7	5.1	6	6.7	3.4	3.9	6.2	7.2
Hipotecarios	2	2.6	1.9	2.6	5.3	4.9	2.3	2.5	3.9	0

Fuente:Reporte de Estabilidad Financiera BCRP Mayo 2016
Elaboración: Propia

Con
desaceleraron la tasa de crecimiento anual de los créditos hipotecarios (de 10,1% a 7,1%) y en los créditos a las grandes empresas (de 11,6% a -2,3%).

Asimismo, los bancos han venido reevaluando sus políticas de otorgamiento de créditos a las empresas, sobre todo en las medianas empresas.

Es importante señalar que, en los últimos doce meses, el saldo de la cartera morosa del sistema financiero creció 15,4% anual en moneda nacional y 3,3% anual en moneda extranjera. En esta última moneda, la cartera morosa en los créditos hipotecarios registra una tasa de crecimiento anual de 19,9%27 y en las grandes empresas, 6,1%. En cambio, se observa una reducción de los créditos morosos en moneda extranjera en los segmentos de empresas corporativas (65,6%), de pequeñas empresas (13%) y de microempresas (20,2%).

CUADRO N° 2.2
RATIO CARTERA MOROSA POR TIPO DE CRÉDITO DEL SISTEMA FINANCIERO (%)

	MN		ME		Total	
	Mar. 15	Mar. 16	Mar. 15	Mar. 16	Mar. 15	Mar. 16
Total	**4.6**	**4.3**	**3.2**	**4**	**4.1**	**4.2**
Total Empresas	**5.4**	**4.5**	**3.3**	**3.9**	**4.4**	**4.3**
Corporativos	0	0	0.1	0	0.1	0
Grandes Empresas	0.8	0.9	1.7	2.4	1.3	1.6
Medianas Empresas	8.7	8.1	6.4	8.3	7.4	8.2
Pequeñas Empresas	11.7	10.7	14.1	22	11.9	11.4
Microempresas	6.7	5.9	8.1	11.1	6.7	6
Total de Hogares	**3.6**	**4**	**2.9**	**4.3**	**3.5**	**4**
Consumo	4.8	5.1	4.2	5.8	4.7	5.2
Tarjetas de crédito	5.1	6.1	3.3	2.9	5	5.9
Préstamos	4.6	4.5	4.6	7.6	4.6	4.7
Hipotecarios	1.7	2.2	2.5	3.8	2	2.6

Fuente:Reporte de Estabilidad Financiera BCRP Mayo 2016
Elaboración: Propia

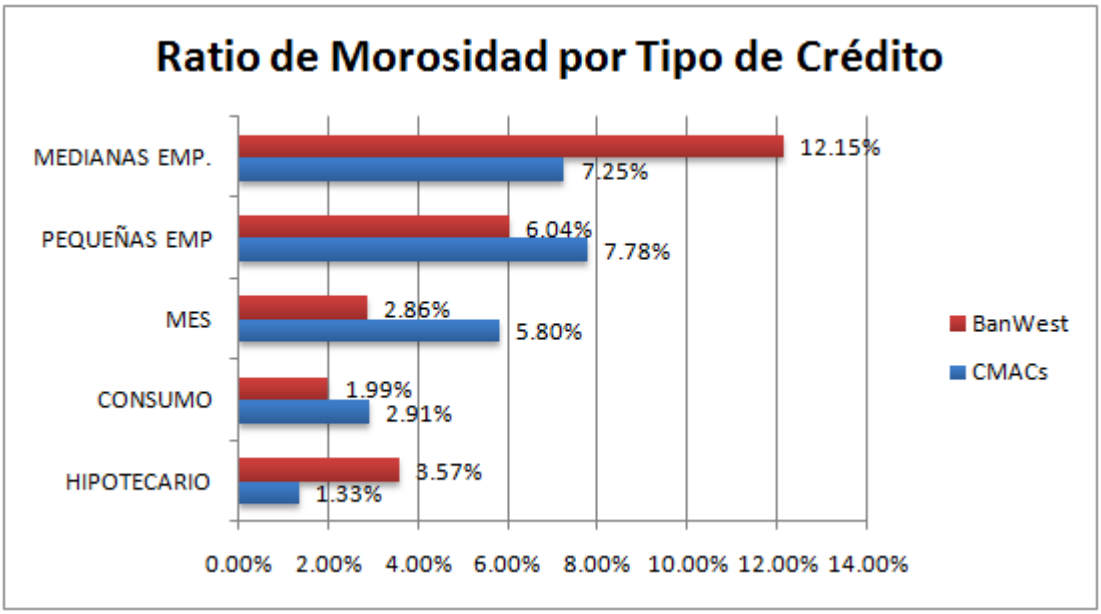
Exigiendo a las entidades financierasnuevos modelos que permitan clasificar la cartera, para ofrecer productos que se ajustaran mejor a las necesidades de los clientes a través de los canales más apropiados.

Actualmente en el BanWest la cartera de colocaciones se incrementó en 5.75%, y esto se debe principalmente al dinamismo de los desembolsos de créditos a pequeña y micro empresa cuyos segmentos representan el 84.51% de la cartera total.

Logrando así una participación del 30.54% en el mercado de las microempresas (Equilibrium- Clasificadora de Riesgo). Asimismo al analizar la mora por el tipo de crédito se observa que los ratios más altos para el BanWest corresponden a los segmentos de mediana y pequeña empresa.

Por lo que se deberá analizar la calidad de Cartera Morosa, y enfatizar las estrategias en aquellosclientes previa clasificación que permitan una rápida recuperación de los créditos.

GRÁFICO 2.1
RATIO DE MOROSIDAD POR TIPO DE CRÉDITO



Fuente:Equilibrium Clasificadora de Riesgo Mayo 2016
Elaboración: Propia

Por otro lado, las entidades financieras al otorgar un crédito, corren el riesgo que el cliente se atrase en el pago y eventualmente no pague, ya sea el crédito completo o una parte de éste. Para prevenir la ocurrencia de crisis financieras, la tendencia actual es adoptar el acuerdo de Basilea II. Este acuerdo se basa en tres pilares: requerimientos de capital mínimos, revisión supervisora y disciplina de mercado.
En particular, el primer pilar obliga a los bancos a reservar capital dado el nivel de riesgo de incumplimiento de sus carteras de clientes.

Asimismo la administración del riesgo crediticono era la apropiada, dado que se pre-aprobaban a clientes propensos a caer en situación de morosa, lo cual ubicaba a la cobranza en un miedo riesgoso. Mientras se enfocaban en las ventas, se descuidaba la cobranza y recuperación.

Por lo que la gestión de procesos se convirtió en un tema fundamental para medir la rentabilidad de las actividades de cobranza y ejecutar estrategias más rentables para cada tipo de clientes.

2.2 Formulación del problema

2.2.1 Problema general

¿Cuál es el mejor método de remuestreo, en la construcción del modelo de Máquina de Vectores de Soporte para la Predicción de la Morosidad de los clientes del Banco Banwest?

2.2.2 Problemas específicos

¿Cuáles son los factores que más influyen en la morosidad de los clientes del BancoBanWest?

¿Cuál es el escenario bajo el cual el modelo de Maquina de Vectores de Soporte, para la Predicción de la Morosidad en el Banwest, presenta un mejor rendimiento mediante el método Bootstrap?

¿Cuál es el escenario bajo el cual el modelo de Maquina de Vectores de Soporte, para la Predicción de la Morosidad en el Banwest, presenta un mejor rendimiento mediante el método Validación Cruzada?

2.3 OBJETIVOS DE LA INVESTIGACIÓN

2.3.1 Objetivo general

Determinar el mejor método de remuestreo, en la construcción del modelo de Maquina de Vectores de Soporte para la predicción de la Morosidad de los clientes del banco Banwest.

2.3.2Objetivos específicos

Determinar mediante el algoritmo Boruta los factores que más influyen en la morosidad de los clientes del BanWest.

