Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Académico Profesional de Ingeniería de Sistemas

Sistema de predicción de clientes desertores de tarjetas de crédito para la banca peruana usando Support Vector Machine

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTORES

Rosa Angela del Carmen ORDOÑEZ CAIRO

Maria Doris PASTOR ZAPATA

ASESOR

David Santos MAURICIO SÁNCHEZ

Lima, Perú

2016

Referencia bibliográfica

Ordoñez, R. & Pastor, M. (2016). *Sistema de predicción de clientes desertores de tarjetas de crédito para la banca peruana usando Support Vector Machine*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Académico Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Académico Profesional de Ingeniería de Sistemas

Acta de Sustentación de Tesis

Siendo las 20:25 horas del día 27 de Mayo del año 2016, se reunieron los docentes designados como miembros de Jurado de la Tesis, presidido por el Mg. Jorge Raúl Díaz Muñante (Presidente), el Lic. Jorge Luis Chávez Soto (Miembro) y el Dr. David Santos Mauricio Sánchez (Miembro Asesor) para la sustentación de la Tesis Intitulada: **"SISTEMA DE PREDICCIÓN DE CLIENTES DESERTORES DE TARJETAS DE CRÉDITO PARA LA BANCA PERUANA USANDO SUPPORT VECTOR MACHINE".** Por la Bachiller: ORDOÑEZ CAIRO, ROSA ANGELA DEL CARMEN; para optar el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el presidente invitó al graduado a dar las respuestas a las preguntas establecidas por los Miembros del Jurado.

EL graduado en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el graduado obtuvo la nota de 17 (En letras) DIECISIETE

A continuación el Presidente de Jurados el Mg. Jorge Raúl Díaz Muñante, declara al graduado **Ingeniero de Sistemas.**

Siendo las 21:15 Horas, se levantó la sesión.

.....
Presidente
Mg. Jorge Raúl Díaz Muñante

.....
Miembro
Lic. Jorge Luis Chávez Soto

.....
Asesor
Dr. David Santos Mauricio Sánchez

700

131

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Académico Profesional de Ingeniería de Sistemas

Acta de Sustentación de Tesis

Siendo las 20:25 horas del día 27 de Mayo del año 2016, se reunieron los docentes designados como miembros de Jurado de la Tesis, presidido por el Mg. Jorge Raúl Díaz Muñante (Presidente), el Lic. Jorge Luis Chávez Soto (Miembro) y el Dr. David Santos Mauricio Sánchez (Miembro Asesor) para la sustentación de la Tesis Intitulada: **"SISTEMA DE PREDICCIÓN DE CLIENTES DESERTORES DE TARJETAS DE CRÉDITO PARA LA BANCA PERUANA USANDO SUPPORT VECTOR MACHINE".** Por la Bachiller: PASTOR ZAPATA, MARIA DORIS; para optar el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el presidente invitó al graduado a dar las respuestas a las preguntas establecidas por los Miembros del Jurado.

EL graduado en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el graduado obtuvo la nota de 17 (En letras) DIECISIETE

A continuación el Presidente de Jurados el Mg. Jorge Raúl Díaz Muñante, declara al graduado **Ingeniero de Sistemas.**

Siendo las 21:15 Horas, se levantó la sesión.

.....
Presidente
Mg. Jorge Raúl Díaz Muñante

.....
Miembro
Lic. Jorge Luis Chávez Soto

.....
Asesor
Dr. David Santos Mauricio Sánchez

Este trabajo está dedicado a toda nuestra familia, en especial a nuestras madres, por su apoyo constante.

AGRADECIMIENTOS

A nuestras familias, fuente de apoyo constante e incondicional en nuestras vidas y aún más en nuestros duros años de carrera profesional, y en especial queremos expresar nuestro más sincero agradecimiento a nuestras madres que sin su ayuda hubiera sido imposible culminar nuestra profesión.

A la Universidad Nacional Mayor de San Marcos por habernos aceptado ser parte de ella y abierto la puerta de su seno científico para poder estudiar y convertirnos en profesionales en lo que tanto nos apasiona.

A los diferentes docentes que brindaron sus conocimientos y su apoyo para seguir adelante día a día.

A nuestro asesor de tesis, Dr. David Mauricio Sánchez, por sus conocimientos, orientaciones y paciencia que han sido fundamentales para nuestra formación como investigadoras.

Sistema de predicción de clientes desertores de tarjetas de crédito para la banca peruana usando Support Vector Machine

RESUMEN

La retención de clientes es un tema importante pues está comprobado que cuesta cinco veces más adquirir un nuevo cliente que retenerlo. El problema tratado en el presente trabajo consiste en la deserción de los clientes de la banca del servicio de tarjeta de crédito dado que, en algunos casos, la banca peruana cuenta con un porcentaje de deserción de hasta 6% produciendo pérdidas considerables dado que el cliente no solo pierde la línea de crédito de la tarjeta, además pierde la probabilidad de acceder a un préstamo en esa entidad generando un ingreso menos. El pronóstico servirá para que la empresa tome medidas preventivas a fin de evitar la deserción y tener éxito en la retención del cliente.

La presente tesis propone la implementación de un sistema de predicción de clientes desertores de tarjeta de crédito usando un modelo de predicción basado en el comportamiento transaccional y datos demográficos de los clientes para la determinación de los patrones de reconocimiento, para ello se definirán las técnicas y algoritmos con los que se validará la propuesta.

Palabras clave: Sistema de predicción, Support Vector Machine.

Prediction system of credit card customer churns for Peruvian banks using Support Vector Machine

ABSTRACT

Customer retention is an important issue because it is proven that costs five times more to acquire a new customer than retain it. The problem addressed in this thesis is the defection of customers banking service credit given that, in some cases, in some cases the Peruvian banking has a dropout rate of up to 6%, resulting in considerable losses since the customer not only lost the credit card also lose the chance to access a loan to that entity generating an income less. The outcome will help the company to take preventive measures to prevent dropout and success in customer retention measures.

This thesis proposes the implementation of a system for predicting deserters credit card customers using a prediction model based on transactional behavior and demographic customer data to determine patterns of recognition algorithms and techniques it will be defined to validate the proposal.

Keywords: Prediction Sytem, Support Vector Machine.

ÍNDICE

LISTA DE CUADROS

LISTA DE FIGURAS

LISTA DE TABLAS

Capítulo 1: Introducción

1.1. Antecedentes

Hoy en día podemos apreciar un aumento considerable de entidades financieras que producen competencia en el mercado. Con el pasar del tiempo estas entidades han producido la mejora y la calidad de sus servicios, incluyendo las mejoras en las ofertas de créditos producido por un menor costo del crédito orientado a la baja, también por la política monetaria del BCR de reducción de encajes y tasa de referencia, produciendo así un aumento de clientes los cuales ahora pueden optar por cambiar de entidad dependiendo de las ventajas que estas les ofrezcan.

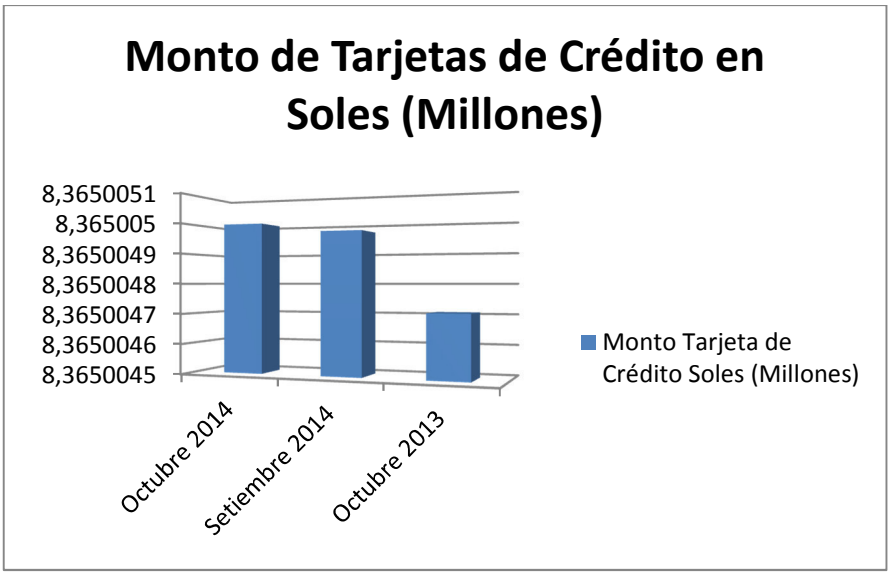
Al cierre del octubre del 2014, el monto de crédito utilizado a través de la tarjetas de crédito de bancos y financieras totalizó S/. 18,816 millones, alcanzando una expansión de S/. 283 millones (1.53%) en comparación con el mes anterior y de S/. 1,414 millones (8.13%) frente a octubre del 2013, afirmó ASBANC.

De acuerdo a cifras entregadas por ASBANC, al término del décimo mes del 2014 el número de tarjetas de crédito sumó S/. 8,365 005millones, superando en 29,313(0.35%) y 291,424 (3.61%) a los resultados observados en setiembre del 2014 y octubre del 2013, respectivamente.

La Asociación de Bancos (ASBANC[1]) informó en su última memoria publicada del año 2014 que la cartera de créditos totales a personas y empresas del sistema bancario peruano privado sumó S/. 193,128 millones al cierre del 2014, representando un crecimiento anual del 10%.

Dada la mayor tasa de crecimiento estimada para la economía nacional en el 2015 (alrededor del 4.8% del PBI, según la Cámara de Comercio de Lima), se esperaría que los préstamos en general registren un incremento anual entre 13% y 15%".

[1] La Asociación de Bancos del Perú representa a los bancos afiliados y ejercer su presencia en las decisiones que afecten al sector financiero.

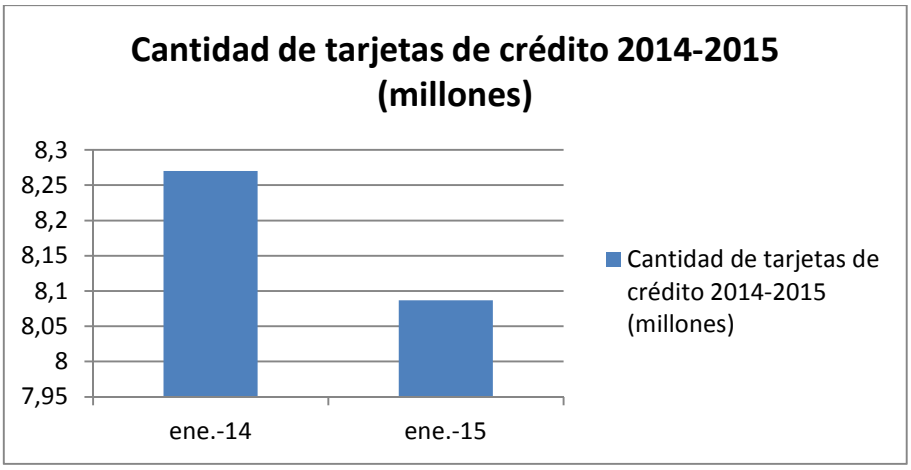


Fuente: ASBANC

Durante enero del 2015, los peruanos (personas naturales y empresas) utilizaron sus tarjetas de crédito por un monto total de S/. 19,369 millones, el cual representa el 32,98% del total del volumen de las líneas de crédito otorgadas a través de este producto financiero.

ASBANC también explicó que del total del financiamiento otorgado a través de tarjetas de crédito de bancos y financieras, a enero del 2015, el 94% fue concedido en moneda nacional y solo el 6% restante en moneda extranjera. Se precisó que en el primer mes del año el 78% del monto utilizado correspondió a las tarjetas de crédito de personas naturales (consumo), mientras que el 22% restante a empresas.

Por el lado del número de tarjetas, a enero del 2015 ascendió a 8,27 millones, cifra que representó un crecimiento de 183.118 unidades (2,26%) frente a enero del 2014. La titularidad del 98% de las tarjetas corresponde a personas naturales.



Fuente: ASBANC

Considerando la fuerte competencia y las medidas de reducción de encaje y de la tasa de interés de referencia por parte del Banco Central de Reserva (BCR), se espera que el costo de financiamiento con tarjetas de crédito se reduzca aún más en los siguientes meses.

Como podemos observar, la cantidad de entidades financieras han ido en aumento en nuestro país, es por eso que el servicio brindado por las entidades financieras ha ido mejorando. Sabemos que el ingreso principal de una entidad financiera son sus clientes, el dinero que ellos depositen en sus entidades. Es por eso que el cliente y la fidelidad de ellos es el factor más importante para dichas entidades. Además, si se logra una mayor permanencia de un cliente en la institución, se obtienen los beneficios asociados a la disminución de los costos operacionales, las referencias y al incremento en las transacciones.

La cartera de clientes es uno de los activos más importantes para una institución financiera, ya que está estrechamente relacionada con las utilidades del negocio. Dos actividades comerciales tienen como objetivo mantener y mejorar dicha cartera: la captación de clientes nuevos y la retención de clientes existentes.

La captación de clientes apunta a aumentar el número de clientes de la cartera a través de la definición e incorporación de nuevos segmentos objetivos. Esta captación se realiza principalmente a través de elaboradas estrategias de publicidad, alta inversión en fuerza de ventas y la generación de ofertas focalizadas. La retención de clientes consiste en la identificación de los clientes con mayores tendencias a la deserción y en la determinación de las estrategias o procedimientos que aumenten el grado de fidelización y bajen los índices de deserción en la cartera.

Existen dos tipos de deserción: la deserción voluntaria y la deserción no voluntaria. Las deserciones voluntarias se asocian a la desafiliación del cliente por iniciativa propia, sin injerencia directa por parte de la institución. A diferencia del caso anterior, las deserciones no voluntarias son desafiliaciones en donde el banco es responsable directo del término de los acuerdos contractuales, donde el cliente no posee ninguna injerencia. Este tipo de cierre se da principalmente por acciones delictuales o por mala utilización de los productos. En el presente trabajo nos centraremos en las deserciones voluntarias.

Según Iván Aquino [59], actualmente las entidades financieras peruanas requieren de tecnologías de bases de datos para explotar la información almacenada producto de sus transacciones y operaciones diarias. Es en este contexto que los tópicos emergentes de bases de datos como Datawarehouse y Data Mining sobresalen como las tecnologías para acumular, transformar y explotar los datos con el fin de descubrir conocimiento útil para la toma de decisiones. El término business intelligence[2] es el que abarca los conceptos de Datawarehouse y Data Mining, y es una tecnología de información que ayudaría mucho a las entidades financieras a resolver uno de sus principales problemas como es: "la deserción de clientes".

1.2. Declaración del Problema

El problema tratado en el presente trabajo consiste en la deserción de los clientes de la banca del servicio de tarjeta de crédito. En consecuencia, es importante la gestión estructurada y uso de los datos de los clientes para la retención de los clientes para evitar pérdidas en la entidad financiera la identificación de los clientes desertores con el fin de generar confianza y transmitirles la seguridad necesaria con el fin de conseguir su no abandono del servicio.

1.3. Objetivos de la Investigación

1.3.1. Objetivo general

Implementar un sistema empleando un modelo de predicción basado en el análisis del comportamiento transaccional y datos demográficos de data histórica de clientes

[2]Se denomina **inteligencia empresarial**, **inteligencia de negocios** o **BI** (del inglés ***business intelligence***) al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa.

que hayan dejado de usar el servicio de tarjeta de crédito para la determinación de los patrones de reconocimiento aplicando la técnica Support Vector Machine, para así permitir predecir e identificar a clientes que estén propensos a dejar de utilizar el servicio de tarjeta de crédito brindado por una entidad financiera con el fin de reducir la tasa de deserción.

1.3.2. Objetivos específicos

Los objetivos específicos son:

- Revisar en la literatura las técnicas más populares aplicadas en casos similares.
- Utilizar perfiles y hábitos de consumo de las personas que poseen una tarjeta de crédito en el Perú. Conocer los hábitos de las personas que no están satisfechas con el servicio brindado por una entidad financiera y, por lo tanto, dejan de utilizar dicho servicio para cambiar por otro proveedor de servicio.
- Aplicar un modelo que permita predecir a los clientes que piensan dejar de utilizar el servicio de tarjetas de crédito brindado por la entidad financiera
- Diseñar y desarrollar un software que alcance un porcentaje de acierto mayor al 90%, tal que garantice su efectividad, y ejecutarlo con escenarios reales o simulados de una entidad financiera.

1.4. Justificación

Cada vez que un competidor nuevo ingresa al mercado peruano de las finanzas, aumenta la posibilidad de que las entidades financieras compartan clientes con otras instituciones. Una situación así impacta en el desempeño financiero e incrementa la probabilidad de que el cliente deje de usar los servicios de estas empresas.

El hábito de consumo del cliente financiero extranjero es diferente a la realidad del consumidor peruano, por lo que se hace necesaria la implementación de un modelo que obtenga patrones representativos de créditos de los clientes dentro del universo de datos transaccionales recopilados de la entidad financiera.

Actualmente, en algunos casos la banca peruana cuenta con un porcentaje de deserción del 6% produciendo pérdidas considerables dado que, de acuerdo a la entidad financiera, el cliente no solo pierde la línea de crédito de la tarjeta, sino que pierde la probabilidad de acceder a un préstamo en esa entidad generando así la reducción en los ingresos.

La puesta en práctica de la solución propuesta pretende obtener un modelo predictivo de deserción de clientes, el cual no solo identificará algunas de las variables más relevantes e influyentes en la deserción de la clientela, sino que, también proporcionará un procedimiento para evaluar fácilmente a los clientes actuales y así determinar quiénes podrían ser futuros desertores. Con la información obtenida, la entidad estará en la facultad de tomar decisiones efectivas a fin de reducir la tasa de deserción y así fidelizar a sus clientes.

1.5. Alcances

El objetivo de la tesis se centra en una entidad financiera peruana. El público objetivo será tarjetahabientes[3] que cuenten con por lo menos una tarjeta de crédito en dicha entidad financiera. El modelo a implementar deberá contemplar reglas del consumidor peruano.

La presente tesis modela una estrategia para la identificación de clientes que sean propietarios de por lo menos una tarjeta de crédito cuyo comportamiento indica que dejará de utilizar el servicio brindado por la entidad financiera.

1.6. Organización de la Tesis

Los capítulos de la presente tesis están organizados de la siguiente manera:

- En el Capítulo 2 se desarrolla el estado del arte y, de manera muy concreta, se explican los métodos y algoritmos utilizados para la predicción de la deserción de clientes.
- El Capítulo 3 expone la técnica a utilizar para enfrentar el problema de manera muy detallada, con el fin de que se pueda entender muy bien la técnica.
- En el Capítulo 4 se desarrolla la metodología de minería de datos CRISP-DM, con el objetivo de guiar en forma estructurada el proceso de descubrimiento de conocimiento basada en las tareas de entendimiento de los datos, preparación de los datos, modelaje, evaluación e implementación. También se explica la técnica enfocada a nuestro problema. Se aplica la técnica identificada anteriormente como la mejor con las variables identificadas para nuestro problema.

[3]Titular de una tarjeta de crédito o débito.

- El Capítulo 5 trata acerca del artefacto, del sistema en sí. Se detalla los requerimientos que se han tenido en cuenta para el sistema, la arquitectura, y los demás datos técnicos del software.
- El Capítulo 6 trata acerca de los experimentos de validación utilizando un algoritmo de aprendizaje supervisado para la predicción de posibles clientes desertores que utilizan tarjetas de crédito en una entidad financiera.
- Finalmente, en el Capítulo 7 se menciona una serie de conclusiones y los trabajos futuros de los resultados obtenidos en esta propuesta de investigación.

Capítulo 2: Estado del arte

2.1. Revisión de la literatura

Altin y Bajram (2006)[1] muestran indicadores para identificar la deserción de los clientes usando la fórmula Waterfield; así, con el fin de identificar y analizar la tasa de deserción, se hace uso de diferentes herramientas como entrevistas personales o llamadas telefónicas que se aplican a las diferentes sucursales del banco en el caso de estudio.

Ning Wang, Dong-xiao Niu (2009) [2] detallan que hoy en día, con el desarrollo del dinero de la informatización los mercados y la competencia siendo intensa, el servicio y gestión de la tarjeta de crédito se convierte en un tema importante para ganancias bancos. Nos mencionan que los clientes se convierten en "desertores" cuando cancelan su tarjeta de crédito. Ning Wang y Dong-xiao Niu citan que “Es más rentable para el banco mantener y satisfacer a los clientes existentes que atraer constantemente nuevos clientes, y se ha demostrado que una pequeña disminución de la tasa de rotación puede dar lugar a cambios significativos en la contribución” [3]. Así, los autores nos explican que retener al cliente de tarjeta de crédito en el grado máximo es el punto de incremento de ganancia. Es por eso que ambos con el fin de controlar al cliente desertor con eficacia mencionan que es importante construir un efectivo y preciso modelo de deserción de clientes que pueda predecir los clientes más propensos a cancelar la tarjeta. El efecto de las estrategias de retención de cliente depende de la exactitud del modelo de predicción. Muchos métodos y algoritmos se utilizan para predecir la pérdida de clientes, tales como árbol de decisión [4], redes neuronales [5], algoritmos genéticos [6], clasificador bayesiano [7], etc. Sin embargo, los autores mencionan que cuando los algoritmos basados en árboles de decisión se amplían para determinar las probabilidades asociadas con las clasificaciones, es posible que algunas hojas en un árbol de decisión tengan similar clase de probabilidades. Así como también, las redes neuronales no expresan explícitamente el patrón descubierto en una manera simbólica, de fácil comprensión. De la misma manera, los algoritmos genéticos no pueden determinar la probabilidad asociada con las predicciones. Finalmente, los autores mencionan que el clasificador bayesiano necesita encontrar máxima verosimilitud usando mucha búsqueda.

El modelo de predicción de la pérdida de los clientes de tarjeta de crédito propuesto por Ning Wang y Dong-xiao Niu combina las ventajas de la RST y LS-SVM. La técnica usada

por los autores es una combinación del Rough Set Theory (RST), que se usó para refinar los datos originales a través de discretización de datos y la reducción de atributos, y del Least Square Support Vector Machine (LS-SVM), que basándose en estos datos se usa para un reconocimiento de patrones basándose en Vapnik-Chervonenks (VC). Estos autores presentaron una comparativa entre métodos como árbol de decisión, regresión Ridge, y el modelo ANN para medir la efectividad de estos. Los resultados de las pruebas demostraron que el modelo LS-SVM tiene un excelente rendimiento, mejor que otros métodos.

K. Hossein Abbasimehr, Mostafa Setak y M. J. Tarokh [24] detallan que en los últimos años, debido a la saturación de los mercados y el entorno empresarial competitivo, la deserción de clientes se convierte en una preocupación central de la mayoría de las empresas en todas las industrias. Neslin et al. [8] definen al cliente rotación como la tendencia de los clientes a dejar de hacer negocios con una empresa en un período de tiempo determinado. La predicción de deserción de clientes es una herramienta útil para predecir quiénes son los clientes con riesgo a irse. Técnicamente, la predicción es clasificar a los clientes en dos tipos: clientes tipo churn (abandonar la empresa) y clientes tipo no churn (los clientes que siguen haciendo sus negocios con la empresa) [9]. Por predicción precisa de churners y no churners, una empresa puede utilizar los limitados recursos de marketing eficaz para dirigirse a los clientes churner en una campaña de marketing de retención. Adquirir un nuevo cliente cuesta 12 veces más que retener la existente [10], por lo que una pequeña mejora en la precisión de la predicción de churn puede resultar un gran beneficio para una empresa [11]. Técnicas de minería de datos se han utilizado ampliamente en la pérdida de clientes en el contexto de la predicción, tales como máquinas de soporte vectorial (SVM) [12, 13,14], árbol de decisión [15], la red neuronal artificial (ANN) [16, 17], la regresión logística [18, 19]. La precisión no es el único aspecto importante en la evaluación de los modelos de predicción. Los modelos de predicción de deserción de clientes deben ser a la vez comprensibles y precisos. La comprensibilidad del modelo hace que se revele algún conocimiento sobre la pérdida de clientes de tipo churn. Tal conocimiento puede extraerse en forma de "si entonces" reglas que permiten el desarrollo de una estrategia de retención más eficaz. En este estudio, los autores aplican Neuro Sistema de Inferencia Difuso Adaptativo (ANFIS, Adaptive Neuro Fuzzy Inference System) como clasificador neuro borroso para la pérdida de clientes predicción. Sistemas neuro difusos se han desplegado con éxito en muchas aplicaciones, y se obtiene un

conjunto de reglas que se deriva de una perspectiva difusa inherente a los datos. De hecho, el principal objetivo de este estudio es comparar los ANFIS como clasificador neuro difuso con otras técnicas de minería de datos.

K. Iyakutti y V. Umayaparvathi [23] explican el término de minería de datos, el cual es muy genérico y se refiere a la minería de datos para descubrir conocimiento (información). En la literatura se define como un proceso de extracción y análisis de patrones, relaciones e información útil a partir de bases de datos masivas. Este proceso de la minería también se llama como descubrimiento de conocimiento en bases de datos (KDD). En cualquier proceso de minería de datos, hay cuatro sub-tareas involucradas. Ellos son: clasificación, agrupamiento, regresión y de reglas de asociación de aprendizaje [20]. Por otra parte, en función del ámbito de aplicación, las técnicas de minería de datos se dividen en dos categorías principales: i) Verificación orientado (el sistema verifica la hipótesis) y ii) Descubrimiento orientado (el sistema encuentra nuevas reglas y los patrones de forma autónoma) [21]. Métodos de verificación frente a la evaluación de una hipótesis propuesta por una fuente externa. Métodos estadísticos como prueba de bondad de ajuste, la prueba t de medias y análisis de varianza viene en esta categoría. Estos métodos están menos asociados con las técnicas de minería de datos que sus homólogos de descubrimiento orientados porque la mayoría de los problemas de minería de datos se refieren a la selección de una hipótesis (de un conjunto de hipótesis) en lugar de probar uno conocido [22]. Sin embargo, los métodos de descubrimiento se utilizan para identificar patrones en los datos de forma automática. Técnicas de minería de datos se aplican en la base de datos de telecomunicaciones para diversos fines. Cada una utiliza diferentes tipos de datos de telecomunicaciones en función de la finalidad. Los datos generados por las industrias de telecomunicaciones se agruparse en 3 tipos. Ellos son: i) los datos de clientes (Demografía), ii) datos de la red y iii) datos de cuentas.

Jaewook Lee Namhyoung Kim, Kyu-Hwan Jung y Yong Seog Kim [27] mencionan la disponibilidad de discos duro de gran espacio en el mercado y la expansión de las tecnologías de recolección de datos potencian muchas empresas de negocios para monitorear fácilmente y visualizar la compra de los clientes todos los días y los patrones de uso a través de procesamiento de transacciones en línea (OLTP, online transaction processing) de bases de datos [25]. Por lo tanto, en estos días, la mayoría de las compañías tienen un montón de datos. Sin embargo, los datos en sí no es información, y los datos

deben convertirse en información para que los usuarios puedan responder a sus propias preguntas con la información correcta en el momento adecuado y en el lugar correcto.

Los autores mencionan que hay que tener en cuenta que muchas compañías en la industria de las telecomunicaciones han estado sufriendo de tasas extremadamente altas de deserción de clientes, es decir, entre el 20% y el 40% de los clientes dejan su actual proveedor de servicios para un año determinado, sobre todo por sus tecnologías y servicios relativamente homogéneos los llevan a competir en términos de bajar cargos por servicio. En tal caso, el papel del marketing se convierte en un factor clave para el éxito. En particular, es bien sabido que es mucho más rentable para una empresa retener un cliente actual y real que reclutar un nuevo cliente considerando un aumento de los costos de comercialización. En particular, los programas de micro marketing u objetivo del marketing con mensajes adaptados son mucho más eficaces en costos que los programas de marketing de masas a través de los canales de comercialización tradicionales como la TV y los periódicos. Por lo tanto, se recomienda que las empresas en un entorno empresarial muy competitivo operen sus propios sistemas de gestión de relaciones con clientes (CRM, Customer Relationship Management), equipadas con la inteligencia empresarial (BI) y herramientas de minería de datos para identificar a un grupo de clientes que tienen más probabilidades de terminar su relación con el servicio actual de proveedores.

Los autores resaltan que hay que tener en cuenta que la identificación y la prevención de deserción de clientes es un tema crítico, ya que el mercado de la telefonía móvil ha llegado a un punto de saturación muy alto y cada empresa se esfuerza por atraer a nuevos suscriptores, y al mismo tiempo se enfocan en la retención de clientes rentables actuales [26]. Para apoyar este esfuerzo los autores introducirán una de estas herramientas de micro marketing para la identificación de clientes a punto de dejar el servicio en nombre de las empresas que pertenecen a la industria de las telecomunicaciones. Los autores notaron que el manejo de deserción de clientes debería comenzar con una identificación precisa de los posibles clientes a punto de dejar de utilizar el servicio, junto con el perfil detallado de la información demográfica, comportamiento transaccional y patrones del comportamiento del usuario. Mientras se desarrollan estrategias de retención y prácticas de gestión específicas para los posibles clientes identificados como desertores (se puede completar un sistema de gestión de clientes desertores), los autores limitan su interés en el desarrollo de un nuevo modelo de conjunto SVM (Support Vector Machine) para identificar con

precisión posibles clientes desertores de sus patrones de uso de servicios recogidos durante un período determinado.

2.2. Modelos de Predicción de Deserción de Clientes en la banca

2.2.1. ROUGH SET THEORY (RST) Y LEAST SQUARE SUPPORT VECTOR MACHINE (LS-SVM)

Ning Wang y Dong-xiao Niu (2009) proponen un modelo de predicción de la pérdida de los clientes de tarjeta de crédito que combina las ventajas de la Rough Set Theory (RST) y Least Square Support Vector Machine (LS-SVM). El algoritmo de predicción emplea RST para refinar los datos iniciales que han sido seleccionados incluyendo discretización de atributos continuos y atributos reducción. Después de ser refinado, el atributo de la fuerte correlación con la pérdida de clientes es más conciso en la estructura y más conveniente para aplicarse en el modelo LS-SVM. La aplicación de la RST puede determinar los principales factores que se traducen en disminución de la lealtad del cliente, y el banco debe prestar mucha atención a mejorar estos puntos clave.

Para efecto de comprender el método, primero definiremos los conceptos de Rough Set Theory (RST) y Least Square Support Virtual Machine (LS-SVM).

Rough Set Theory (RST)

El matemático polaco Z. Pawlak introdujo por primera vez la Teoría de Conjuntos Aproximados (RST) como una nueva herramienta matemática para analizar los datos y hacer frente a los problemas de incertidumbre. Bajo la premisa de mantener la capacidad de clasificación de los datos originales, RST puede refinar los datos originales a través de discretización de datos y la reducción de atributos. A través de la clasificación de la información imprecisa, incompleta e inconsistente con eficacia, RST puede retener reglas inherentes e información útil. Aquí solo se introduce varias definiciones importantes de RST propuestos por este documento, de la siguiente manera:

Sistemas de representación del conocimiento asumidas S = (U, A, V, f), donde U es un conjunto finito de objetos y A es conjunto finito de atributos. Asimismo $V = \bigcup_{a \in A} V_a$, donde $V a$ es un dominio del atributo a. Además, f denota una función de información, que dota cada atributo con un valor numérico, como el siguiente:

$$f : U \times A \rightarrow V \ , f(x, a) = \in V_a, \forall a \in A, \forall x \in U \quad (2.1)$$

Como regla, S = (U, A) puede sustituir a S = (U, A, V, F). El sistema de representación del conocimiento denota los datos como las relaciones. Cuando $A = C \cup D, C \cap D = \Phi$, entonces C se llama como atributo condición y D se llama como atributo decisión. El sistema de representación del conocimiento que posee el atributo de estado y el atributo de decisión se denomina tabla de decisión.