Encontrar el escenario bajo el cual el modelo de Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método Bootstrap, presente un mejor rendimiento.

Encontrar el escenario bajo el cual el modelo Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método de Validación Cruzada, presente un mejor rendimiento.

2.4 PLANTEAMIENTO DE HIPOTESIS

2.4.1 Hipótesis general

El método de Bootstrap, mejora en un 5% la sensibilidad del modelo de Maquina de Vectores de Soporte, en comparación al método Validación Cruzada, en la predicción de la Morosidad de los clientes del banco Banwest.

2.4.2 Hipótesis específicas

Los factores que más influyen en la morosidad haciendo uso del algoritmo de BORUTA de los clientes son: Deuda total, Días de Mora, Calificación SBS, Saldo Capital y Tipo de Crédito.

El escenario bajo el cual el modelo Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método Bootstrap, presente un mejor rendimiento, ocurre cuando el número de muestras es 40 y un Índice de Gini mayor a 85%.

El escenario bajo el cual el modelo Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método de Validación Cruzada, presente un mejor rendimiento, ocurre cuando el número de particioneses 10 y un Índice de Gini mayor a 85%.

2.5JUSTIFICACIÓN

Según el Reporte de Estabilidad Financiera del BCRP (Mayo 2016), El ratio de morosidad de los créditos del sistema financiero se elevó ligeramente los últimos doce meses, por la menor calidad de la cartera en los segmentos de medianas empresas, de consumo e hipotecario.
Además, el sistema financiero podría enfrentar un escenario de mayor morosidad en caso el crecimiento económico sea menor al esperado.

Ello se dio en un contexto en el cual las entidades financieras adoptaron medidas correctivas en su política crediticia para realizar una mejor selección de deudores y una recuperación más eficiente de los préstamos. Así, algunas entidades ajustaron sus modelos de scoring y fortalecieron sus áreas de cobranzas y de riesgos, mientras que otras se reorganizaron internamente para salvaguardar la calidad de la cartera.

Por tales motivos el BanWest deberá salvaguardar su calidad de cartera, mediante técnicas y herramientas de análisis, con el fin de llevar a cabo una evaluación más precisa del riesgo asociado a cada cliente. Con ello se pretenderá identificar las acciones más efectivas de cobranza y enfocar los esfuerzos hacia donde puede haber una mayor recuperación.

Por lo que se deberá clasificar de manera efectiva a los clientes de la Cartera Morosa, con el fin de elaborar estrategias diferenciadas que permitan la Recuperación de la Cartera.
Dichas estrategias dependerán del canal, de los productos que se ofrezcan y de los individuos encargados de administrar la cobranza de aquellas cuentas que tienen asignadas.

MATRIZ CONSISTENCIA

PROBLEMA	OBJETIVOS	HIPOTESIS	VARIABLES
GENERAL	GENERAL	GENERAL	
¿Cuál es el mejor método de remuestreo, en la construcción del modelo de Máquina de Vectores de Soporte para la Predicción de la Morosidad de los clientes del Banco Banwest?	Determinar el mejor método de remuestreo, en la construcción del modelo de Maquina de Vectores de Soporte para la predicción de la Morosidad de los clientes del banco Banwest.	El método de Bootstrap, mejora en un 5% la sensibilidad del modelo de Maquina de Vectores de Soporte, en comparación al método Validación Cruzada, en la predicción de la Morosidad de los clientes del banco Banwest.	MOROSIDAD
ESPECIFICOS	ESPECIFICOS	ESPECIFICOS	
¿Cuáles son los factores que más influyen en la morosidad de los clientes del BancoBanWest?	Determinar mediante el algoritmo Boruta los factores que más influyen en la morosidad de los clientes del BanWest.	Los factores que más influyen en la morosidad haciendo uso del algoritmo de BORUTA de los clientes son: Deuda total, Días de Mora, Calificación SBS, Saldo Capital y Tipo de Crédito.	**Dias de atraso, Deuda Total, Saldo Capital, Calificacion SBS,**
¿Cuál es el escenario bajo el cual el modelo de Maquina de Vectores de Soporte, para la Predicción de la Morosidad en el Banwest, presenta un mejor rendimiento mediante el método Bootstrap?	Encontrar el escenario bajo el cual el modelo de Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método Bootstrap, presente un mejor rendimiento.	El escenario bajo el cual el modelo Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método Bootstrap, presente un mejor rendimiento, ocurre cuando el número de muestras es 40 y un Índice de Gini mayor a 85%.	
¿Cuál es el escenario bajo el cual el modelo de Maquina de Vectores de Soporte, para la Predicción de la Morosidad en el Banwest, presenta un mejor rendimiento mediante el método Validación Cruzada?	Encontrar el escenario bajo el cual el modelo Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método de Validación Cruzada, presente un mejor rendimiento.	El escenario bajo el cual el modelo Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método de Validación Cruzada, presente un mejor rendimiento, ocurre cuando el número de particioneses 10 y un Índice de Gini mayor a 85%.	

CAPÍTULO III

MARCO TEORICO

3.1 MODELOS SUPERVISADOS

El aprendizaje estadísticosupervisadobusca aprender de los datos bajo la guía de una variable objetivo que se busca predecir.
El problema de clasificación es un típico problema de análisis supervisado, que cuenta con una variedad de técnicas disponibles tales como regresión logística, el algoritmo del vecino más cercano, los árboles de clasificación, las redes neuronales, las máquinas de vectores de soporte, y todo tipo de métodos lineales.
Asimismo se combinan estas técnicas con otras un poco más sofisticadas como boosting, subbaging, combinación de modelos óptima, y técnicas bayesianas entre otras.

3.2 Regresión Logística Binaria

Es una técnica estadística de "*dependencia*", un tipo especial de regresión que se utiliza para predecir y explicar una variable categórica binaria (dos grupos) en lugar de una medida dependiente métrica.

Donde la variable respuesta puede expresar de la siguiente forma:

$$Y_i = \begin{cases} 1, Prob(Y_i = 1) = p_i \\ 0\ , Prob(Y_i = 0) = 1 - p_i \end{cases}$$

La representación matemática del modelo es la siguiente:

$$Z_i = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Z_i: $variabledependientedelmodelo$: "Diabético" "No Diabético"
p_i: $Probabilidad\ de\ éxito$
β_i: $Parámetros\ del\ modelo\ a\ estimar.$
X_i : $Variables\ explicativas\ del\ modelo.$

3.2.1. Método de Estimación

Para modelos de regresión logística, los parámetros se estiman a través de los métodos de Máxima Verosimilitud.

Puesto que el modelo es no lineal, se necesita un algoritmo iterativo para esta estimación. El método iterativo que se aplica es el método de Newton-Raphson.

3.2.2. Contraste de significación

Para contrastar la significatividad global en los modelos de regresión logística, se utiliza el estadístico de razón de verosimilitud (RV). Normalmente se usa la prueba ómnibus.

Las hipótesis que se plantean:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_P = 0$
$H_1: Algún\ \beta_J \neq 0$

3.2.3. Contraste de bondad de ajuste de Hosmer y Lemeshow

Se dividen todos los casos en deciles basados en las probabilidades predichas, el primer decil se cuentan los casos con las probabilidades más altas, siendo el estadístico:

$$HL = \sum_{i=1}^{10} \frac{[O_i - N_i\pi_i]^2}{N_i\pi_i(1-\pi_i)}$$

$$HL\ se\ distribuye\ como\ una\ Chicuadrado\ con\ 8\ grados\ de\ libertad$$

$$O_i : Número\ de\ unos\ en\ el\ decil\ i-ésimo$$
$$\pi_i : Media\ de\ probabilidades\ en\ el\ decil\ i-ésimo$$
$$N_i : número\ de\ observaciones\ en\ el\ decil\ i-ésimo$$

Las hipótesis nula y alternante son:

H_0: Noexisten diferencias entre los valores observados y predichos
H_1: Existen diferencias entre los valores observados y predichos

Si rechazamos la H_0, implica que el modelo ajustado no es el adecuado.

3.2.4. Tablade Clasificación

La tabla de clasificación muestra la distribución de valores observados y estimados. Los valores estimados se obtienen a partir del modelo.

3.4. ÁRBOL DE DECISIÓN

Es un modelo de aprendizaje supervisado que permite la clasificación y predicción. Un árbol de decisión es una estructura que puede ser usada para dividir una gran cantidad de registros en conjuntos más pequeños de registros aplicando una secuencia de reglas simples de decisión. De este modo, los miembros de los conjuntos resultantes se vuelven más similares entre ellos. En otras palabras, los grupos que se forman tienen un mayor grado de “pureza" en función de la variable objetivo (predomina más uno solo de los valores). Sin perder generalidad, es posible suponer que la variable objetivo es binaria yCategórica.
Los algoritmos más destacados son CHAID, los algoritmos de poda (CART, ID3 y C5) y QUEST.

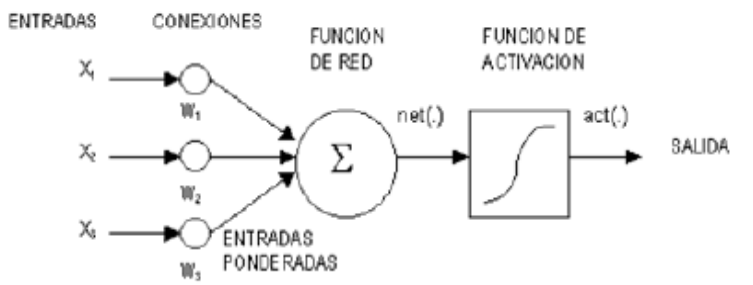
3.5. REDES NEURONALES ARTIFICIALES (RNAS)

Las RNAS son el resultado de los intentos por reproducir mediante computadoras el funcionamiento del cerebro humano. La capacidad más buscada en las RNAs es la capacidad de aprender de la experiencia y generalizar a partir de ella. Las RNAs han pasado a ser una buena herramienta para comprender y poder predecir un complejo sistema de variables que están intercorrelacionados.

La información se procesa en elementos simples llamados neuronas (nodos). Para procesar la información, las neuronas se organizan en capas: Capa de entrada, que es la que transmite las variables inputs utilizadas, las capa de salida, que presentan el output final y las capas ocultas que se encuentran entre la capa de entrada y de salida y son las que procesan la información.

Cada neurona tiene asociado un peso Wij que representa la intensidad de la señal recibida corresponde al input i en la neurona j.

El aprendizaje se realiza mediante el ajuste de los pesos que ponderan las conexiones entra las neuronas que componen la red.



Las redes supervisadas es un tipo de red neuronal que consiste en disminuir el error que comete la red cuando se introducen datos en la fase de entrenamiento, esta red puede calcular el error que se comete y modificar los pesos sinápticos.

Las arquitecturas más utilizadas son la de PERCEPTRON MULTICAPA y la de BASE RADIAL

3.5.1Las RNA de Base Radial:

La función de base radial es una función que calcula la distancia euclideana de un vector de entrada x respecto de un centro c, de tal manera que resulta la siguiente función:

$$f(x) = (||x - c_i||)$$

A cada neurona de la capa de entrada le corresponde una función de base radial ϕ(x) y un peso de salida wi. El patrón de salida ingresa a una neurona de salida que suma las entradas y da como resultado una salida. La función de una red RBF final resulta:

$$F(x) = \sum_{i=1}^{N} w_i \phi(\left|\left|x - c_i\right|\right|)$$

Las redes RBF tienen una construcción rígida de tres capas: Capa de entrada, capa oculta y capa de salida (a diferencia de otras redes backpropagation)

3.5.2. Arquitectura

Cada red de base radial tiene 3 capas diferentes en total:

- ✓ **Capa de entrada:** Transmiten las señales de entrada a las neuronas ocultas sin realizar procesamiento, es decir, las conexiones de la capa de entrada a la capa oculta no llevan pesos asociados.

- ✓ **Capa oculta:** Realizan una transformación local y no lineal de dichas señales.

3.6. MAQUINA DE VECTORES DE SOPORTE

Modelo supervisado que sirve para clasificar datos, son consideradas más sencillas que las RNAs y desde el punto de vista algebraico son planos (hiperplanos). Pertenecen a la familia de clasificadores lineales puesto que inducen separadores lineales o hiperplanos en espacios de características de muy alta dimensionalidad (introducidos por funciones núcleo o Kernel) con un sesgo inductivo muy particular.

Trata de encontrar el $\mathbf{w}^T\varphi(\boldsymbol{x}_i) + b$ hiperplano de máximo margen que separa los puntos de datos positivos y negativo.

Dado un conjunto de datos de entrenamiento $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$,

donde N es el número de puntos de datos de formación, x_i es un vector de características de entrada y y_i es la etiqueta de clase de destino correspondiente, un SVM puede ser formulado como el siguiente problema de optimización:

$$\begin{aligned} \underset{\mathbf{w},b,\xi_i}{\text{minimize}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_i \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\varphi(\boldsymbol{x}_i) + b) \geqslant 1 - \xi_i, \\ & \xi_i \geqslant 0, i = 1, \ldots, N, \end{aligned} \tag{1}$$

Donde C> 0 es el parámetro que controla el equilibrio entre los errores de formación y la complejidad de modelo. Al introducir el multiplicador de **Lagrange$\boldsymbol{\alpha_i}$,** un problema dual correspondiente se puede derivar como el problema de programación cuadrática (QP) siguiente:

$$\begin{aligned} \underset{\alpha_i}{\text{maximize}} \quad & -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_i \alpha_i \\ \text{subject to} \quad & \sum_i \alpha_i y_i = 0, \\ & 0 \leqslant \alpha_i \leqslant C, i = 1, \ldots, N, \end{aligned} \tag{2}$$

Una vez que el problema dual se resuelve, la función de decisión resultante en cualquier punto de datos de prueba x es el siguiente:

$$f(\boldsymbol{x}) = \mathbf{w}^T\varphi(\boldsymbol{x}) + b = \sum_{i=1}^{N}\alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b = \sum_{i\in SV}\alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b. \tag{3}$$

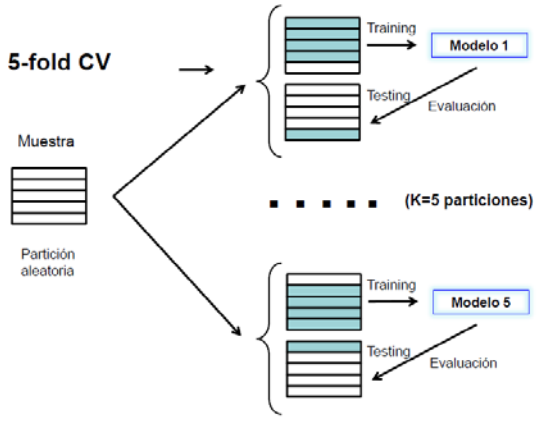
3.7. TÉCNICAS DE REMUESTRO

Las técnicas de remuestreo constituyen una propuesta muy robusta y de gran validez para la estimación de la capacidad de generalización de los modelos desarrollados. Estos modelos consideran múltiples muestras de aprendizaje (k>1), obtenidas a partir de la muestra original, que se utilizan para el desarrollo de los distintos modelos, empleándose los individuos incluidos en cada sub-muestra para la validación de los resultados obtenidos.