La reducción del conocimiento puede eliminar la información irrelevante o poco importante, al mismo tiempo de preservar la capacidad de clasificación, incluida la reducción de atributos y valores de atributo de reducción. Hay dos tipos de métodos para la reducción de Atributo: una es la operación de definición de núcleo y reducción sobre la base de las relaciones indiscernibles y la otra es la operación lógica basada en matrices discernibles y funciones discernibles.

LS-SVM

Support Vector Machine (SVM) es una nueva tecnología de reconocimiento de patrones que se basa en Vapnik-Chervonenks (VC). La dimensión VC (del inglés Vapnik-Chervonenkis) es una medida de la capacidad de los algoritmos de clasificación estadística, definida como la cardinalidad del mayor conjunto de puntos que el algoritmo puede separar. El algoritmo SVM estándar transforma un problema práctico en un problema de programación cuadrática convexa con restricciones de desigualdad, por otra parte Least Square Support Vector Machine (LS-SVM) transforma un problema práctico en la solución de un sistema de ecuaciones lineales, lo que simplifica el cálculo y aumenta la velocidad de la convergencia. El algoritmo de regresión de LS-SVM es de la siguiente manera:

Teniendo en cuenta el conjunto de datos:

$$s = \left((x_1, y_,), (x_2, y_2), \dots, (x_n, y_n)\right) \in R^N x\, R \qquad (2.2)$$

El objetivo es el de estimar el modelo predictivo como la ecuación:

$$f(x) = W^T \cdot \varphi(x) + b \qquad (2.3)$$

Donde φ (x) es el mapeo no lineal a una alta característica dimensional espacio y w es el vector de peso, b es un desplazamiento. De acuerdo con el principio de minimización del riesgo estructural, el problema de optimización de LS-SVM se formula como la ecuación:

$$\min \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{l} e_1^2 \quad \text{s.t. } y_i = w^T \phi(X_i) + b + e_i \qquad (2.4)$$

Donde $y = [y_1, y_2, \dots, y_n]^T$ es el vector de variables dependientes, e_i es el término de error y $\phi = [(\phi(x_1)^T, \phi(x_2)^T, \dots, \phi(x_n)^T]^T$ es la matriz de regresores no lineal.

El modelo de predicción de la pérdida de clientes requiere dos tipos de tasas de precisión: una es tasa de éxito de predicción, el otro es una tasa cubierta de predicción.

Para resolver los problemas de optimización con restricciones, se transforma la ecuación Figura 3.1.4 en el espacio dual mediante la introducción de la función de Lagrange.

La Función común del Kernel incluye los siguientes tipos:

$$\text{Polynomial Kernel: } k(x,y) = (\sum_{i=1}^{n} x_i y_i)^d$$

$$\text{RBF Kernel: } k(x,y) = \exp(-\frac{\|x_i - x_j\|}{2\sigma^2})$$

$$\text{Cauchy Kernel: } k(x,y) = \prod_{i=1}^{n} \frac{1}{1 + d(x_i - y_i)^2}$$

$$\text{Sigmoid Kernel: } k(x,y) = \tan(\gamma(x_i, x) + c)$$

Ning Wang y Dong-xiao Niu mencionan que resolviendo la ecuación con el método de mínimos cuadrados, a y b pueden ser extraídas, y luego la función LS-SVM resultante en doble que es obtenida es:

$$f(x) = \sum_{i=1}^{l} a_i\, k(x_i, x) + b \quad (2.5)$$

De la ecuación (2.5), podemos ver que un solo parámetro "γ" es incierto, lo que indica que los parámetros a ser optimizados de LS-SVM son menos que la del SVM estándar. Y no es necesario para asumir la exactitud aproximación de ε. Por lo tanto, el Algoritmo LS-SVM tiene las características de una operación simple, velocidad de cálculo rápido y de alta precisión.

A continuación se detallará la aplicación del método en el artículo:

Los autores nos presentan 3 etapas para conseguir el modelo predictivo. Estos son:

- Preparación de la Data

  Los autores utilizan una base de datos desde octubre del 2007 a julio del 2008. Del total, seleccionan 1500 registros, cada registro con 12 atributos.

- Reducción de atributos con RST

  En la etapa de Reducción de atributos con RST se define el conjunto de atributos asociados (incluyendo atributo condiciones y atributos decisión).

  Los autores establecieron un sistema de representación del conocimiento S = (U, A, V, f): conjunto objeto U={1,2...,1500}, atributo de decisión: D = {0,1}, donde "0" indica la pérdida de clientes, "1" indica que el cliente es permanente, atributo de condición $c = \{c_1, c_{2,\dots}, c_{12}\}$, donde por ejemplo c_1 =edad,c_2 = formación académica, etc. (Ning Wang y Dong-xiao Niu, 2009).

  Para luego aplicar RST, anteriormente mencionado, y así reducir los atributos de la pérdida cliente: Bajo la realización de software de Rosetta y quedarse con el conjunto $\{C_3, C_4, C_7, C_8C_9C_{10}, C_{12}\}$ que se convierte en el vector de entrada de LS-SVM (Tabla 2.1).

Attribute type	Attribute name	Code	Classification	Annotation
condition attribute	age	C_1	0	from 20 to30
			1	more than 30
	educational background	C_2	0	senior high school and below
			1	reverse
	monthly income	C_3	1	lower than 2000 Yuan
			2	from 2000-5000 Yuan
			3	more than 5000 Yuan
	time for holding credit card	C_4	1	lower than one year
			2	from one to three year
			3	more than three year
	whether possess the account of BC (current deposit or fixed deposit)	C_5	0	no
			1	yes
	whether the loan customers of BC	C_6	0	no
			1	yes
	transaction times(during observation time)	C_7	0	less than 20 times
			1	reverse
	the amount of transaction money (during observation time)	C_8	0	lower than 10000 Yuan
			1	reverse
	overdraft times(during observation time)	C_9	0	less than 20 times
			1	reverse
	the line of credit	C_{10}	0	lower than 10000 Yuan
			1	reverse
	late fee(during observation time)	C_{11}	0	less than 200 Yuan
			1	reverse
	time since the last transaction	C_{12}	1	lower than 30 days
			2	from30 to 90 days
			3	more than 90 days
decision attribute	customer churn	D	0	customer churn
			1	customer retain

Tabla 2.1 Tabla de clasificación y definición de atributos del cliente

(Niu & Dong-xiao, 2009, págs. 275 - 279)

- Predicción de la Pérdida de Clientes con LS-SVM

  Dado 1.000 muestras de entrenamiento, el conjunto de datos es $\{z_i = (x_i, y_i)\}_{i=1}^{1000}$, donde x_i es el vector de entrada de siete dimensiones, y_i es la salida vector, y = 1 cliente denota clientes perdido, y=-1 denota cliente retenidos.

  El modelo de predicción de la pérdida de clientes requiere dos tipos de tasas de precisión: una es tasa de éxito de predicción, el otro es una tasa cubierta de predicción. La matriz de evaluación de la rotación de clientes para verificar el modelo de predicción se define como en la **Tabla 2.2**:

TABLE III: THE EVALUATION MATRIX OF CUSTOMER CHURN

customer number	predicted retention	predicted churn
real retention	A	C
real churn	B	D

Tabla 2.2 Tabla de la matriz de evaluación de los clientes desertores
(Niu & Dong-xiao, 2009, págs. 275 - 279)

Donde la tasa de precisión = (A+D)/(A+B+C+D), la tasa de éxito = D/(C+D); tasa cubierta = D/(B+D). La estructura de los costes de contabilidad de la tarjeta de crédito requiere que la tasa de éxito de predicción sea más de 65% y la tasa de cobertura de la predicción sea más del 75%.

Los resultados como los de la Tabla 2.3 muestran que el núcleo RBF puede obtener la tasa con más alta precisión para el conjunto de datos Modelo LS-SVM construida en este trabajo, por lo tanto, el núcleo RBF es lo esperado en este modelo.

TABLE IV:
THE RESULTS OF LS-SVM PREDICTION MODEL ON DIFFERENT KERNEL FUCTIONS

kernel function	parameter	accuracy rate
Polynomial Kernel	c=10 d=3	78.7%
RBF Kernel	c=10 σ=0.28	86.2%
Cauchy Kernel	c=10 d=0.4	80.3%
Sigmoid Kernel	c=5 γ=1	45.9%

Tabla 2.3 Tabla de resultados del modelo LS-SVM en las diferentes funciones Kernel
(Niu & Dong-xiao, 2009, págs. 275 - 279)

Los autores, con el fin de comparar con otros métodos, seleccionan el árbol de decisión, regresión Ridge, y el modelo ANN para hacer pruebas comparativas con el modelo LS-SVM. El modelo de árbol de decisión selecciona el modelo común C4.5; Regresión Ridge se basa en el principio de minimización del riesgo estructural; modelo ANN utiliza un modelo no lineal con una sola capa oculta, la condición de parada es si el error ha llegado mínimo local. De acuerdo con 1.000 muestras de entrenamiento y 500 muestras de prueba, los resultados de la predicción de los diferentes modelos se muestran en la **Tabla 2.4** y la **Figura 2.1** (Ning Wang y Dong-xiao Niu, 2009).

TABLE V: THE PREDICTION RESULTS OF DIFFERENT MODELS

models	accuracy rate	hit rate	covering rate
The Decision Tree	80.50%	55.66%	76.62%
Ridge Regression	75.10%	47.02%	61.47%
ANN	84.30%	65.04%	69.26%
LS-SVM	89.90%	74.62%	85.28%

Tabla 2.4 Tabla de resultados de la predicción de los diferentes modelos

(Niu & Dong-xiao, 2009, págs. 275 - 279)

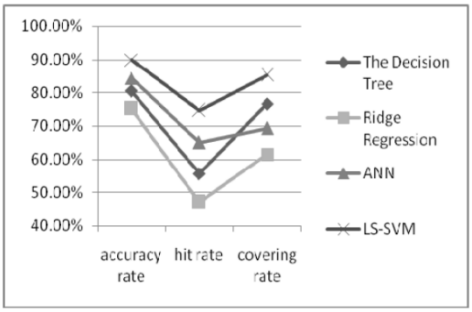


Figura 2.1 Gráfico de comparación de los diferentes modelos

(Niu & Dong-xiao, 2009, págs. 275 - 279)

Comparando los resultados del experimento de la Fig. 2.1.1.4 y la Fig. 2.1.1.5, muestran que el modelo LS-SVM tiene una mayor tasa de precisión de la predicción que otros métodos. La tasa de éxito del modelo LS-SVM y del modelo de ANN es 74.62% y 65.04%, respectivamente, que son más que el estándar de 65%. La tasa de cobertura del modelo LS-SVM y el modelo de árbol de decisión es de 85,28% y 76,62%, respectivamente, que son más que el estándar de 75%. Estos

resultados demostraron que el modelo LS-SVM tiene un excelente rendimiento mejor que otros métodos en la predicción de pérdidas de clientes de tarjeta de crédito. (Ning Wang y Dong-xiao Niu, 2009).

2.2.2. ADAPTATIVE NEURO FUZZY INFERENCE SYSTEM (ANFIS)

K. Hossein Abbasimehr, Mostafa Setak y M. J. Tarokh en el paper "A Neuro-Fuzzy Classifier for Customer Churn Prediction" utilizan la técnica de agrupación de sustracción (Subtractive Clustering method) con la función (genfis2). Teniendo en cuenta conjuntos separados de datos de entrada y de salida, la genfis2 utiliza un método de agrupación sustractiva para generar un sistema de inferencia difusa (FIS, Fuzzy Inference System). Cuando solo hay una salida, genfis2 puede ser utilizado para generar una cantidad inicial de FIS para la formación ANFIS (Adaptative Neuro Fuzzy Inference System) mediante la implementación de primera agrupación de sustracción en los datos. La función genfis2 utiliza la función subclust para estimar las funciones de pertenencia antecedentes y un conjunto de reglas. Esta función devuelve una estructura FIS que contiene un conjunto de reglas difusas para cubrir el espacio de características. Los parámetros de la agrupación sustractiva se establecieron de la siguiente manera: el rango de influencia es 0,5, el factor de calabaza es 1,25, aceptar relación es de 0,5; ratio de rechazo es 0,15. El número de época es igual a 100. Llamamos a los FIS generados por la agrupación de sustracción y entrenado por ANFIS como modelo ANFIS-sustractiva. También utilizaron la técnica de agrupación FCM (Fuzzy c-means) con la función (genfis3). genfis3 genera un FIS utilizando c-means clustering difuso (FCM) mediante la extracción de un conjunto de reglas que modela el comportamiento de los datos. Similar a genfis2, esta función requiere conjuntos separados de datos de entrada y salida como argumentos de entrada. Cuando solo hay una salida, puede utilizar genfis3 para generar una cantidad inicial de FIS para la formación ANFIS. El método de extracción de reglas utiliza primero la función fcm para determinar el número de reglas y funciones de pertenencia de los antecedentes y consecuentes. Los autores han establecido el número de clúster para FCM igual a 6; y el número de época es igual a 100. Los autores llaman a los FIS generados por la agrupación FCM y entrenado por ANFIS como modelo ANFIS-FCM.

Adaptative Neuro Fuzzy Inference System (ANFIS)

La lógica difusa (FL, Fuzzy Logic) y los sistemas de inferencia difusos (FIS), propuesto por primera vez por Zadeh, proporcionan una solución para la toma de decisiones basadas

en datos imprecisos, ambiguos, imprecisos o faltantes. FL representa modelos o conocimientos que utilizan reglas si-entonces bajo la forma de "si X e Y entonces Z". Un sistema de inferencia difuso consiste principalmente en reglas difusas, las funciones de pertenencia y las operaciones fusificación y de-fusificación. Mediante la aplicación de la inferencia difusa, los datos de entrada nítidas ordinarios produce salida exacto ordinaria, que es fácil de entender y de interpretar.

Hay dos tipos de sistemas de inferencia difusos que pueden ser implementadas: tipo Mamdani y Sugeno. Debido a que el sistema de Sugeno es más compacto y computacionalmente eficiente que un sistema de Mamdani, que se presta a la utilización de las técnicas de adaptación para la construcción de los modelos difusos, se utilizará el tipo Sugeno.

Una regla difusa en un modelo difuso Sugeno tiene la forma de, si X es A y Y es B, entonces z = f (x, y), donde A y B son conjuntos difusos de entrada en el antecedente y por lo general z = f (x, y), es una función polinómica de orden cero o primero en el consecuente.

El proceso de razonamiento difuso para el modelo difuso Sugeno de primer orden se muestra en la Figura 2.2 (a).

Para que una FIS sea madura y bien establecida de modo que pueda trabajar adecuadamente en modo de predicción, necesitan ser sintonizadas o adaptadas a través de un proceso de aprendizaje usando un patrón input/output suficiente de los datos de su estructura y los parámetros inicial (lineal y no lineal). Uno de los sistemas de aprendizaje más utilizados para la adaptación de los parámetros lineales y no lineales de un FIS, en particular el modelo difuso Sugeno de primer orden, es el ANFIS. ANFIS es una clase de redes de adaptación que son funcionalmente equivalentes a los sistemas de inferencia difusos.

Arquitectura ANFIS:

Supongamos un sistema de inferencia borrosa con dos entradas "x", "y" y "z" con una salida de primer orden del modelo Sugeno difuso. Conjunto de reglas difusas con dos reglas de "Si entonces" (if A then B) difusas son como se muestra a continuación:

Si "x" es A1 y "y" es B1, entonces f1 = p1x + q1 + r1.

Si "x" es A2 y "y" es B2, entonces f2 = p2x + q2 + r2.

Donde (p1, q1, r1) y (p2, q2, r2) son parámetros de las funciones de salida.

Como se muestra en la Figura 2.2 (b), se puede implementar el mecanismo de razonamiento en un avance de alimentación de la red neuronal con capacidad de aprendizaje supervisado, que se conoce como la arquitectura ANFIS. El ANFIS tiene las siguientes capas como se ilustra en la Figura 2.2.2.1 (b).