3.7.1 VALIDACION CRUZADA

El método de Validación Cruzada (Cross-Validation, CV) consiste endividir de forma aleatoria la muestra en K subconjuntos de similartamaño, y ajustar K modelos, dejando cada vez una partición como conjunto de test, y construyendo el modelo con las K-1 restantes
El número K de particiones (folds) se elige dependiendo del tamaño de lamuestra. Por lo general suele elegir K=10 (10-fold CV) o K=5.
La estimación del error se calcula como promedio de los K errores evaluadosen las muestras de testing de las K particiones. (Machine Learning, Neural and StatisticalClassificationMichie)

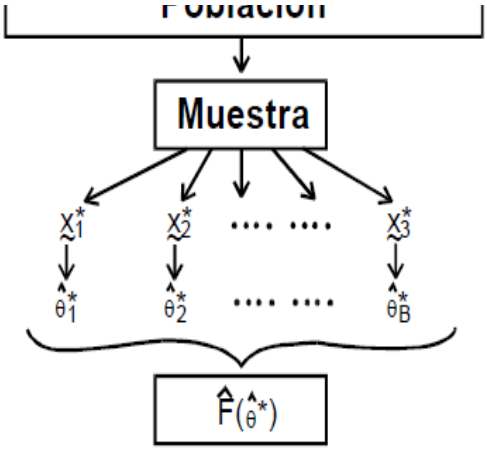
GRÁFICO 3.7.1
TÉCNICAS DE REMUESTREO. VALIDACIÓN CRUZADA



3.7.2BOOTSTRAP

Método de remuestreo que se suele utilizar en muestras pequeñas. Se elige una muestra aleatoria con reemplazamiento de tamaño N para laconstrucción del modelo.
Se repite este proceso un número Bprefijado de veces.

GRÁFICO 3.7.2
ESQUEMA BOOTSTRAP

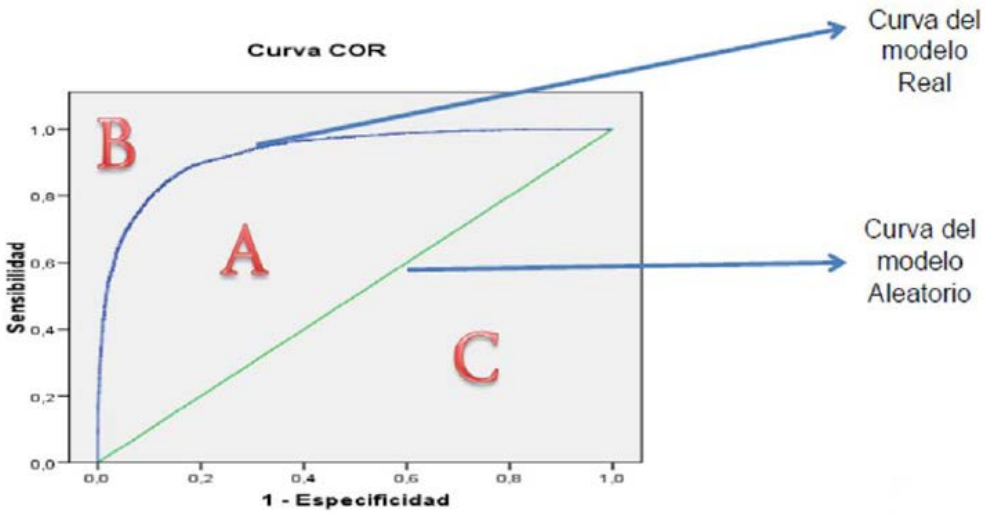


El reemplazamiento de las observaciones permite crear tantas sub-muestras como se desee (B), de tamaño P, que pueden analizarse de forma independiente y permiten estimar medidas centrales robustas de error e intervalos de confianza asociados a los resultados obtenidos.

A medida que B tiende a infinito, el error estimado se aproxima al error real de generalización. No obstante, a efectos prácticos se considera que si B varía entre 25 y 200 replicaciones los resultados obtenidos resultan suficientemente robustos (MERLER y FURLANELLO, 1997: SHAKHNAROVICH. 2001). Este método permite obtener estimaciones muy buenas del error verdadero.

3.8 CURVA ROC / INDICE DE GINI

Mide la probabilidad de que dos elementos elegidos al azar de la misma población se encuentren en la misma clase. En el caso de una población pura, esta probabilidad es 1.
La curva ROC indica que cuanto más alejada este de la diagonal principal mejor es el método de diagnóstico, ya que la curva ROC ideal sería la que con una especificidad de 1 tuviera una sensibilidad de 1,y cuanto más cercana esté a dicha diagonal peor será el método de diagnóstico.



3.9TERMINOLOGIA BASICA

SENSIBILIDAD: La capacidad predictiva del modelo para detectar el suceso de interés cuyo valor es 1.

ESPECIFICIDAD: La capacidad predictiva del modelo de no detectar el suceso de interés cuyo valor es 0.

CARTERA MOROSA: La cartera morosa neta incluye los créditos vencidos, en cobranza judicial, refinanciados y reestructurados netos de provisiones.

RECUPERACION: La recuperación de cartera morosa consiste en rescatar los adeudos pendientes por créditos que otorgan las instituciones financieras.

COLOCACIONES: Son los préstamos de dinero que un banco otorga a su clientes con el compromiso de que el cliente devolverá dicho préstamo.

CAPÍTULO IV

METODOLOGÍA

4.1 POBLACIÓN EN ESTUDIO

La población en estudio la constituye la Cartera Morosa de Clientes del Banwest durante el periodo de marzo y agosto del 2016 con total de 15000 registros.

4.2 FUENTES DE INFORMACIÓN

La información corresponde a la Base de Datos del Banwest, la cual contiene la información histórica de la Cartera Morosa de Clientes.

4.3 DEFINICIÓN DE VARIABLES

El registro de cada cliente moroso, incluye variables que pueden clasificarse en las siguientes categorías:

1. DATOS BÁSICOS

Permite tener detalles de los préstamos y nombres asignados a cada cliente.

CUADRO N° 4.1
VARIABLES ASOCIADAS AL CLIENTE

NRO_PRESTAMO	Código identificador asignado al préstamo.
CODIGO_CLIENTE	Código identificador asignado a los clientes.
NOMBRE_CLIENTE	Contiene los datos de nombres y apellidos.
SEXO	Representa el género del cliente moroso.

2. DATOS DEL CREDITO

Estos hacen referencia al detalle del crédito que la persona solicita, contiene tanto información detallada del producto como información financiera que responden al cliente.

CUADRO N° 4.2
VARIABLES ASOCIADAS AL CRÉDITO

TOTAL_DEUDA	Representa el monto de deuda en soles que incluye intereses y otros.
IMPORTE_DESEMBOLSO	Representa el monto total en soles desembolsado por el Banwest.
SALDO_VENCIDO	Representa el monto en soles que esta por pagar.
DIAS_ATRASO	Representa los días de mora por parte del cliente.
CUOTAS_PAGADAS	Numero de cuotas pagadas
TASA_INTERES	% de tasa de interés.
FECHA_ULTIMO_PAGO	Fecha en que se realizó el último pago.
SUBFAMILIA	Área para el cual se utiliza el crédito desembolsado.
PRODUCTO	Representa el tipo de producto ofrecido por el Banwest.
TIPO_PRESTAMO	Individual, Grupal.
SEGMENTACION	Calificación propia del Banwest (1,2,3,4,5)
SECTOR	Sector de Destino para el cual se desembolsó el crédito
CALIFICACION_SBS	Clasificación crediticia por parte de la Sbs. (0,1,2,3,4)

4.4 DISEÑO DE MUESTREO Y PREPARACION DE DATOS

Se utilizará un tipo de muestreo probabilístico, en este caso el muestreo a utilizar será un muestreo Aleatorio Simple. En donde la unidad muestral estaría representada por cada cliente de la Cartera Morosa del BanWest.

Se definió un tamaño de muestra $n = 1500$. Que representan el 10% de población total para el periodo de Marzo y Agosto del 2016.

4.5 PROCESAMIENTO ESTADÍSTICO

El procesamiento estadístico, estará dado por los siguientes pasos:

1. Seleccionar las variables que aporten más a la construcción de los modelos mediante el algoritmo de Boruta.
2. Determinar los conjuntos de entrenamiento y prueba.

   Consiste en dividir a los datos en dos submuestras:

   I. Una muestra de entrenamiento (training set) que se usa para la construcción del modelo
   II. Una muestra de validación (testing set) donde se realizan las predicciones y donde se evalúa la capacidad predictiva del modelo.

   Por lo general la partición más utilizada es la 70% (muestra construcción) y 30% (Muestra Validación)

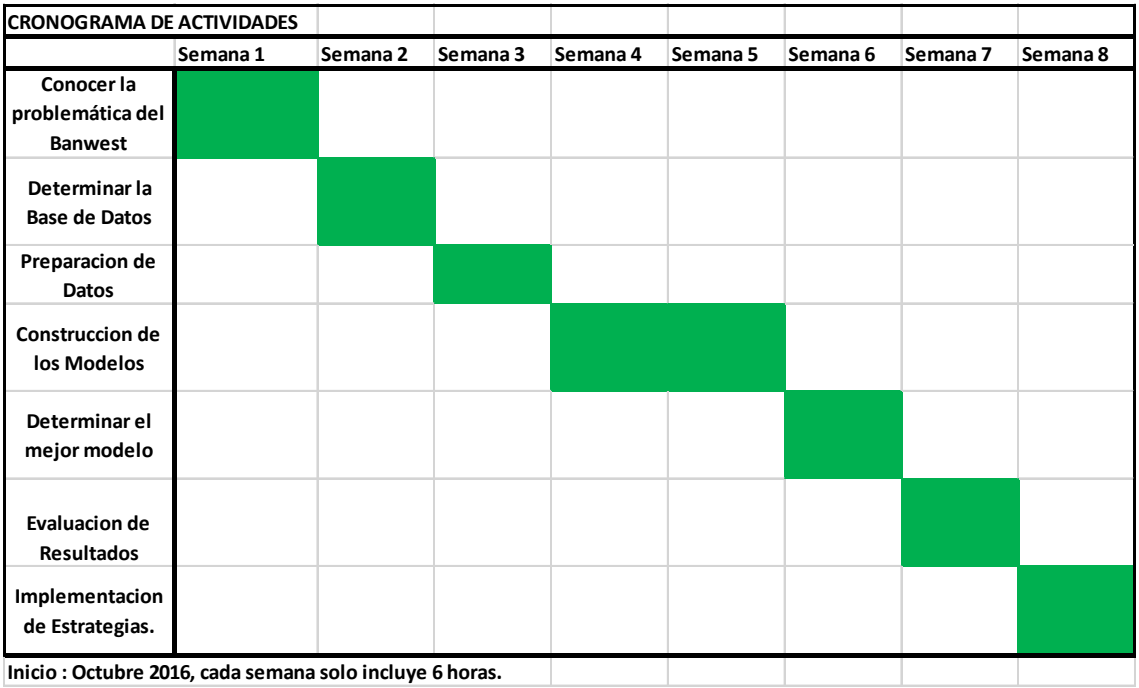
3. Elegir el número de muestras Bootstrap y particiones en la validación cruzada.
4. Desarrollar los modelos Maquina de Vectores Soporte Bootstrap y Validación Cruzada.

5. Comparar los modelos obtenidos mediante indicadores de sensibilidad, especificidad, porcentaje global eÍndice de Gini.
6. Elegir el mejor método de remuestreoque se ajuste a la problemática de predecir la morosidad en la construcción del Modelo de Maquina de Vectores de Soporte.

DIAGRAMA DE GANT

Con el propósito de planificar de manera eficiente nuestras actividades, se planteará el siguiente cronograma:

1. **Conocer la problemática del Banwest,** nos permitirá identificar las necesidades que afrenta el banco.
2. **Determinar la técnica estadística** que mejor se ajuste a la necesidad encontrada.
3. **Determinar la Base de Datos adecuada**, de tal manera que está pueda utilizada en el análisis de manera eficiente.
4. **Preparación de Datos**implicará una previa revisión de la Base de Datos, con el propósito de facilitar la construcción de los modelos.
5. **Construcción de los Modelos,** en esta etapa se procederá a implementación de los modelos.
6. **Determinar el mejor modelo,** mediante una comparación de indicadores tales como sensibilidad, porcentaje global e Índice de Gini.
7. **Evaluación de Resultados,** en cuanto a indicadores financieros del banco.
8. **Implementación de Estrategias,** nos permitirá aplicar estrategias diferenciadas según las necesidades encontradas.



COSTEO Y PRESUPUESTO

En cuanto a gastos, se comprará una unidad USB para el almacenamiento de información. Asimismo tendremos gastos en los corresponde a utilización de pappers, impresiones, internet y licencias de software. Siendo un total asignado de 410 soles.

DESCRIPCIÓN	COSTO
Compra USB	S/. 40.00
Licencias Software	S/. 120.00
Pappers	S/. 100.00
Impresiones	S/. 50.00
Internet	S/. 100.00
TOTAL	S/. 410.00

RESULTADOS

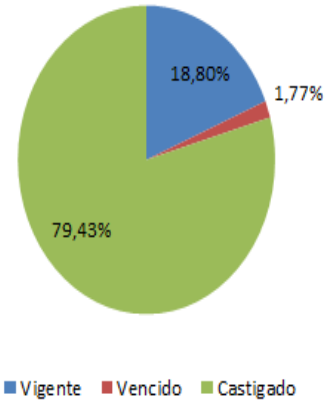
Análisis Descriptivo

El análisis de las variables independientes nos permitirá conocer la estructura de estas en la población de la cartera morosa del BanWest.
Dichas variables que serán tomadas en cuenta son:

- Estado del Préstamo
- Calificación_SBS
- Tipo de Credito
- Codigo_Segmen_Int
- Sector
- Tipo de Persona
- Refinanciado

Las proporciones según el estado del préstamo, nos indica que el 79,43% de la muestra elegida pertenecen a créditos que han sido castigados. Por lo tanto nuestras estrategias deberían estar enfocadas a este grupo de créditos.

GRAFICO Nº 5.1
Distribución por estado del préstamo



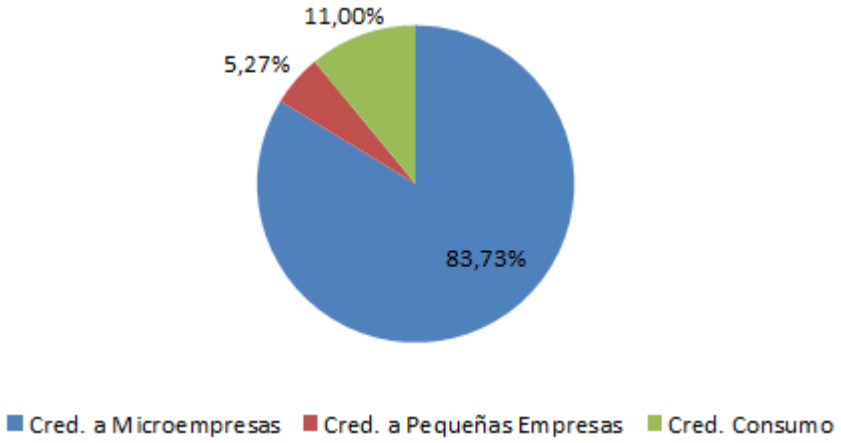
En el caso de la Calificación crediticia por parte de Superintendencia de Banca y Seguros (SBS), el 80% del total de la muestra presenta una calificación de Pérdida, siendo esto perjudicable para el banco; por tal motivo se debe tomar medidas que ayuden en la recuperación de la cartera.

CUADRO Nº 5.1
Calificación Crediticia SBS

CALIFICACION_SBS	Porcentaje
Normal	17,63
CPP	1,07
Deficiente	,33
Dudoso	,93
Pérdida	80,03

En el tipo de crédito, se observa que alrededor del 94% del total de la muestra forman parte de los créditos a microempresas y créditos de consumo. Por lo que se deberá fortalecer la gestión de cobranza a este tipo de créditos.