- Capa 0: Consiste en la entrada del conjunto de variables.
- Capa 1: La función de nodo para cada nodo "i" en esta capa toma la forma:

$$O_i^1 = \mu A_i(x) \qquad (2.2.5)$$

Donde "x" es la entrada al nodo i, μA_i es la función de pertenencia (que pueden ser triangular, trapezoidal, función gaussiana o de otra forma) de la etiqueta lingüística A_i asociado con este nodo. En otras palabras, O_i^1 es la función de pertenencia de A_i y especifica el grado en el que "x" satisface el cuantificador A_i. En este estudio, la forma gaussiana MFs se define a continuación se utilizan:

$$\mu A_i(x) = \exp\left(\frac{(x-c_i)^2}{2\sigma_i^2}\right) \qquad (2.2.6)$$

Donde $\{c_i, \sigma_i\}$ son los parámetros de la función gausseana MFs. Los parámetros en esta capa son usualmente referidos como parámetros premisa.

- Capa 2: Cada nodo en esta capa multiplica señales de entrada de la capa 1 y envían productos a cabo de la siguiente manera:

$$W_i = \mu A_i(x) \times \mu B_i(x) \quad (2.2.7)$$

  Donde la salida de esta capa W_i representa la fuerza de disparo de una regla.

- Capa 3: Cada nodo "i" en esta capa determina la relación de la fuerza de disparo de la regla i-ésima a la suma de los puntos fuertes de disparo de todas las reglas como:

$$W_i = \frac{w_i}{w_1+w_2} \; i = 1,2 \quad (2.2.8)$$

  Donde la salida de esta capa representa la fuerza de disparo normalizada.

- Capa 4: Cada nodo "i" en esta capa es un nodo de adaptación con una función de nodo de la forma:

$$0_i^4 = W_i f_i = W_i(p_i x + q_i x + r_i) \qquad (2.2.9)$$

Donde W_i es la salida de la capa 3, y $\{p_i, q_i, r_i\}$ es el conjunto de parámetros. Los parámetros en esta capa son referidos como parámetros consecuentes.

- Capa 5: Esta capa consta de un único nodo que calcula la producción total como la suma de todas las señales entrantes desde la capa 4 así:

$$\text{Salida global} = \sum_i W_i f_i = \frac{\sum_i W_i f_i}{\sum_i W_i} \quad (2.2.10)$$

Ambas premisas y los parámetros consecuentes de los ANFIS deben estar sintonizados, usando un algoritmo de aprendizaje de manera óptima la relación entre el espacio de entrada y el espacio de salida. Básicamente, ANFIS toma el modelo difuso inicial y se sintoniza por medio de una técnica híbrida que combina descenso de gradiente de propagación hacia atrás y la media de mínimos cuadrados algoritmos de optimización. En cada época, una medida de error, por lo general se define como la suma del cuadrado de la diferencia entre la salida real y la deseada, se reduce. La formación se detiene cuando se obtiene o bien el número época predefinido o la tasa de error. Hay dos pases en el procedimiento de aprendizaje híbrido para ANFIS. En el pase hacia adelante del algoritmo de aprendizaje híbrido, señales funcionales van hacia adelante hasta que la capa 4 y los parámetros consiguientes se identifican por la estimación de mínimos cuadrados. En el paso hacia atrás, las tasas de error se propagan hacia atrás y hacia los parámetros premisa, se actualizan por el método de descenso de gradiente.

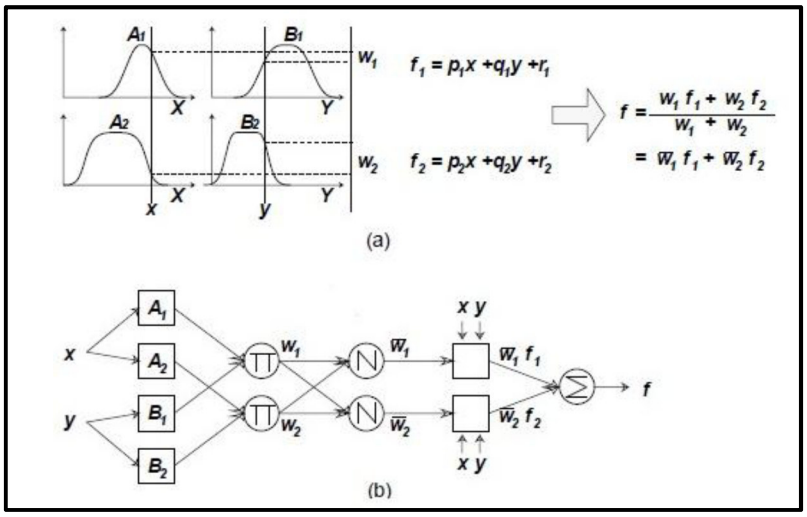


Figura 2.2 (a) Modelo difuso de razonamiento Sugeno, (b) La estructura ANFIS

Los resultados que obtuvieron estos autores fueron con base en un conjunto de datos a disposición del público descargado del repositorio UCI de bases de datos de la máquina de aprendizaje de la Universidad de California, Irvine1. El conjunto de datos contiene 20 variables de valor de la información sobre 5.000 clientes, junto con una indicación de si ese cliente a punto de irse (salió de la empresa). La proporción de churner (término que se le da a los clientes que están a punto de dejar de utilizar los servicios de una entidad) en el conjunto de datos es del 14,3%. En primer lugar, los autores dividieron el conjunto de datos en un 67% y 33% de entrenamiento y prueba de conjunto división, respectivamente. La proporción de usuarios que abandonan fue muestreado con el fin de dar el modelo predictivo una mejor capacidad de los patrones de discriminación exigentes. Por lo tanto, la proporción de churner y no churner en el conjunto de datos de entrenamiento es 50% - 50%. El equipo de prueba no fue muestreado para proporcionar un conjunto de pruebas más real, la tasa de pérdida de clientes se mantuvo un 14,3%. Todos los modelos construidos durante este trabajo fueron evaluados en este conjunto de pruebas.

Los autores muestran una tabla de resultados en la cual podemos ver que se obtiene la más alta precisión utilizando RIPPER (Precisión = 95%). Sin embargo, C4.5 árbol de decisión, ANFIS-sustractivo y ANFIS-FCM siguen muy de cerca, y con excepción de regresión logística, todos los resultados se encuentra en el intervalo de entre el 91% y el 95%. Debido a la precisión supone implícitamente una distribución de clase relativamente equilibrada entre las observaciones y los costos de la igualdad de errores de clasificación, que por sí sola no es una medida de rendimiento adecuado para evaluar los resultados experimentales. El conjunto de datos de los churn reales, tienen una distribución

asimétrica, por lo tanto, el supuesto de igualdad de costes de clasificación errónea no se puede sostener. Por lo general, un gerente de relaciones con clientes, la cuestión más importante es la detección correcta de futuro churner. Dado que los costes relacionados con la clasificación errónea de churners son claramente superiores a los costos relacionados con la clasificación errónea de un no churner, debemos asumir costes de clasificación errónea desiguales. Como resultado, una alta sensibilidad se prefiere en vez de una alta especificidad dependiendo del punto de vista de una compañía. Por supuesto, esto no significa que la especificidad puede ser completamente ignorada. En efecto, una solución de compromiso razonable tiene que ser hecha entre la especificidad y la sensibilidad. Un modelo de predicción que pueda predecir todos los clientes calificados como churners para que el personal de marketing pueda batir una buena campaña de retención. La sensibilidad más alta en nuestros experimentos se obtiene con C4.5 (sensibilidad = 87%). RIPPER, ANFIS-sustractivo, ANFIS-FCM y de regresión logística no son calificados como los peores. La especificidad más alta en nuestros experimentos se alcanza con RIPPER (especificidad = 97,5%). C4.5, y los modelos ANFIS-FCM, ANFIS-sustractivos difieren significativamente en términos de especificidad, a excepción de regresión logística, todos los resultados se encuentra en el intervalo de entre el 92% y el 95,6%. En resumen, los modelos ANFIS-sustractivos y ANFIS-FCM tienen un rendimiento razonable en términos de precisión, especificidad y sensibilidad, como se muestra en la Tabla 2.5.

Technique	Accuracy	Specificity	Sensitivity	#rules
C4.5	94%	95.6%	87%	25
RIPPER	95%	97.5%	85.7%	18
Logistic regression	77.3%	76.6%	82%	----
ANFIS-Subtractive	92%	93%	84%	6
ANFIS-FCM	91%	92%	84%	6

Tabla 2.5 Rendimiento de los algoritmos

2.2.3. KNOWLEGE DISCOVERY IN DATABASES (KDD)

Los autores K. Iyakutti y V. Umayaparvathi proponen en su paper “Applications of Data Mining Techniques in Telecom Churn Prediction” la aplicación de técnicas de Minería de

Datos para la predicción de deserción de clientes en una empresa de telecomunicaciones. Para ello explican en qué consiste el proceso de minería de datos.

Se define como "el proceso no trivial de identificación válida, novedosa, los patrones potencialmente útiles y en última instancia comprensible de los datos". El problema de nuestras ofertas de discusión con la variable objetivo valioso discreta, y nuestro objetivo final es declarar cada suscriptor como "potencialmente churner" (cliente potencial que se puede deserciónr) o "potencialmente no churner" (cliente potencial fiel a la organización), por lo que la función de KDD para nuestro problema se define como el problema de clasificación. El primer paso en el modelado predictivo es la adquisición y preparación de los datos. Tener la información correcta es tan importante como tener el método correcto.

- **Adquisición de la data**

  Los modelos de predicción requieren la historia pasada o el comportamiento de uso de los clientes durante un período específico de tiempo para predecir su comportamiento en el futuro cercano, que no se pueden aplicar directamente al conjunto de datos real. Por lo tanto, es la práctica habitual para llevar a cabo algún tipo de agregación en el conjunto de datos. Durante el proceso de agregación, además de las variables reales, se generaran nuevas variables que exhiben el comportamiento consume periódica de los clientes. Estas variables poseen información vital para ser utilizado por los modelos de predicción para pronosticar el comportamiento de los clientes con antelación.

- **Preparación de la data**

  En los problemas de minería de datos, la preparación de datos consume una considerable cantidad de tiempo. En la fase de preparación de los datos, la información se recopila, integra y limpiado. La integración de datos puede requerir la extracción de datos de diversas fuentes. Una vez que los datos han sido dispuestos en forma de tablas, tiene que ser totalmente caracterizados. Los datos necesitan ser limpiados por resolver las ambigüedades, errores. También los elementos de datos redundantes y problemáticas deben ser eliminados en esta etapa.

  - **Variables derivadas**

Variables derivadas de nuevas variables basadas en las variables originales. Las variables derivadas más eficaces son aquellos que representan algo en el mundo real, tales como la descripción de un comportamiento del cliente subyacente [8]. Debido a las propias variables originales que se agregan, también pueden ser llamados como variables derivadas. En nuestro conjunto de datos, el DAYS_TO_CONTRACT_EXPIRY variable es una variable derivada que se calcula restando la fecha de inicio de la conexión desde la fecha actual en la que se aplica esta predicción. Hay algunas clases generales de variables derivadas, como los valores totales, los valores promedio y las proporciones. Nuestro estudio considera el valor promedio de los últimos seis meses a un tipo variable derivada. Además, la relación entre la media de los últimos tres meses y la media de todos los meses antes de que se utiliza como una variable derivada. Además, se utiliza un número de variables derivadas específicos.

Las variables derivadas explican el comportamiento del cliente de una manera mejor que las variables originales.

Con fines de formación, un conjunto de datos de 18.000 clientes se consideran y esta cuenta es más que suficiente para entrenar el modelo. Y para el propósito de prueba, se utiliza el conjunto de datos con 6.000 registros. Cada conjunto de datos consta de 252 atributos y casi el 50% de ellos se derivan atributos. No todos los atributos 252 se utilizan para el modelado. Solo los atributos relevantes se extraen a partir de (incluyendo tanto grupo actual y derivado) el conjunto de datos.

- **Extracción de variables - Análisis Exploratorio de Datos**

Al referirse a los trabajos de investigación anteriores sobre este estudio, basado en la inferencia manual y la información obtenida de la empresa "s personales de telecomunicaciones, hemos seleccionado las posibles variables para modelar el árbol de decisión. Entre ellos, las variables más importantes que tienen una mayor contribución a predecir el churn se seleccionan. Las variables seleccionadas se agrupan en 4 categorías y se describen a continuación.

a) Demografía Cliente

   - ✓ Edad - Se encontró que los clientes entre el grupo de edad de 45 a 48 tienen una alta probabilidad de pérdida de clientes.

- ✓ Line_Tenure - Los clientes de 25 a 30 meses de período de tenencia a punto de batir.
- ✓ Customer_Class - En general, la probabilidad de deserción de los titulares de las cuentas corporativas es alta. Esto se debe al hecho de que su cuenta se mantiene por la empresa y los clientes que dejan la compañía churn. El Customer_Class puede ser cualquiera de VIP / Individual / Empresa.
- ✓ Days_to_Contract_Expiry - La mayoría de los clientes se suscriben a un nuevo servicio con la intención de adquirir nuevos HAND_SET. Estas personas salen de la red después de que expire el contrato.

b) Declaración y pago

- ✓ Avg_Pay_Amount - Si el promedio de los pagos que hace el cliente mensualmente en los últimos 6 meses es de menos de $100 o está entre $520 o $550, el cliente tiene una alta propensión a cambiar de banco.
- ✓ Overdue_Payment_Count - Si la cantidad de pagos es mayor que 0 y menor que 4 en los últimos 6 meses, el cliente puede cambiar de banco.

- **Construcción del modelo**

El modelo creado para este estudio se muestra en la Figura 2.3. El conjunto de reglas descrito anteriormente para las variables de características se utilizan para la formación del modelo de árbol de decisión y el modelo de red neuronal. Como no existen métodos estadísticos aplicados a la selección de conjunto de características, el aumento de la información y la entropía de los atributos se calculan para probar la eficacia en la búsqueda de la pérdida de clientes. Como ya se ha mencionado, los datos se agregan durante seis meses. Esto significa que el comportamiento de los clientes "s durante los últimos 6 meses se utilizó para predecir los churners durante el séptimo mes. Los siguientes son los pasos a seguir para la pérdida de clientes predicción

I. Inicialmente, para cada atributo, se asigna un valor de umbral.
II. Los valores de los atributos de la formación de datos se comparan con el "umbral s atributo para declarar que un cliente va a batir o no. Simple if ... then ... else normas se aplican en este proceso.
III. Un modelo se construye a continuación para la formación de datos.

IV. El modelo se aplica a continuación, en el conjunto de datos de prueba y los resultados se enumeran.

V. Los pasos anteriores pueden ser repetidos mediante la variación de los valores de umbral de los atributos seleccionados.

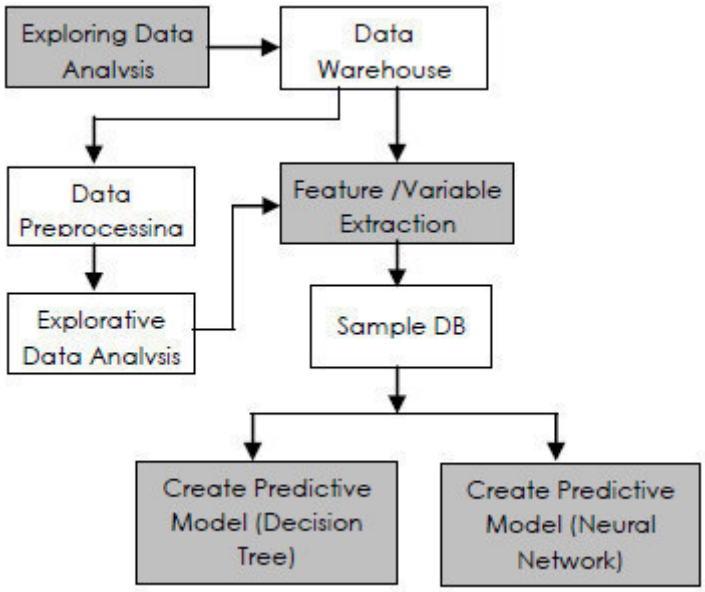


Figura 2.3 Modelo de predicción churner

Resultados de la evaluación

Los autores evalúan el rendimiento de dos técnicas de minería de datos: árboles de decisión y redes neuronales, que se utilizan en este estudio para construir el modelo de predicción de churn. En la **Figura 2.4** se muestra el árbol de decisión generado para nuestra base de datos.