GRAFICO Nº 5.2
Distribución por tipo de préstamo



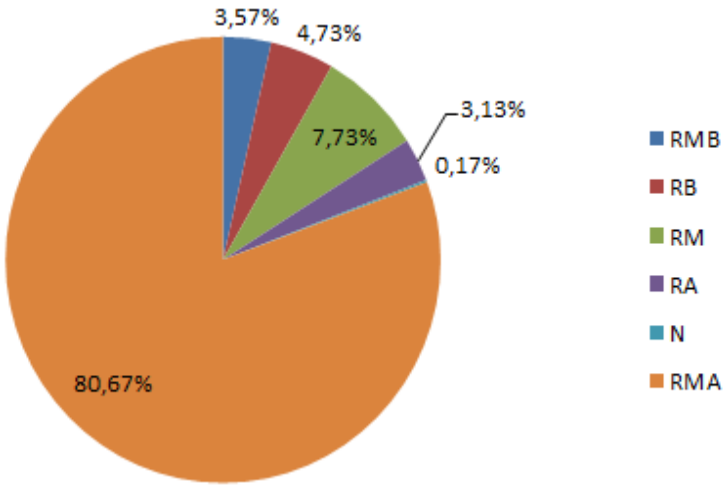
En lo que corresponde al sector para los cuales son adquiridos los créditos, se observa que el 65% corresponden al sector de Comercio, seguido por el Sector Servicios con un 25% y finalmente el sector Producción que representa el 10% de la muestra total.

CUADRO Nº 5.2
Sector de Adquisición

SECTOR	Porcentaje
Comercio	65,2
Produccion	10,3
Servicios	24,5

Finalmente en cuanto al tipo de segmentación interna, alrededor del 81% de clientes de la muestra en estudio presentan una segmentación con un Riesgo muy Alto (RMA). Por lo que deberá desarrollarse estrategias con el propósito de impactar en estos clientes.

GRAFICO Nº 5.3
Distribución por Segmentación Interna



Análisis entre la variable dependiente e independiente

Mediante la prueba Chi Cuadrado, analizaremos la relación de interdependencia o dependencia entre la variable dependiente e independientes.

En el cuadro Nº 5.3, observamos que bajo un nivel de significancia del 5% la prueba Chicuadrado entre la variable dependiente y las independientes, resultan ser significativas. Por lo tanto podemos afirmar que existe una relación de dependencia entre las variables mencionadas, lo que permitirá construir un modelo dado que resultan ser variables influyentes en el target.

CUADRO Nº 5.3
PRUEBA CHICUADRADO PEARSON

TARGET VS VARIABLES	Estadistico Chi-cuadrado de Pearson	gl	p-value
ESTADO_PRESTAMO	1971,89	2	0,00
CALIFICACION_SBS	1887,88	4	0,00
FAMILIA	107,48	2	0,00
TIPO_CREDITO	428,32	2	0,00
SECTOR	7,30	2	0,03
COD_SEG_INT	1816,20	5	0,00
SEXO	28,09	2	0,00

Análisis de las variables independientes cuantitativas, mediante la clasificación de grupos morosos y no morosos.

Mediante la Comparación de medias para muestras independientes, analizaremos si existen diferencias significativas entre los grupos de morosos y no morosos. Consideran un nivel de significancia del 5% podemos afirmar que los grupos son estadísticamente diferentes, tal como se muestra en el cuadro Nº 5.4, por lo que tiene sentido trabajar con grupos que son diferentes.

CUADRO Nº 5.4
PRUEBA DE MEDIAS

Prueba de muestras independientes

VARIABLES	igualdad de varianzas		Prueba T para la igualdad de medias		
	F	Sig.	t	gl	Sig. (bilateral)
CUOTAS	193,655	,000	-24,166	2998	,000
TASA_INTERES	39,289	,000	14,836	2998	,000
DIAS_ATRASO	177,894	,000	60,500	2998	,000
IMPORTE_DESEMBOLSO	553,033	,000	-19,036	2998	,000
SALDO_VENCIDO	18,830	,000	8,360	2998	,000
SALDO_VIGENTE	499,809	,000	-17,433	2998	,000
IMPORTE_SALDO	303,941	,000	-13,140	2998	,000
INTERESES_DEV_VIGENTE	921,690	,000	-24,128	2998	,000
INTERESES_DEV_SUSTENTAT	71,300	,000	7,298	2998	,000
CUOTAS_PAGADAS	152,587	,000	-20,817	2998	,000
TOTAL_DEUDA	368,286	,000	-14,081	2998	,000
CANTIDAD_PRESTAMOS	199,487	,000	-13,338	2998	,000
PROV_ALINEADA_SOLES	142,808	,000	-7,368	2998	,000

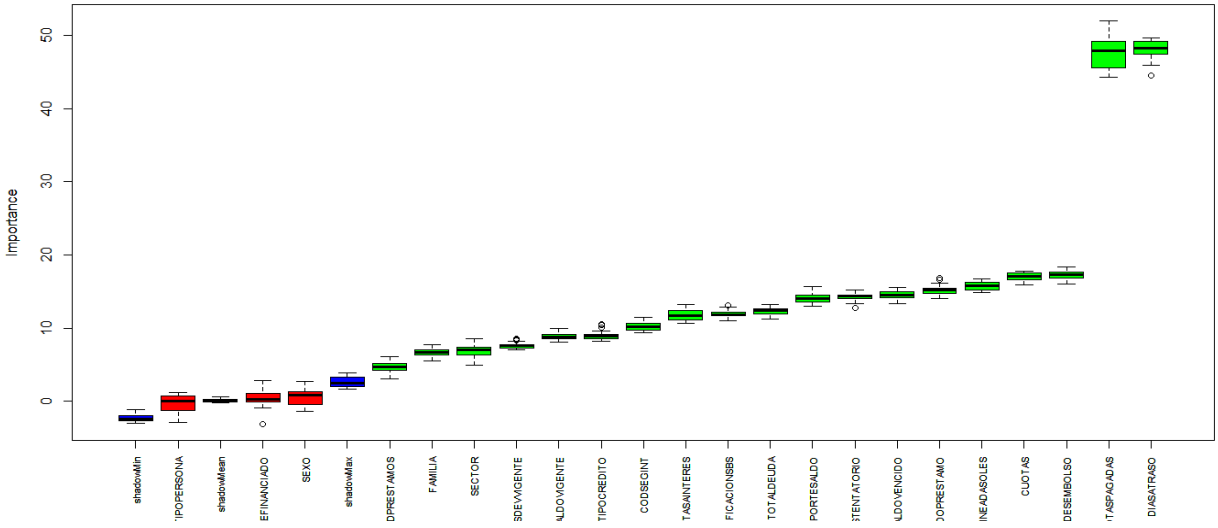
Se han asumido varianzas iguales

Análisis de Maquina de Vectores de Soporte

1. Selección de variables mediante el Algoritmo Boruta.

Mediante el algoritmo Boruta, un método alternativo de selección de variables, determinaremos las variables que influyen más en la variable morosidad.
El algoritmo confirma que 19 variables son importantes, rechazándose 3 variables como no influyentes mediante 19 iteraciones. Por lo que se considerará, aquellas que tengan una importancia mayor al 15 % (ver anexo 4).

CUADRO Nº 5.4
IMPORTANCIA POR VARIABLES –ALGORITMO BORUTA



Por lo tanto las variables que más influyen en la morosidad de clientes del banco Banwest son:

Días de atraso, cuotas pagadas, importe del desembolso, cuotas, estado del préstamo, deuda total ,tipo de credito, tasa de intereses, calificacion de la Superintendencia de Banca y Seguros.

2. Determinación de los conjuntos de entrenamiento y prueba.

La muestra seleccionada se divide en dos conjuntos:

MUESTRA CONSTRUCCIÓN.- nos permitirá obtener los modelos de Máquina de Vectores de Soporte, se considerará el 70% de los datos.

MUESTRA DE VALIDACIÓN.- servirá para poder evaluar los modelos obtenidos en base a indicadores de sensibilidad, especificidad y Índice de Gini. Representará el 30% de los datos.

CUADRO Nº 5.5
MUESTRA DE VALIDACION Y CONSTRUCCION

		TARGET		
		No Moroso	Moroso	Total
MUESTRA VALIDACION	Recuento	189	704	893
	% dentro de FLAG_MUESTRA	21.2%	78.8%	100.0%
	% dentro de TARGET	30.0%	29.7%	29.8%
	% del total	6.3%	23.5%	29.8%
MUESTRA CONSTRUCCIÓN	Recuento	440	1667	2107
	% dentro de FLAG_MUESTRA	20.9%	79.1%	100.0%
	% dentro de TARGET	70.0%	70.3%	70.2%
	% del total	14.7%	55.6%	70.2%
TOTAL	Recuento	629	2371	3000
	% dentro de FLAG_MUESTRA	21.0%	79.0%	100.0%
	% dentro de TARGET	100.0%	100.0%	100.0%
	% del total	21.0%	79.0%	100.0%

3. Determinación delnúmero de muestras Bootstrap, y particiones en la validación cruzada.

La construcción del modelo mediante validación cruzada, consistirá en dividir a la población en 10, 20,30particiones de tal manera que dicho proceso se repita 5 veces; mientras que el remuestreoBootstrap constará de 20,30 y 40 muestras.

4. Desarrollar los modelos de Máquina de Vectores Soporte Bootstrap y Validación Cruzada.

Para la construcción de los modelos de Máquina de Vectores de Soporte con Función Radial, se hará uso delSoftware R, con el propósito de obtener información sobre dichosmodelosyhacer la comparación entre ellos.

Para poder el obtener el ROC del Modelo Máquina de Vectores de Soporte Bootstrap, se consideraran 20,30 y 40 muestras, de tal manera que el modelo de SVM con BASE RADIAL con sigma=0.005 y C=5 es el modelo con mayor ROC igual a 88.89%,queresulta ser el promedio de los 40 ROCs obtenidos de los modelos del re-muestreo para esos parámetros. (Ver Anexo 2).

En el cuadro 5.6 se muestran los resultados del Área bajo la CURVA ROC y el INDICE GINI, sensibilidad, especificidad y porcentaje global mediante el modelo de Máquina de Vectores Bootstrap.

CUADRO Nº 5.6
INDICADORES DEL MODELO SVM BOOTSTRAP

ROC	88.89%
GINI	77.78%
SENSIBILIDAD	86.78%
ESPECIFICIDAD	62.75%
PORCENTAJE GLOBAL	74.85%
ERROR	891.123

Para poder el obtener el ROC del Modelo Máquina de Vectores con validación cruzada, se considerarán 10,20 y 30particiones, de tal manera que este proceso se repetirá 5 veces. Siendo el modelo de SVM con BASE RADIAL con sigma=0.001 y C=5 es el modelo con mayor ROC igual a 88.63%, que resulta ser el promedio de los 5 iteracciones de los ROCs obtenidos de los modelos del re-muestreo para esos parámetros. (Ver Anexo 3).

CUADRO Nº 5.7
INDICADORES DEL MODELO SVM - CV

ROC	88.63%
GINI	77.26%
SENSIBILIDAD	77.80%
ESPECIFICIDAD	66.36%
PORCENTAJE GLOBAL	73.89%
ERROR	1547.450

CONCLUSIONES

El mejor método de remuestreo, para la construcción del modelo de Maquina de Vectores de Soporte, para predecir la morosidad de los clientes del banco Banwest, resulta ser el método Bootstrap, presenta un indicador de sensibilidad del 86.78%, por lo que resulta ser mejor en un 8.9% en comparación con el método Validacion Cruzada (77.80%) , es decir , mediante el modelo de Máquina de Vectores de Soporte con el método Bootstrap se logró clasificar a los clientes de la Cartera Morosa del BanWesten un 86.78% como clientes potenciales morosos.

Con el algoritmo Boruta, se determinó que los factores que más influyen en la morosidadde los clientes de la Cartera Morosa son el estado del préstamo, Calificación de la Superintendencia de Banca y Seguros, Segmentación interna, Total Deuda, Tipo de crédito, Importe de Desembolso, Tasa de interés, cuotas pagadas, Importe de Saldo, Días de mora, Saldo Vencido.

El escenario, bajo el cual el modelo Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método Bootstrap, presentó un mejor rendimiento, considerando 40 muestras Bootstrap, para lo cual se presentó un indicador de Gini del 77.78%.Mientras que la capacidad de predecir el evento moroso fue 86.78% con un error de estimación del 891.123.

El escenario, bajo el cual el modelo Maquina de Vectores de Soporte, para la predicción de la morosidad de los clientes del banco Banwest, mediante el método Validación Cruzada, presentó un mejor rendimiento, considerando 10 particiones, para lo cual se presentó un indicador de Gini del 77.26%.Mientras que la capacidad de predecir el evento moroso fue 77.80% con un error de estimación del 1547.45.

RECOMENDACIONES

Considerando la problemática de predecir de la morosidad de los clientes en un banco, se podría utilizar técnicas opcionales, como árboles de decisión, regresión logística, bosques aleatorios; que puedan ayudar a comparar el rendimiento de los modelos mediante los métodos de remuestreo.

Establecer intervalos, tanto para el número de muestras Bootstrap como para el número de particiones, que según el modelo a utilizar, permitan ver en que rangos resultan ser más potentes.

No sólo utilizar este tipo de metodología en el sector financiero, sino que podría tener aplicación en otras áreas de la investigación tales como salud, marketing e investigación de mercados.