El rendimiento de un modelo de clasificación se basa en los recuentos de registros de prueba correcta e incorrectamente predicho por él. Estos recuentos se tabulan como una tabla llamada matriz de confusión. En la Tabla 2.6 y Tabla 2.7 se presenta la matriz de confusión para los modelos de red neuronal de árbol de decisión y la base, respectivamente, en los atributos de datos demográficos. Esta matriz le ayuda a encontrar el valor predictivo y la tasa de error de los modelos de clasificación.

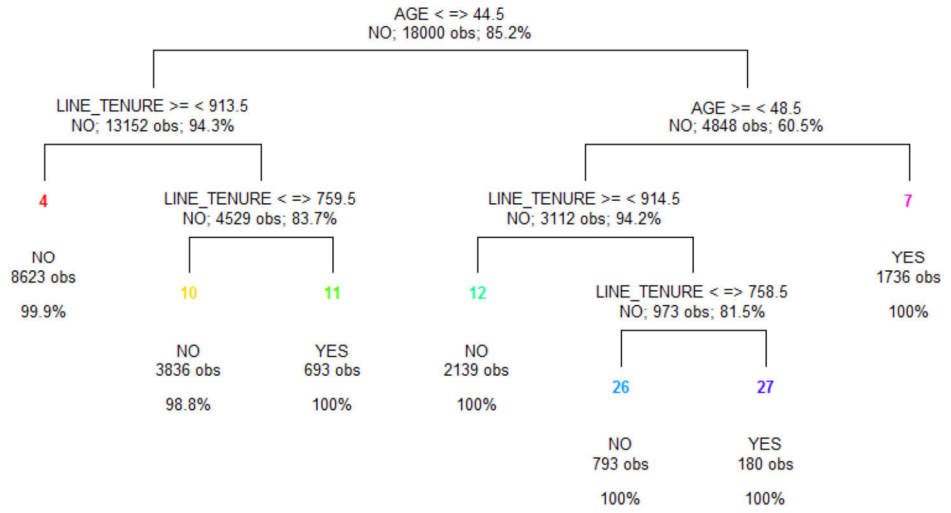


Figura 2.4 Árbol de decisión de la formación de datos para los atributos demográficos de los clientes.

Decision Tree		Predicted Class	
		Churn	Non Churn
Actual	Class = Churn	833	19
	Class = Non Churn	48	5100

Tabla 2.6 Matriz de confusión para el modelo de árbol de decisión para los atributos de los datos demográficos

La precisión del modelo se calcula utilizando la siguiente fórmula:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{833 + 5100}{833 + 19 + 48 + 5100} = 98.88\%.$$

$$\text{Error rate} = \frac{\text{Number of wrong predictions}}{\text{Total number of predictions}} = \frac{48 + 19}{6000} = 1.116667\%$$

De los cálculos anteriores, se observó la exactitud de predicción de 98,88% y la tasa de error del 1,11167% de nuestro modelo de árbol de decisión. También tiene la falsa positiva de 0,93% y falso negativo de 2,23%. Del mismo modo, la exactitud de predicción y otras medidas se calculan para el modelo de red neuronal.

Neural Network		Predicted Class	
		Churn	Non Churn
Actual	Class = Churn	823	29
	Class = Non Churn	65	5083

Tabla 2.7 Matriz de confusión para el modelo de red neuronal para atributos de datos demográficos.

En la Tabla 2.7, se observa que la precisión prevista es 98.43%, falso positivo es 1,26% y falso negativo es 3,40%. Y la tasa de error es de 1,5616%.

2.2.4. MÉTODO UNIFORMLY SUBSAMPLED ENSEMBLE (USE SVM + PCA)

Los autores Jaewook Lee, Namhyoung Kim, Kyu-Hwan Jung y Yong Seog Kim proponen un nuevo modelo de conjunto "Uniformly subsampled ensemble", denominado USE, también presentan la estructura del modelo y describen sus características únicas en términos de toma de muestras y sistemas de ponderación. La **Figura 2.5** presenta gráficamente la estructura del modelo USE. El primer paso en la construcción de USE es dividir el conjunto de datos en subconjuntos para capacitar a un único clasificador correspondiente. Una vez que un solo clasificador está calibrado para producir la puntuación estimada (por ejemplo, probabilidad de deserción) para cada registro del cliente de cada partición, el modelo de conjunto USE agrega las puntuaciones de cada clasificador y produce el resultado final del modelo de conjunto.

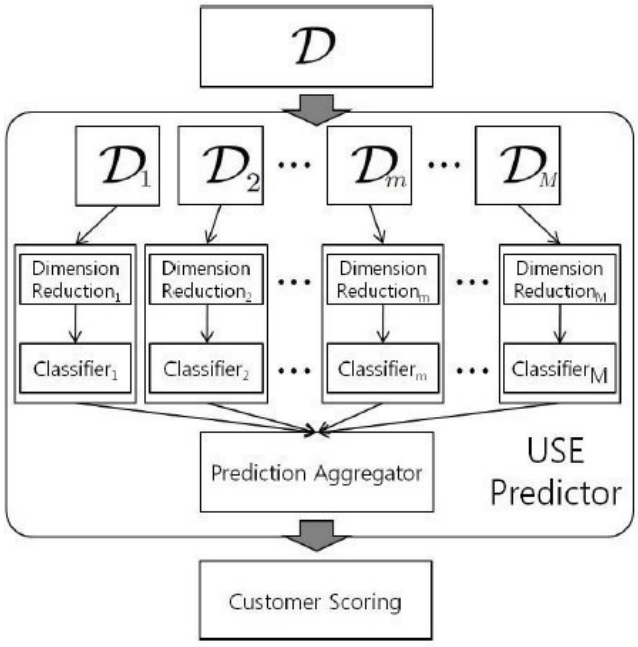


Figura 2.5 La estructura de la propuesta método de conjunto.

Los autores presentan el modelo USE trabajando juntamente con SVM, para lo cual en los experimentos realizados presentan el proceso y los resultados que el conjunto uniformemente submuestreado SVM (USE SVM) propuesto se aplica al mercado de las telecomunicaciones datos. La Figura 2.6 muestra la matriz de correlación de las variables. Se puede apreciar que hay una alta correlación entre características. Es compatible con la necesidad de extraer correlacionadas nuevas características.

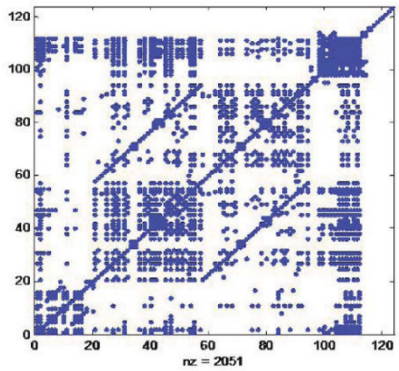


Figura 2.6 La matriz de correlación con los valores de mayor que 0,5.

Los autores aplican PCA para la reducción de la dimensión de los datos. Existen algunos tipos de métodos para seleccionar el número óptimo de PCs. Entre ellos se consideran tres enfoques que son más de uso común.

La Figura 2.7 es la trama de valores propios. Los números de PCs de cada enfoque son los siguientes.

- El valor propio-uno de los criterios: 27 PCs
- Prueba Scree: 4 PCs
- Proporción de varianza explicada: 36 (90%), 48 (95%) PCs

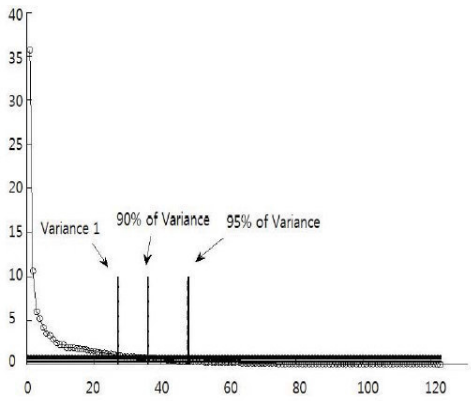


Figura 2.7 Trama de valores propios

Los autores aplicaron el método propuesto con el anterior número de PCs, luego se compararon sus tasas de acierto. Los resultados para un número diferente de PCs se presentan en la Figura 2.8. Como se muestra en el gráfico, la tasa de éxito al 30 % es mayor cuando se utilizan 48 PCs. Tiende a aumentar cuando el número de PCs aumenta. Por lo tanto, 48 PCs son seleccionados en este estudio. Después de elegir el número de PCs, el óptimo número de SVMs, M debe ser considerado. Se exploró los efectos de número de clasificadores en la exactitud de predicción, mientras que el número de PCs fue fijado como 48. El conjunto de datos de entrenamiento es dividido en los diferentes grupos de M por un muestreador al azar. La tasa de éxito al 10% es mayor cuando M es 49, es decir 49, SVMs pero 25 SVMs muestran mejor tasa de éxito al 30%. Por lo tanto, nuestro último modelo óptimo es un modelo de conjunto de 25 SVMs con 48 equipos. Por otro lado, también se muestra los resultados para un número de clasificadores (Figura 2.9).

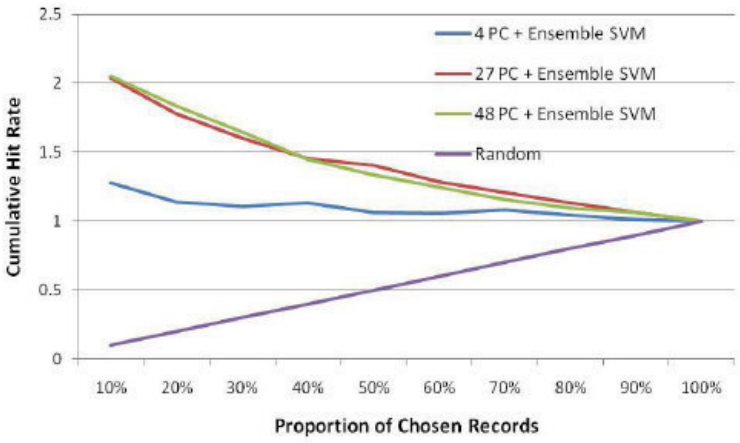


Figura 2.8 Efecto del número de PCs

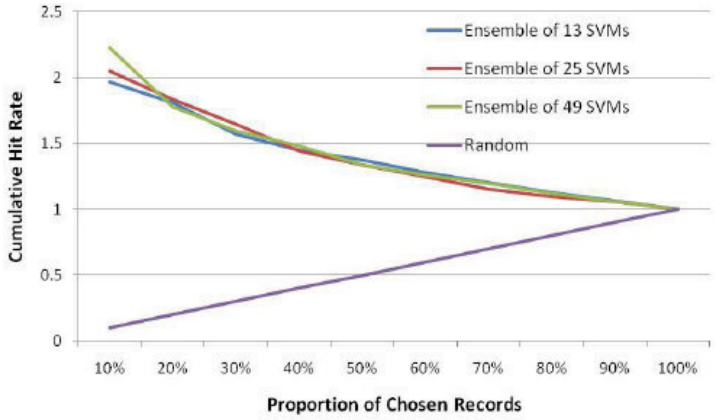


Figura 2.9 Efecto del número de clasificadores

También los autores analizaron el efecto de los métodos de ponderación. La Figura 2.10 representa un gráfico de la tasa de éxito acumulativa para diferentes métodos de ponderación. PCA en la gráfica es un método de peso uniforme. Como se muestra en la figura, los métodos de ponderación no afectan en gran medida al rendimiento. Sin embargo, el método de peso uniforme es fácil de aplicar y su rendimiento es un poco mejor que otros métodos. Así que los autores decidieron aplicar el método propuesto que utiliza el método de peso uniforme.

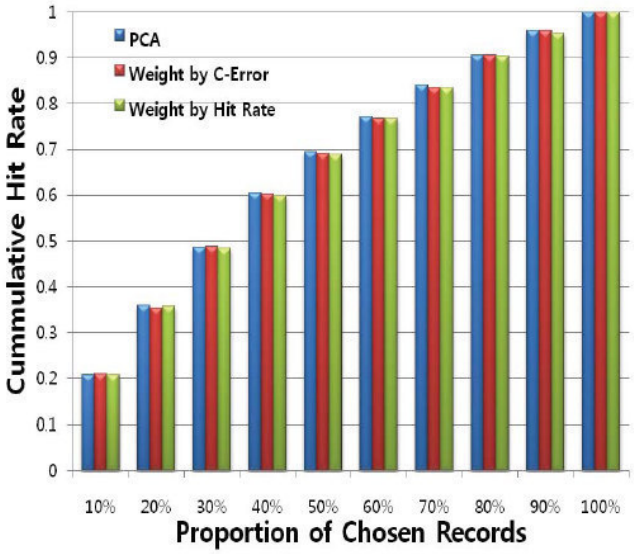


Figura 2.10 Efecto de métodos de ponderación

Para explorar hasta qué punto el propuesto PCA y el modelo de conjunto contribuyen al aumento de rendimiento en comparación con un solo SVM, los autores compararon las actuaciones de cinco modelos: USE SVM + PCA, USE SVM, solo SVM + PCA, SVM, y un modelo al azar. En la Figura 2.11, las tasas de éxito fueron notablemente mejoradas en ambos casos: USE SVM y USE SVM + PCA. En el caso de utilizar SVM con el PCA, el rendimiento se incrementa ligeramente en comparación con los resultados de los dos anteriores métodos.

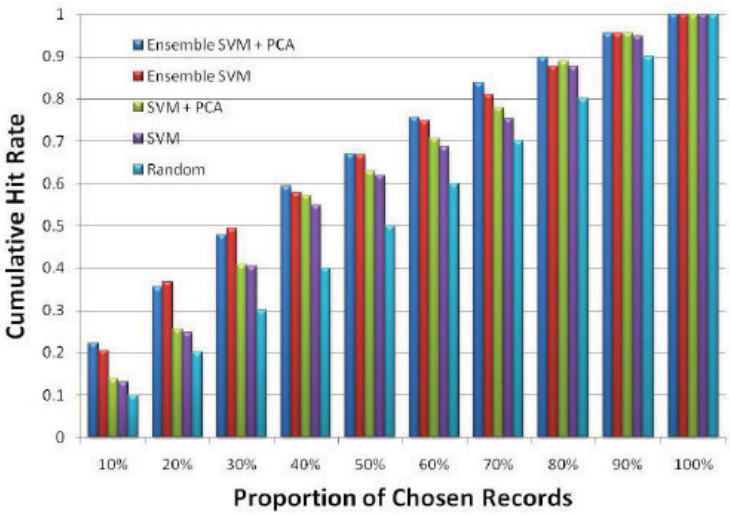


Figura 2.11 Ganancia por PCA y Conjunto

El rendimiento del método propuesto era comparado con el rendimiento de otros clasificadores. Teniendo en cuenta que el conjunto de datos es a gran escala y altamente desequilibrado, solo el 1,8% de las observaciones no eran desertores. Así, métodos convencionales simples no funcionarán correctamente.

En los estudios anteriores, el modelo Partial Least Square (PLS) y el modelo logístico, los modelos más populares en área de comercialización, se han propuesto para resolver este problema [13]. También los autores aplican el modelo de conjunto de varios SVDD (Support Vector Domain Descripción) a nuestro problema. La Figura 2.12 presenta una tasa de éxito de cinco modelos diferentes, respectivamente: USE SVM + PCA, Conjunto Multi-SVDD, PLSall, el modelo logístico y un modelo al azar. El uso propuesto SVM + PCA superado a otros métodos, y se nota un mayor la mejora del rendimiento en baja proporción. Como los autores mencionaron antes, la tasa de éxito en baja proporción es más importante medida que su en una gran proporción. A través de este método propuesto, USE supera el método convencional no solo teóricamente, sino también prácticamente.

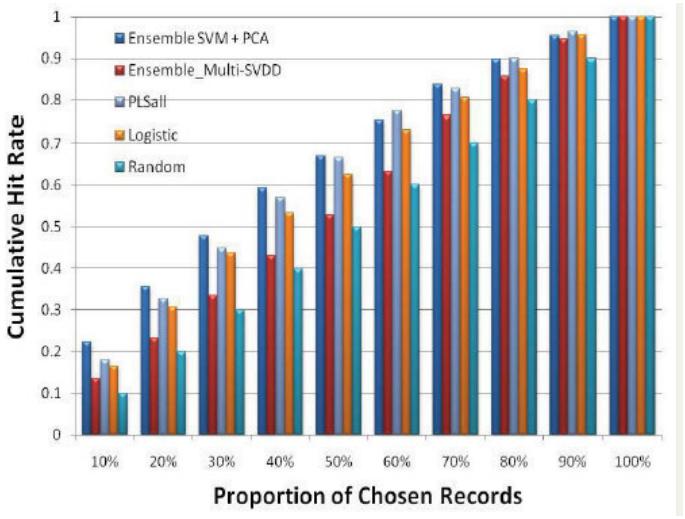


Figura 2.12 Comparación con otros métodos

2.2.5. COMPARACION DE METODOS DE LA REVISION DE LA LITERATURA

Año	Nombre de la técnica	Fuente	Resultados (Efectividad %)
2009	ROUGH SET THEORY (RST) Y LEAST SQUARE SUPPORT VECTOR MACHINE (LS-SVM)	Ning Wang y Dong-xiao Niu	89.90%
2011	ADAPTIVE NEURO FUZZY INFERENCE SYSTEM (ANFIS)	K. Hossein Abbasimehr, Mostafa Setak y M. J. Tarokh	91%
2012	REDES NEURONALES Y ÁRBOLES DE DESICIÓN – METODOLOGÍA KDD	Applications of Data Mining Techniques in Telecom Churn Prediction	98.43%
2012	MÉTODO USE SVM + PCA	Namhyoung Kim, Jaewook Lee, Kyu-Hwan Jung, Yong Seog Kim	99.75%

Tabla 2.2.5.1 Comparación de técnicas de la revisión de la literatura – Fuente Propia

Capítulo 3: Modelo USE SVM + PCA

En el presente capítulo nos enfocaremos en la descripción de la técnica a usar en el diseño del modelo predictivo para la identificación del cliente desertor.

Para una mejor comprensión del modelo se presentará la definición del Análisis de Componentes Principales (PCA) y Máquina de Soporte Vectorial (SVM).

3.1 Análisis De Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica cuya finalidad es transformar un conjunto de variables, a las que se las denomina variables originales interrelacionadas, en un nuevo conjunto de variables que son la combinación lineal de las originales, denominadas componentes principales. Para obtener tales combinaciones, es necesario construir la matriz de varianzas y covarianzas de esas variables. Estas nuevas variables tienen la característica de no estar correlacionadas entre sí.

En el PCA, se persigue explicar la mayor parte de la variabilidad total con el menor número de componentes, en donde cada componente como se dijo anteriormente está expresada en función de las variables observadas y es muy adecuado para resumir y reducir datos.

Algebraicamente, las componentes principales son una combinación lineal de las *p* variables aleatorias originales $X_1, X_2,\ldots, X_p$ y geométricamente esta combinación lineal representa la elección de un nuevo sistema de coordenadas obtenidas al rotar el sistema original. Estos nuevos ejes representan la dirección de máxima variabilidad. Por lo tanto, el PCA permite describir la estructura e interrelación de variables originales consideradas simultáneamente, determinando *q* combinaciones lineales de las *p* -variables originales que expliquen la mayor parte de la variación total, y de esta forma resumir y reducir los datos.

Sea $X^T = [X_1 X_2 \ \ldots \ X_p]$ un vector aleatorio p-variado, donde las variables que lo componen son las variables aleatorias originales y no necesariamente normales. El vector *p*-variado **X** tiene como matriz de varianzas y covarianzas a Σ, donde se tiene que $\lambda_1 \geq \lambda_2 \geq \ldots. \geq \lambda_p \geq 0$ y $a_1, a_2, \ldots a_p$ son los valores y vectores propios de Σ, respectivamente.

Ahora, consideremos las siguientes combinaciones lineales:

$$Y_1 = a_1{}^T X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$
$$Y_2 = a_2{}^T X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$
$$\vdots$$
$$Y_p = a_p{}^T X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

Entonces las variables Y_1, Y_2,…,Y_p son las componentes principales, las mismas que no están correlacionadas entre sí, son ortonormales entre ellas y además se cumple que:

$$Var(Y_i) = a_i^T \sum a_i = \lambda_i \; i = 1,2,...., p \quad \ldots (3.1.1)$$

$$Cov(Y_i, Y_j) = a_i^T \sum a_j = 0 \quad i \neq j, \quad i, j = 1,2,...., p \quad \ldots (3.1.2)$$

Donde se cumple que:

$\|a_i\| = 1$ para i=1,2,….,p y $\langle a_i, a_j \rangle = 0$ para $i \neq j$.

$\|a_i\|$ es la norma del vector a_i y $\langle a_i, a_j \rangle$ es el producto interno entre los vectores a_i y a_j.

La primera componente principal es la combinación lineal de $Y_1 = a_1^T X$ que maximiza la varianza de Y_1, donde $\|a_1\|=1$.

La segunda componente principal es la combinación lineal $Y_2 = a_2^T X$ que maximiza la varianza de Y_2, donde $\|a_2\|=1$ y la Cov (Y_1, Y_2)=0.

En general, la i-ésima componente principal es la combinación lineal que maximiza la varianza de $Y_i = a_i^T X_i$, sujeta a que la norma del vector a_i sea unitaria y que la $Cov(Y_i, Y_k) = 0$ para k < i.

Resumiendo, tenemos que $\sum$ es la matriz de varianzas y covarianzas asociada con el vector aleatorio, X^T=[X_1 X_2 … X_p] $\in R^p$, y que $\sum$ tiene los pares de valores y vectores propios $(\lambda_1, a_i), (\lambda_2, a_2), ..., (\lambda_p, a_p)$, donde $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

El porcentaje total de la varianza contenida por la i-ésima componente principal o su explicación está dado por:

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \quad \ldots \text{(3.1.3)}$$

Y el porcentaje total de la varianza contenida por las *q* primeras componentes principales se define así:

$$\frac{\sum_{i=1}^{q} \lambda_q}{\sum_{i=1}^{p} \lambda_i} \quad \ldots \text{(3.1.4)}$$

Existen algunos criterios para determinar el número de componentes principales a retener, los cuales son:

- **En general (Proportion of variance accounted)**, el criterio más sencillo para obtener el número *m* de componentes principales a retener debe ser tal que $\lambda_1, \lambda_2, \cdots, \lambda_m$ en conjunto expliquen más del 75% de la información total de la muestra.
- **Gráfico de sedimentación (Scree Test).** En este gráfico en el eje Y se representan los valores propios o raíces características y en el eje X el número de componentes principales correspondientes a cada valor propio en orden decreciente, de acuerdo a este gráfico se retienen aquellas componentes que se encuentran antes de que el gráfico presente un "quiebre" o “codo”.
- **Media aritmética (The eigenvalue-one criterion)**. Según este criterio se retienen aquellas componentes tales que :

$$\lambda_q > \overline{\lambda} = \frac{\sum_{q=1}^{p} \lambda_q}{p} \quad \ldots \text{(3.1.5)}$$

  Y se seleccionan aquellas componentes cuya raíz característica excede de la media de las raíces características.

En nuestro caso se usará el primer criterio debido a que es el que muestra mejor resultado al trabajar con el modelo.

3.2 Máquinas de Soporte Vectorial (SVM)

Las Máquinas de Soporte Vectorial (SVM) son una moderna y efectiva técnica de IA que ha tenido un formidable desarrollo en los últimos años. A continuación se presentarán los fundamentos teóricos que definen estos sistemas de aprendizaje.

Uno de los conceptos fundamentales en esta técnica es el algoritmo Vector de Soporte (VS) es una generalización no-lineal del algoritmo Semblanza Generalizada, desarrollado en la Rusia en los años sesenta. El desarrollo de los VS trae consigo el surgimiento de las Máquinas de Soporte Vectorial. Estas son sistemas de aprendizaje que usan un espacio de hipótesis de funciones lineales en un espacio de rasgos de mayor dimensión, entrenadas por un algoritmo proveniente de la teoría de optimización.

El algoritmo se enfoca en el problema general de aprender a discriminar entre miembro positivos y negativos de una clase de vectores de n-dimensional dada. Las SVM pertenecen a la familia de clasificadores lineales. Mediante una función matemática denominada kernel, los datos originales se redimensionan para buscar una separabilidad lineal de los mismos. Una característica de las SVM es que realiza un mapeo de los vectores de entrada para determinar la linealidad o no de los casos los cuales serán integrados a los Multiplicadores de Lagrange para minimizar el Riesgo Empírico y la Dimensión de Vapnik-Chervonenkis. De manera general, las Máquinas de Soporte Vectorial permiten encontrar un hiperplano óptimo que separe las clases.

En esta sección se hará una revisión de la teoría básica de las SVM en problemas de clasificación [45]- [46].

Caso linealmente separable

Supongamos que nos han dado un conjunto *S* de puntos etiquetados para entrenamiento como se aprecia en la Figura 3.1.

$$s = \Big((x_1, y_,), (x_2, y_2), \dots, (x_n, y_n)\Big) \in R^N x\, R \ldots (3.2.1)$$

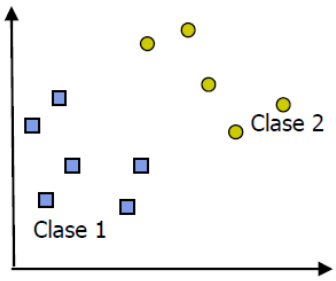


Figura 3.1Caso linealmente separable.

Cada punto de entrenamiento $x_i = \Re^N$ pertenece a alguna de dos clases y se le ha dado una etiqueta $y_i \in$ {-1,1 }para i=*1,2, . . . l*. En la mayoría de los casos, la búsqueda de un hyperplano adecuado en un espacio de entrada es demasiado restrictiva para ser de uso práctico. Una solución a esta situación es mapear el espacio de entrada en un espacio de características de una dimensión mayor y buscar el hyperplano óptimo allí. Sea $Z = \varphi(x)$ la notación del correspondiente vector en el espacio de características con un mapeo $\varphi\ de\ \Re^N$ a un espacio de características *Z,* deseamos encontrar el hyperplano:

$$w \bullet z + b = 0 \ldots (3.2.2)$$

Definido por el par (*w, b*), tal que podamos separar el punto x_i de acuerdo a la función

$$f(x_i) = sign(w \bullet z + b) = \begin{cases} 1 & y_i = 1 \\ -1 & y_i = -1 \end{cases} \ldots (3.2.3)$$

Donde w ϵ Z y b ϵ ℜ. Más precisamente, el conjunto *S* se dice que es linealmente separable si existe *(w, b)* tal que las inecuaciones

$$\begin{cases} w \bullet z + b \geq 1, & y_i = 1 \\ w \bullet z + b \geq 1, & y_i = 1 \end{cases} \quad i = 1, \ldots, l \quad (3.2.4)$$

sean válidas para todos los elementos del conjunto *S*. Para el caso linealmente separable de *S*, podemos encontrar un único hyperplano óptimo, para el cual, el margen entre las proyecciones de los puntos de entrenamiento de dos diferentes clases es maximizado.

Caso no linealmente separable

Si el conjunto *S* no es linealmente separable (ver Figura 3.2), violaciones a la clasificación deben ser permitidas en la formulación de la SVM.

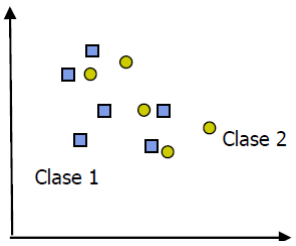


Figura 3.2 Caso no linealmente separable

Para tratar con datos que no son linealmente separables, el análisis previo puede ser generalizado introduciendo algunas variables no-negativas $\xi i \geq 0$, de tal modo que la fórmula (3.2.4) es modificado a

$$y_i(w \bullet z + b) \geq 1 - \xi i \ , \quad i = 1, \dots l \quad (3.2.5)$$

Los $\xi i \neq 0$ en la fórmula (3.2.5) son aquellos para los cuales el punto *x*, no satisface la fórmula (3.2.4). Entonces el término $\sum_{i=1}^{l} \xi_i$ puede ser tomado como algún tipo de medida del error en la clasificación.

El problema del hyperplano óptimo es entonces redefinido como la solución al problema

$$\min\left\{\frac{1}{2} w \cdot w + C \sum_{i=1}^{l} \xi_i\right\}$$

$$s.a y_i(w \bullet z + b) \geq 1 - \xi i \ , \quad i = 1, \dots l (3.2.6)$$

$$\xi i \geq 0 \ , \quad i = 1, \dots l$$

Donde *C* es una constante. El parámetro *C* puede ser definido como un parámetro de regularización. Este es el único parámetro libre de ser ajustado en la formulación de la SVM. El ajuste de éste parámetro puede hacer un balance entre la maximización del margen y la violación a la clasificación. Más detalles se pueden encontrar en [46], [47].

Buscando el hyperplano óptimo en la fórmula (3.2.6) es un problema QP, que puede ser resuelto construyendo un Lagrangiano y transformándolo en el dual.

$$\text{Max} \quad W(\propto) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j \, y_i y_j z_i \bullet z_j$$

…(3.2.7)

$$s.a \quad \sum_{j=1}^{l} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i=1, \ldots l$$

Donde $\alpha \quad = (\alpha_1, .., \alpha_l)$ es una vector de multiplicadores de Lagrange positivos asociados con las constantes en la fórmula (3.2.5).

El teorema de Khun-Tucker juega un papel importante en la teoría de las SVM. De acuerdo a este teorema, la solución $\overline{\alpha_1}$ del problema (3.2.7) satisface:

$$\bar{\alpha}_i \left(y_i \left(\bar{w} \bullet z_i + \bar{b} \right) - 1 + \bar{\xi}_1 \right) = 0, \quad i = 1, .., l \quad \ldots (3.2.8)$$

$$(C - \bar{\alpha}_i)\bar{\xi}_1 = 0, \quad i = 1, .., l \quad \ldots (3.2.9)$$

De esta igualdad se deduce que los únicos valores $\bar{\alpha}_i \neq 0$ (3.2.9) son aquellos que para las constantes en (3.2.5) son satisfechas con el signo de igualdad. El punto $x_i$correspondiente con $\bar{\alpha}_i$>0 es llamado *vector de soporte.* Pero hay dos tipos de vectores de soporte en un caso no separable. En el caso 0 <$\bar{\alpha}_i$<*C*, el correspondiente vector de soporte x_i, satisface las igualdades $y_i \left(\bar{w} \bullet z_i + \bar{b} \right) = 1$ y $\xi_i = 0$. En el caso $\bar{\alpha}_i = C$, el correspondiente ξ_i es diferente de cero y el correspondiente vector de soporte $x_i$no satisface (3.2.4). Nos referimos a estos vectores de soporte como errores. El punto x_i correspondiente con $\bar{\alpha}_i = 0$ es clasificado correctamente y está claramente alejado del margen de decisión (ver Figura 3.3).

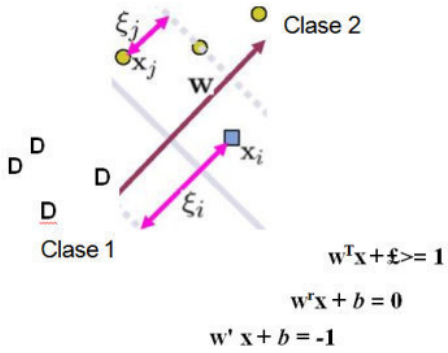


Figura 3.3 Aparición del parámetro de error $\xi_i$en el error de clasificación

Para construir el hyperplano óptimo $\bar{w} \bullet z_i + \bar{b}$, se utiliza:

$$\bar{w} = \sum_{j=1}^{l} \bar{\alpha}_i y_i z_i \ \ldots (3.2.10)$$

Y el escalar *b* puede ser determinado de las condiciones de Kuhn-Tucker (3.2.9).