ANEXOS

ANEXO Nº 1
SINTAXIS DEL MODELO DE MAQUINA DE VECTORES BOOTSTRAP EN R

```
library(caret)
library(corrplot)
library(lattice)
library("pROC")
library(glmet)
install.packages("Matrix")
install.packages("lattice")
install.packages("ggplot2")
install.packages("caret")

data=read.table(file.choose(),T)
names(data)
dim(data)
data$TARGET<- as.factor(data$TARGET)
indY = which ( "TARGET" == names(data) )
xx.train<-data[data$FLAG_MUESTRA==1,-24]
xx.test<-data[data$FLAG_MUESTRA==0,-24]
dim(xx.train)
dim(xx.test)
## Distribución de la Variable Respuesta
table(xx.train$TARGET)
## Control de la Técnica de Remuestreo
fiveStats = function(...) c (twoClassSummary(...), defaultSummary(...))
##trControl: controla la construcción del modelo y el proceso de la técnica de remuestreo usada
cv.ctrl = trainControl(method ="repeatedcv",number=10,repeats=5,classProbs = TRUE,summaryFunction = fiveStats)
##summaryFunction : una función que evalúa la capacidad predictiva.
##expand.grid() crea un dataframe con todas las combinaciones de los parametros.
svmGrid = expand.grid ( .C = c ( 1, 5, 10, 50) , .sigma = c ( 0.001, 0.005, 0.01,0.05) )
##Antes de correr el Modelo instalar el paquete Kernlab
install.packages("kernlab")
library(kernlab)
library("pROC")
## Construcción del Modelo Predictivo
svm.fit=train(xx.train[ ,-indY],xx.train$Target,method = "svmRadial",tuneGrid = svmGrid,trControl = cv.ctrl,metric = "ROC",prob.model = TRUE )
svm.fit
##El modelo de SVM con sigma=0.001 y C=5 es el modelo con mayor AUC, 0.9507578 que
##es el promedio de las 40 ROCs de los modelos del remuestreo para esos parámetros
## Parámetros óptimos
```

```
svm.fit$bestTune
## Gráfico del AUC respecto a los 2 parámetros
dev.new()
plot(svm.fit )
dev.new()
plot(svm.fit, metric= "Kappa" )
### Modelo Final, construido con los parámetros óptimos
#$finalModel contiene el modelo predictivo construido con el paquete.
#kernlab en la muestra de training con los parámetros óptimos
svm.fit$finalModel
class(svm.fit)
## Clasespredichas
pred.train.class = predict(svm.fit$finalModel, newdata = xx.train [ , -indY] )
pred.test.class = predict ( svm.fit$finalModel, newdata = xx.test [ , -indY] )
head(pred.test.class)
B<-data.frame(pred.test.class)
write.table(B,"CLASESPREDICHAS.txt",sep="\t")
## Probabilidadespredichas
pred.test.prob = predict(svm.fit$finalModel , newdata = xx.test [ , -indY] , type =
"prob" )
A<-data.frame(pred.test.prob)
write.table(A,"PROBABILIDADESPREDICHAS.txt",sep="\t")
#ANALISIS DE SENSIBILIDAD
confusionMatrix ( pred.test.class, xx.test$Target,dnn = c("Prediction",
"Reference") )
confusionMatrix ( pred.train.class , xx.train$Target ,dnn = c("Prediction",
"Reference"))
dev.new()
#GRAFICANDO CURVA ROC
roc (xx.test$Target, pred.test.prob [ , 2] , plot=T ,main="CURVA
ROC",col=3,pch=19)
#IMPORTANCIA DE LAS VARIABLES
svm.imp = varImp(svm.fit , scale = F )
svm.imp
head(svm.imp$importance)
## Gráfico segun la importancia de las variables
dev.new()
plot(svm.imp, top=8,main="IMPORTANCIA DE LAS VARIABLES")
## Control de la Técnica de Muestreo
svm.fit$control$index
svm.fit$control$index$Fold01.Rep1
names(svm.fit$control$index)
## Remuestreo del Modelo Final
dim(svm.fit$resample)
svm.fit$resample
C<-data.frame(svm.fit$resample)
#El objeto $resample tiene información del proceso de remuestreo en el
modelo final
#construido con los parámetros óptimos
mean(svm.fit$resample$ROC)
```

ANEXO Nº 2
PROMEDIO DE LOS ROCS OBTENIDOS EN LAS 40 MUESTRAS BOOTSTRAP

MUESTRA	ROC	Sens	Spec
1	87.851%	73.300%	91.703%
2	98.615%	85.800%	91.831%
3	89.002%	72.124%	91.703%
4	94.752%	86.633%	88.418%
5	94.269%	72.124%	93.018%
6	91.742%	70.443%	92.949%
7	95.120%	88.300%	92.949%
8	90.148%	66.241%	89.105%
9	97.354%	84.729%	91.890%
10	97.106%	85.522%	93.095%
11	94.587%	82.050%	91.668%
12	95.903%	77.586%	91.668%
13	95.088%	68.856%	91.919%
14	97.388%	90.223%	92.967%
15	95.926%	77.586%	94.300%
16	90.060%	74.838%	90.891%
17	93.263%	79.967%	89.538%
18	92.751%	71.633%	90.596%
19	97.475%	73.300%	93.050%
20	90.914%	63.300%	91.890%
21	94.401%	79.967%	91.736%
22	98.691%	90.223%	94.300%
23	91.867%	91.871%	91.947%
24	96.126%	86.027%	93.095%
25	96.599%	78.006%	92.984%
26	93.579%	82.531%	91.039%
27	97.767%	75.800%	93.151%
28	97.552%	67.146%	94.300%
29	93.956%	88.300%	93.080%
30	98.412%	91.078%	91.947%
31	94.909%	72.124%	92.967%
32	98.265%	98.300%	89.422%
33	97.183%	79.967%	91.947%
34	96.794%	78.006%	93.001%
35	97.782%	91.871%	91.668%
36	88.949%	79.967%	88.895%
37	97.383%	82.050%	91.668%
38	95.632%	96.633%	90.596%
39	95.122%	77.586%	92.892%
40	95.397%	99.014%	89.105%

ANEXO Nº 3
PROMEDIO DE LOS ROCS OBTENIDOS EN EL METODO VALIDACION CRUZADA

	ROC	SENSIBILIDAD	ESPECIFICIDAD
1	87.037%	72.195%	75.286%
2	87.086%	78.143%	70.536%
3	92.484%	71.429%	68.714%
4	87.932%	78.889%	57.214%
5	88.616%	88.326%	60.071%

ANEXO Nº 4
IMPORTANCIA DE LA VARIABLES SEGÚN EL AGORITMO BORUTA

	row.names	meanImp	medianImp	minImp	maxImp	normHits	decision
1	ESTADOPRESTAMO	15.2236346	15.20896391	14.060353	16.839997	1.00000000	Confirmed
2	CALIFICACIONSBS	11.9304982	11.85806023	10.974315	13.106751	1.00000000	Confirmed
3	CUOTAS	17.0423398	17.03335452	15.955689	17.818045	1.00000000	Confirmed
4	TASAINTERES	11.8466954	11.73143912	10.654383	13.202794	1.00000000	Confirmed
5	DIASATRASO	48.1106809	48.27888682	44.553924	49.636176	1.00000000	Confirmed
6	IMPORTEDESEMBOLSO	17.3094537	17.28870804	16.007596	18.403738	1.00000000	Confirmed
7	SALDOVENCIDO	14.5364974	14.44307598	13.369080	15.573674	1.00000000	Confirmed
8	SALDOVIGENTE	8.8196936	8.82697463	8.113527	9.896331	1.00000000	Confirmed
9	IMPORTESALDO	14.1089447	13.99623990	12.969350	15.698179	1.00000000	Confirmed
10	INTERESESDEVVIGENTE	7.6144189	7.56578340	6.975087	8.562385	1.00000000	Confirmed
11	INTERESESDEVSUSTENTATORIO	14.2215344	14.32930377	12.721209	15.238830	1.00000000	Confirmed
12	CUOTASPAGADAS	47.6698564	47.87048599	44.308561	51.965126	1.00000000	Confirmed
13	TOTALDEUDA	12.2936206	12.38951815	11.280502	13.211916	1.00000000	Confirmed
14	CANTIDADPRESTAMOS	4.7051543	4.63355416	3.045491	6.066703	0.94736842	Confirmed
15	REFINANCIADO	0.4540209	0.20008447	-3.100600	2.858036	0.10526316	Rejected
16	FAMILIA	6.7286396	6.69929070	5.494085	7.733455	1.00000000	Confirmed
17	TIPOPERSONA	-0.2905578	-0.01428159	-2.920847	1.162772	0.00000000	Rejected
18	SEXO	0.5746356	0.87978722	-1.440818	2.762260	0.05263158	Rejected
19	SECTOR	6.8934983	7.04620576	4.868351	8.506843	1.00000000	Confirmed
20	PROVALINEADASOLES	15.7923332	15.77374608	14.857733	16.766013	1.00000000	Confirmed
21	TIPOCREDITO	9.0484141	8.91769862	8.138504	10.571464	1.00000000	Confirmed
22	CODSEGINT	10.2106724	10.12800337	9.390374	11.489405	1.00000000	Confirmed

REFERENCIAS BIBLIOGRAFICAS

[1] Berry, M., &Linoff, G. (1997). Data mining techniques for marketing sales and customer support. Wiley Computer Pub.

[2] Cristianini, N., & Taylor, J. S. (2000). An introduction to support vector machines. Cambridge, MA: Cambridge University Press

[3] Hung, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications

[4] Schebesch, K. B., &Stecking, R. (2005). Support vector machines for classifying and describing credit applicants: Detecting typical and critical regions. Journal of the Operational Research Society, 56(9), 1082–1088.

[5] Machine Learning, Statistical and classification D. Michie, DjSpiegelhalter, C.C Taylor, 1994.