La función de decisión generalizada de (3.2.3) y (3.2.10) es tal que:

$$f(x_i)=sign(w\bullet z+b)=sign\left(\sum_{i=1}^{l}\alpha_i y_i z_i\right) \qquad (3.2.11)$$

Función Kernel para el caso no linealmente separable

Como no tenemos ningún conocimiento deφ, el cálculo del problema es (3.2.7) y (3.2.11). Hay una buena propiedad del SVM, la cual es que no es necesario tener ningún conocimiento acerca de φ. Nosotros solo necesitamos una función $K(\cdot,\cdot)$ llamada *kernel* (ver **Figura 3.4**) que calcule el producto punto de los puntos de entrada en el espacio de características Z, esto es:

$$z_i \bullet z_i = \varphi(x_i) \bullet \varphi(x_j) = K(x_i, x_j)$$

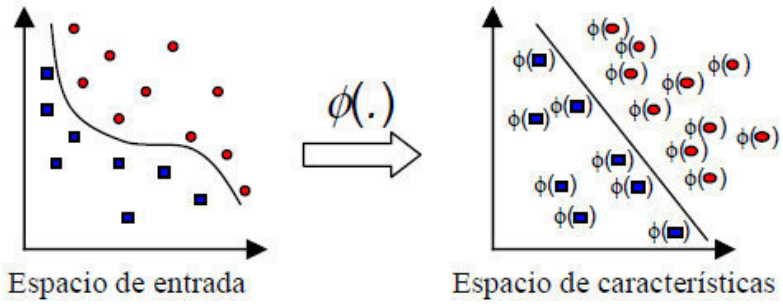


Figura 3.4 Idea del uso de un *kernel* para transformación del espacio de los datos

Función común kernel incluye los siguientes tipos:

$$Polynomial\ Kernel: k(x,y) = (\sum_{i=1}^{n} x_i y_i)^d$$

$$RBF\ Kernel: k(x,y) = exp(-\frac{\|x_i - x_j\|}{2\sigma^2})$$

$$Cauchy\ Kernel: k(x,y) = \prod_{i=1}^{n} \frac{1}{1 + d(x_i - y_i)^2}$$

$$Sigmoid\ Kernel: k(x,y) = \tan(\gamma(x_i, x) + c)$$

Las funciones que satisfacen el teorema de Mercer pueden ser usadas como productos punto y, por ende, pueden ser usadas como kernels. Podemos usar el kernel polinomial de grado d

$$K(x_i x_i)=(1 + x_i \bullet x_i)^d \quad (3.2.13)$$

para construir un clasificador SVM.

Entonces el hyperplano no lineal de separación puede ser encontrado como la solución de:

$$\text{Max} \quad \text{W}(\propto) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j \, y_i y_j K(x_i x_i) \quad (3.2.4)$$

$$s.a \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i=1,\ldots,l$$

y la función de decisión es:

$$f(x) = sign(w \bullet z + b) = sign\left(\sum_{i=1}^{l} \alpha_i y_i \, K(x_i x_i) + b\right) \quad (3.2.15)$$

Para nuestro trabajo se usará el RBF Kernel, dado que está demostrado que produce una mejor performance cuando trabaja con casos de predicción.

3.3 Modelo USE SVM + PCA

Este modelo es un clasificador ensemble SVM combinado con un análisis de componentes principales (PCA) no solo para reducir la alta dimensionalidad de los datos, sino también para aumentar la fiabilidad y la precisión de modelos calibrados en los conjuntos de datos con muy desigual distribuciones de clase. De acuerdo con experimentos, el rendimiento del modelo de USE SVM con PCA es superior a todos los modelos de comparación y el número de componentes principales (PCs) afectan a la precisión de los modelos de conjunto.

Para la creación del Modelo USE SVM se describirá el conjunto de datos, el procedimiento de procesamiento y la evaluación de la métrica.

1. Obtención de la Data

   Para predecir la pérdida de clientes, tenemos que establecer los criterios de pérdida al principio. De esta manera, se seleccionará el conjunto de datos que se usará.

2. Pre procesamiento de datos

   Para un análisis más detallado, se realiza el procesamiento previo de los datos en bruto antes de aplicar el método propuesto de la siguiente manera. En primer lugar, se elimina las variables continuas con más de 20% de valores faltantes. En segundo lugar, las variables categóricas con una alta tasa de falta también se eliminaron

porque cada variable categórica tiene muy poco poder de predicción en general [28].

3. Evaluación

   Se usa la tasa de éxito como una métrica de evaluación para nuestra investigación. La tasa de éxito es una medida popular para evaluar la capacidad de predicción de los modelos numéricos para el campo de la comercialización [18]. La tasa de éxito se calcula como:

$$Tasa\ de\ éxito\ = \sum_{i=!}^{n} {H_i}/{n} \quad \ldots \quad (3.3.1)$$

   Donde Hi es 1 si la predicción es correcta y 0 de otro modo. 'n' representa el número de muestras en los conjuntos de datos. En otras palabras, la tasa de éxito representa el porcentaje de los clientes correctamente predichos de los clientes candidatos. La tasa de éxito está asociada con un punto objetivo. Por ejemplo, una tasa de aciertos en un punto objetivo de x% es una tasa de éxito solo cuando el x% de los mejores clientes son considerados para la evaluación en función de sus probabilidades de deserción. Por lo tanto, si asumimos que se tiene 10,000 observaciones, una tasa de éxito en un punto objetivo del 10% es el porcentaje de predicciones correctas de desertores de 1.000 clientes que tienen más probabilidades para desertar. Teniendo en cuenta las tasas de éxito con los puntos objetivos, es importante porque los directores de marketing tienen que centrarse solo en el porcentaje superior de los clientes debido a presupuesto limitado y la falta de tiempo.

Para el entendimiento de la técnica se explicará el método conjunto propuesto:

3.3.1 Método de Conjunto Uniforme de Submuestras (USE)

Se presentará la estructura del nuevo modelo conjunto, **USE**, y se describirá sus características únicas en términos de toma de muestras y sistemas de ponderación. La Figura 3.5 presenta gráficamente la estructura del modelo **USE**. El primer paso en la construcción de **USE** es dividir el conjunto de datos en subconjuntos para capacitar a un único clasificador correspondiente. Una vez que un solo clasificador está calibrado para producir la puntuación estimada (por ejemplo, probabilidad de deserción) para cada registro del cliente de cada partición, el modelo de conjunto **USE** agrega las puntuaciones de cada clasificador y produce el resultado final del modelo de conjunto.

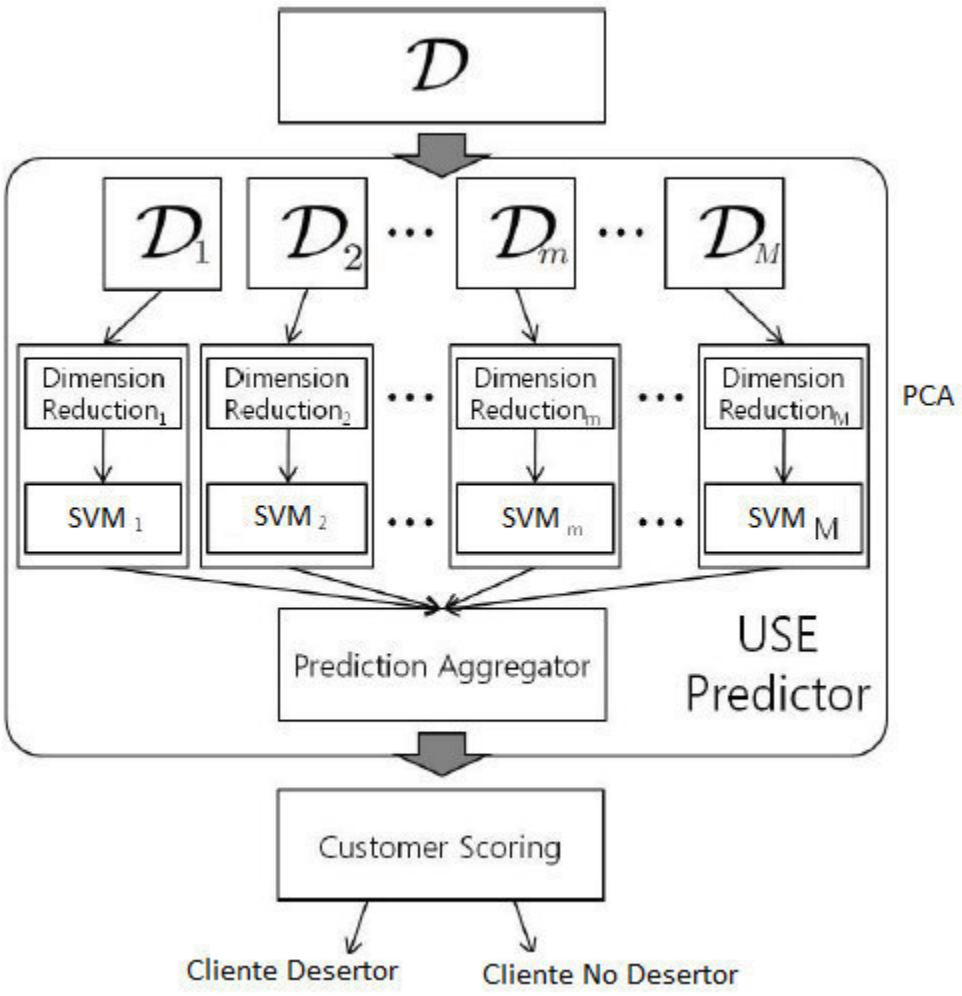


Figura 3.5 La estructura del método conjunto propuesto

3.3.2 Métodos de ponderación

Para generar una decisión colectiva, consideramos varias formas de agregar las predicciones de los modelos capacitados de clasificación a través de varios esquemas de ponderación como pesos uniformes, ponderados por clasificación de rendimiento, o ponderados por el índice de aciertos. El sistema de ponderación más simple es el método de peso uniforme que aplica el mismo peso (= 1/M) a la predicción a partir de todos los clasificadores. La predicción de cada clasificador individual puede ser ponderada en función en el rendimiento de la clasificación binaria o la tasa de éxito en los datos de validación de la muestra de los datos de entrenamiento. Para aplicar el sistema de ponderación basado en rendimiento de clasificación, la exactitud de la clasificación de los datos de validación de cada clasificador es normalizada para facilitar sumando a 1 y la predicción final en el conjunto de datos de prueba se pondera de acuerdo a este peso normalizado. En el sistema de ponderación sobre la base de la tasa de éxito, la tasa de éxito

en el 10%, 20%, y 30% se suman para medir el rendimiento. Posteriormente, se normalizan antes de la suma a 1. La predicción final sobre el conjunto de datos de prueba es ponderada según este peso normalizado como sigue:

$\hat{f}(x) = \sum_{m=1}^{M} w_m f_m(x)$ ...(3.3.2.1)

3.3.3 Bagging y Boosting vs. USE

Para construir un modelo de conjunto preciso basado en nuestro método de uso propuesto, se divide un completo conjunto de datos de entrenamiento en M submuestras dimensionados igualmente que no se superponen usando un muestreador al azar.

Por consiguiente, cualquier clasificador único (por ejemplo, un clasificador SVM) se puede calibrar en cada uno de los datos submuestreados establecidos para determinar los patrones ocultos. Por último, la predicción de todos los clasificadores se agregará a través de una suma ponderada para la construcción de la predicción final como un modelo de conjunto para cada registro en una base de datos de prueba. En este sentido, el método propuesto **USE** es muy similar a dos métodos populares de conjuntos, llamados, Bagging [29] y Boosting [30] que han sido conocidos por obtener mejores resultados que los clasificadores individuales [31], [32]. Por ejemplo, los modelos basados en conjuntos Bagging entrenan cada clasificador en el grupo de aprendizaje extraído al azar que consiste en el mismo número de ejemplos al azar tomados del conjunto de entrenamiento original, con la probabilidad de sacar cualquier ejemplo dado es iguales. Dado que las muestras se toman con el reemplazo, algunos ejemplos se pueden seleccionar múltiples veces mientras que otros no se pueden seleccionar en absoluto. Bagging combina las predicciones de varios clasificadores usando una votación con un peso igual. En resumen, la principal diferencia entre el Bagging y el propuesto método **USE** es que si las muestras se toman o no con sustitución y si el tamaño de la muestra conjunto de entrenamiento para cada clasificador individual es igual o no al tamaño del conjunto de entrenamiento original.

Por otro lado, nuestro método propuesto **USE** es diferente del método Boosting [29] que produce una serie de clasificadores con cada conjunto de entrenamiento basado en el rendimiento de los clasificadores anteriores. A través de remuestreo adaptativo en el Boosting, ejemplos que son incorrectamente predichos por los clasificadores anteriores son muestreados con más frecuencia, mientras que el método uniforme de submuestreo sin sustitución es explotado en el **USE**. En general, cada clasificador en el modelo USE es calibrado en el grupo de aprendizaje más pequeño en comparación con los clasificadores

de Bagging y Boosting, que requiere menos energía de la CPU y la memoria principal. El modelo USE todavía puede reducir el esperado error de predicción de un único predictor.

Los tres modelos de conjuntos- Bagging, Boosting, y USE-comparten una característica común: la eficacia y la mejora de la precisión de la propuesta modelo de conjunto viene sobre todo de la diversidad causada por remuestreo ejemplos de entrenamiento.

Si bien es perfectamente razonable calibrar un solo clasificador en un conjunto de muestra de entrenamiento sin más procesamiento previo, consideramos un método de reducción de dimensión de datos tal como el análisis de componentes principales (PCA). Se tiene en cuenta que la PCA es un procedimiento matemático para transformar un conjunto de predictores correlacionados en un conjunto de nuevas variables no correlacionadas llamada componentes principales (PC) que capturan la máxima cantidad de variación en los datos. Desde que el número de PC es menos que o igual al número de variables originales, y cada PC no está correlacionado con otros PC, el método PCA puede ser particularmente útil para reducir la alta dimensionalidad de los conjuntos de datos en la que se correlacionan muchas variables de entrada. Se tiene en cuenta que la reducción de dimensionalidad se puede lograr seleccionando un menor número de PCs que las originales variables de entrada, y que tres métodos han sido ampliamente utilizados para determinar el número de PC. El primer criterio, "el valor propio de un criterio" o el criterio de Kaiser-Guttman [33] selecciona todos los PCs con un valor propio mayor que 1. El segundo enfoque es basado en "La Prueba Scree" [34] y se selecciona todas las PCs teniendo en cuenta la ruptura definitiva entre ordenados valores propios de las PC. El último criterio conserva componentes si se supera una determinada proporción de variación en los datos, donde la proporción es calculada como sigue:

$$Proporcion = \text{valor propio para el componente de intereses} \Big/ \text{valores propios totales de la matriz de correlación}$$

En la implementación actual del modelo USE en el presente trabajo, se construyó un conjunto de clasificador SVM. Principalmente, los clasificadores SVM se utilizan para construir un modelo de conjunto debido a su popularidad entre los investigadores, un rendimiento superior en comparación con otros clasificadores [35], [36]. Por otro lado, el método propuesto USE puede ser combinado con cualquier otro clasificador. Además, clasificadores SVM a menudo requieren un poder de cómputo adicional y muestran pobre

espectáculo rendimiento cuando se aplican a gran escala datos [37], [38], [39], [40]. Por lo tanto, la SVM clasificador es un candidato perfecto para probar la eficacia del método de uso a través de los datos submuestreo si el objetivo es reducir el requisito de alta potencia de cálculo.

A continuación se detallan los criterios anteriormente mencionados:

- Criterio "el valor propio de un criterio" (Kaiser-Guttman) [33]:
  selecciona todos los PCs con una valor propio mayor que 1.
- El segundo enfoque es basado en "La Prueba Scree" [34] y se selecciona todos las PCs teniendo en cuenta la ruptura definitiva entre ordenados valores propios de las PCs.
- El último criterio conserva componentes si se supera una determinada proporción de variación en los datos, donde la proporción es calculada como sigue:

$$Proporcion = \text{valor propio para el componente de intereses} \Big/ \text{valores propios totales de la matriz de correlación}$$

En nuestro caso se usará el último criterio debido que es el que muestra mejor resultado al trabajar con el modelo.

Finalmente, se expone un resumen de los pasos a seguir para la aplicación del modelo:

Paso 1. PCA para la reducción de la dimensión de los datos.
Se aplica los métodos propuestos para la reducción para luego comparar sus tasas de acierto.

Paso 2. Escoger número de PCs.
Se aplican los tres criterios mencionados anteriormente y se elige el de mayor tasa de éxito.

Paso 3. Escoger el óptimo número de SVMs 'M'
- El conjunto de datos de entrenamiento es dividido en los diferentes grupos de M por un muestreador al azar.

Paso 4. Aplicar SVM a los grupos

Paso 5. Se elige el de mayor tasa de éxito.

Capítulo 4: DISEÑO DEL MODELO USE SVM + PCA

4.1 Selección del método CRISP-DM

La metodología elegida para la implementación de Minería de Datos es CRISP-DM, proceso estándar entre industrias para este campo. Esta metodología ha sido definida por un grupo de compañías con amplia trayectoria en el uso de minería de datos. CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Minería de Datos, como se puede constatar en la gráfica presentada en la Figura 4.1. Según varios autores, entre los que se destacan Gamberger y otros (2001), Ramos y Giménez (2004), esta metodología consta de seis fases: (a) Entendimiento del negocio, (b) Entendimiento de los datos, (c) Preparación de los Datos, (d) Modelado y (e) Evaluación.

En la fase de modelado se emplea una técnica basada en algoritmos de aprendizaje supervisado, la resultante del proceso de entrenamiento del algoritmo determina a qué clasificación pertenecen los clientes del banco, para ello estamos considerando dos clasificaciones: clientes desertores (clientes que van a dejar de utilizar el servicio de préstamos del banco) y clientes no desertores (clientes fieles al servicio de préstamos del banco, ya que van a continuar utilizando dicho servicio). En la fase de evaluación se verifica la tasa de acierto con respecto a la cantidad de aciertos obtenidos en la fase de entrenamiento para cada clasificación.

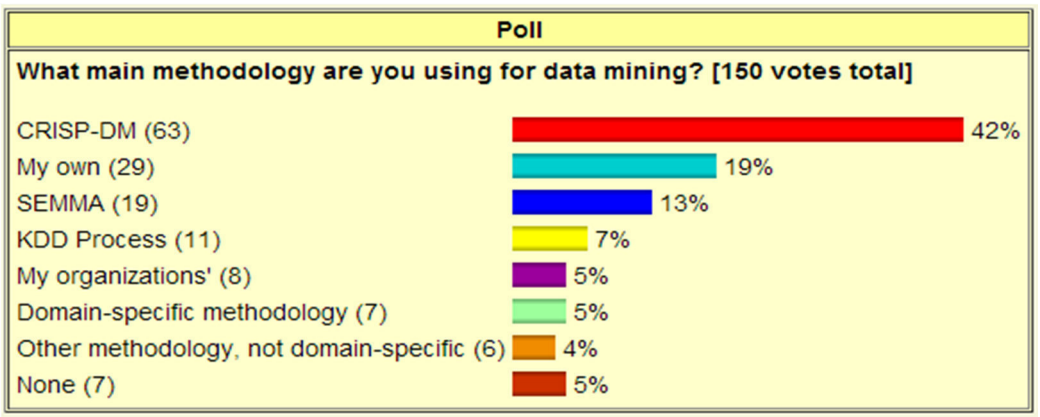


Figura 4.1 Metodología utilizada en Minería de Datos (Kdnuggets, 2007)

4.2 Desarrollo metodológico CRISP-DM de la propuestaFase I: Entendimiento del negocio

Es de interés de las entidades financieras peruanas implementar mecanismos y herramientas tecnológicas de inteligencia de negocios que contribuyan a la identificación de patrones de acciones o comportamientos mostrados por los potenciales clientes desertores a fin de evitar la deserción y tener éxito en la retención del cliente para tomar acciones preventivas para así evitar la deserción y tener éxito en la retención del cliente.

Ante la reducción de la confianza mundial en la banca y el aumento del porcentaje de clientes que planea cambiar de banco, se debe de tomar medidas que permitirán retener a los clientes. En el Perú, los bancos invierten miles de dólares para evitar el fraude electrónico pero ninguna herramienta permite reconocer este patrón y determinar los potenciales clientes desertores en un futuro establecido de los servicios de una entidad financiera.

Los bancos desean obtener nuevos clientes e incrementar la lealtad de los que ya se enrolaron a sus servicios, para este fin, las tecnologías de información están jugando un papel importante brindando sistemas atractivos para los clientes, así como oferta de productos dinámicos y apertura de nuevos canales que permitan satisfacer las necesidades de los clientes.

Determinar Objetivos del Negocio

Esta tarea corresponde a una labor de comprensión de qué es lo que los altos funcionarios quieren conseguir con la implementación de un proyecto de minería de datos para la identificación de los clientes desertores.

4.2.2 Fase II: Entendimiento de los datos

Recolectar datos iniciales

En el proceso de detección de los conjuntos de datos relevantes para formar la hipótesis de los hábitos de compra de los tarjetahabientes se ha recolectado información de las siguientes fuentes:

- BD1 - Base de datos de datos demográficos del cliente.
- BD2 - Base de datos de cuentas de ahorros, ingresos y saldos de la cuenta.
- BD3 - Base de datos de tarjetas.
- BD4 - Base de datos del motor transaccional.

- BD5 - Base de datos de créditos.
- BD6 - Base de datos de operaciones por internet

Atributos candidatos por categoría:

- Atributos demográficos:

- Identificador del cliente	- Tipo documento	- Número de documento
- Apellidos y Nombres	- Estado civil	- Sexo
- Fecha de nacimiento	- Ubigeo de residencia	- Ubigeo nacimiento
- Correo electrónico	- Teléfono	- Grado de instrucción
- Profesión	- Sector económico	- Ocupación
- Dirección de domicilio	- Dirección laboral	- Interdicto
- Fallecido	- Mancomuno	- Carga familiar

- Atributos de vínculo con el cliente:

- Productos afiliados	- Cuenta aperturada	- Tipo de moneda cta.
- Fecha de apertura cta.	- Estado cuenta	- Fecha cuenta cerrada
- Estado de la tarjeta	- Fecha tarjeta cerrada / bloqueada	- Tarjetas vinculadas

- Atributos de ingresos, saldos y préstamos:

- Identificador del crédito	- Monto de crédito	- Monto línea de crédito
- Saldo vigente de crédito	- Cuotas por pagar	- Aval/garante
- Promedio remuneración	- Saldo contable	- Saldo disponible

- Atributos de compra del cliente:

- Número de tarjeta	- Identificador de la transacción	- Fecha de transacción
- Hora de transacción	- Moneda de transacción	- Monto de la transacción
- Código de mensaje	- Código de proceso	- Situación de transacción
- Resultado de transacción	- Tipo canal de transacción	- Código de autorización
- Monto de Comisión	- Identificador del terminal	- Categoría de establecimiento
-Nombre de establecimiento	- Ciudad de establecimiento	-País de establecimiento
- Código ZIP del establecimiento	- Código de departamento y provincia del establecimiento	- Identificador del producto
- Ciudad de entrega	- País de entrega	- Correo electrónico

4.2.3 Fase III: Preparación de los datos

Seleccionar datos

En esta actividad se definen las variables a tomar en cuenta en el modelo meta-heurístico. Las variables se obtuvieron del dataset [4] de Business Intelligence Cup 2004, organizado por la Universidad de Chile en 2004. Este dataset es de un banco latinoamericano que sufrió un incremento de número de desertores que utilizaban sus tarjetas de crédito y decidieron mejorar su sistema de retención de clientes. La Tabla 1 presenta la descripción de las variables del conjunto de datos.

[4] Conjunto de datos

El conjunto de datos consiste de 21 variables independientes y 01 variable dependiente. Tiene 14814 registros de clientes que utilizan tarjetas de crédito, de los cuales 13812 son clientes fieles a la entidad financiera y 1002 son desertores, lo que significa que el 93.24% de clientes son fieles a la entidad financiera y el 6.76% son desertores.

Todas estas variables nos ayudarán a determinar si un cliente se encuentra en la clasificación de "churner"[1] o"no churner"[2]. Las variables que se presentan han sido clasificadas desde cuatro (04) perspectivas: información personal del cliente, información básica de la tarjeta, información de riesgo e información transaccional. La Figura 4.2 esquematiza los atributos agrupados en 4 categorías.

Figura 4.2 Categoría de los atributos (de las variables seleccionadas)

Las variables del 7 al 11 pertenecen al sector de información personal del cliente, o también denominada información demográfica; las variables del número 1 al 3 pertenecen a la información básica de la tarjeta; las variables del número 4 al 6 están relacionadas con el riesgo; y, finalmente, las demás variables (del número 12 al 21) están relacionadas a la información de las transacciones de la tarjeta de crédito. Por último, la variable 22 es la salida del sistema, es decir, la clasificación a la cual pertenece un cliente ya sea “Desertor” o “No Desertor”. En la Tabla 4.1 se detalla las variables a utilizar en el modelo.

Número	**Código**	**Descripción**	**Referencia Técnica**	**Valor**
1	CRED_T	Crédito en el mes T	(Nekuri Naveen,	Número real

			Vadlamani Ravi, Dudyala Anil Kumar, 2009)	positivo
2	CRED_T-1	Crédito en el mes T-1	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real positivo
3	CRED_T-2	Crédito en el mes T-2	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real positivo
4	NCC_T	Número de tarjetas de crédito en el mes T	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Valor entero positivo
5	NCC_T-1	Número de tarjetas de crédito en el mes T-1	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Valor entero positivo
6	NCC_T-2	Número de tarjetas de crédito en el mes T-2	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Valor entero positivo
7	INCOME	Ingresos del cliente[5]	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real positivo
8	N_EDUC	Nivel educacional del cliente	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	1-4
9	AGE	Edad del cliente	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número entero positivo

[5] Ingresos económicos que tiene el cliente mensualmente.

10	SX	Género del cliente	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	0-1
11	E_CIV	Estado civil del cliente	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	1-4
12	T_WEB_T	Número de transacciones web en el mes T	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número entero positivo
13	T_WEB_T-1	Número de transacciones web en el mes T-1	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número entero positivo
14	T_WEB_T-2	Número de transacciones web en el mes T-2	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número entero positivo
15	MAR_T	Margen del cliente [6]para la empresa en el mes T	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real
16	MAR_T-1	Margen del cliente para la empresa en el mes T-1	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real
17	MAR_T-2	Margen del cliente para la empresa en el mes T-2	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real

[6]Un método de crear una garantía por parte de los compradores y vendedores de contratos de futuros que se cumplieron todas las obligaciones del contrato. Márgenes de clientes se establecen de forma individual sobre la base de la cantidad de exposición al riesgo y el tamaño del contrato.

18	MAR_T-3	Margen del cliente para la empresa en el mes T-3	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real
19	MAR_T-4	Margen del cliente para la empresa en el mes T-4	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real
20	MAR_T-5	Margen del cliente para la empresa en el mes T-5	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real
21	MAR_T--6	Margen del cliente para la empresa en el mes T-6	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	Número real
VARIABLE DE SALIDA				
22	TARGET	La clasificación a la cual pertenece el cliente, ya sea "churner" o "no churner".	(Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar, 2009)	0-1

Tabla 4.1 Detalle de las variables del modelo de predicción

Para las variables N_EDUC, SX y E_CIV se detalla el significado de cada número correspondiente al rango establecido (entre valor mínimo y máximo) en la Tabla 4.1. En la Tabla 4.2 se detalla la información correspondiente a la variable N_EDUC (nivel educacional), en la Tabla 4.3 se detalla la información correspondiente a la variable SX (género) y en la Tabla 4.3se detalla la información correspondiente a la variable E_CIV (estado civil). Para el siguiente trabajo se toma un periodo mínimo de 6 meses de datos de los clientes.

NIVEL EDUCACIONAL	
Categoría	**Valor**
Estudiante Universitario	1
Bachiller	2
Técnico	3
Grado Universitario (Titulado)	4

Tabla 4.2 Detalle de la información de la variable N_EDUC

GÉNERO	
Categoría	**Valor**
Masculino	0
Femenino	1

Tabla 4.3 Detalle de la información de la variable SX

ESTADO CIVIL	
Categoría	**Valor**
Soltero	1
Casado	2
Viudo	3
Divorciado	4

Tabla 4.4 Detalle de la información de la variable E_CIV

4.2.4 Fase IV: Modelaje

Seleccionar técnica de modelaje

En el Capítulo 2 se realiza una descripción y comparación de estrategias empleadas en investigaciones científicas de temas relacionados al dominio problema identificación de patrones de comportamiento de clientes desertores, del análisis obtenido se concluyó que la mejor técnica es el algoritmo de clasificación de auto aprendizaje Support Vector Machine (SVM). La técnica a usar es el USE SVM + PCA, la cual es una mejora al SVM tradicional.

El modelo a usar será el USE SVM + PCA debido a que ha demostrado el más alto porcentaje de acierto comparado con otros modelos de clasificación [27]. También elegimos esta técnica porque la combinación de los métodos USE y PCA confrontan dos debilidades del SVM que son: requiere un alto energía computacional y muestran rendimiento medio cuando son aplicados a una gran cantidad de datos [37, 38, 39, 40].

El detalle de la aplicación de la técnica a nuestro problema se desarrollará en el punto 4.3.

4.2.5 Fase V: Evaluación

En el Capítulo 6 se detalla los experimentos realizados en la prueba del modelo USE PCA SVM.

4.2.6 Fase VI: Implementación

Esta fase consiste en la presentación del conocimiento obtenido de forma clara y precisa a todos los actores dentro de la organización, quienes utilizarán el software. Se puede presentar un informe de resultados, reportes que necesite la organización para la toma de decisiones e instruir al usuario para que pueda manejar de forma correcta el sistema y obtener los resultados deseados.

Es importante al final de esta fase tener desarrollada la documentación del proyecto para dar independencia al usuario final en la utilización y generación de nuevos procesos de explotación de datos. El ámbito definido en el presente trabajo no abarca esta fase.

4.3 Aplicación del modelo USE SVM + PCA

4.3.1 Justificación

La elección de esta técnica se basó en el porcentaje de la tasa de acierto que mostraban en problemas similares al nuestro [27].

La técnica Support Vector Machine (SVM) por sí sola genera un porcentaje de acierto muy alto comparado con otras técnicas de clasificación y algoritmos de aprendizaje. [60][61] El método USE y la técnica del PCA ayudan a que la tasa de precisión de acierto sea mayor a la anterior, ya que ambas refuerzan las dos desventajas que tiene el SVM, las cuales son: Alto consumo de recursos para su ejecución y su rendimiento medio al ingresar una cantidad elevada de datos.

4.3.2 Aplicación de la técnica USE SVM + PCA al problema

El problema planteado se enfrentó con un enfoque de clasificación binaria [51]. Este tipo de procedimiento se basa en la determinación de una función clasificadora que permite asignar a cada objeto a una de las dos clases definidas a priori. En nuestro caso, cada cliente será asignado a una de las clases "deserción" o "no deserción".

La construcción del modelo se lleva a cabo en dos etapas: entrenamiento y test. Para cada una de las etapas se considera un subconjunto del total de los objetos (clientes) a clasificar. Estos subconjuntos de objetos forman una partición del conjunto total de objetos y son llamados, conjunto de entrenamiento y conjunto de test, respectivamente.

En la etapa de entrenamiento se estima la mejor función clasificadora considerando algún criterio (por ejemplo, el error de clasificación) en el conjunto de entrenamiento. En la etapa de test se valida la efectividad del modelo respecto de objetos no utilizados en el entrenamiento. Para esto, se utiliza el modelo obtenido para clasificar los elementos del conjunto de test. El modelo asigna cada objeto a una de las clases definidas, la que llamaremos la "clase generada" del objeto en contraposición a la "clase real" que es la clase a la que el objeto efectivamente pertenece. Considerando los objetos "mal clasificados" (aquellos cuya clase generada es diferente a su clase real) se estima un error de clasificación. Dependiendo de este error, se revisa el modelo propuesto.

Existen diversas técnicas que tratan el problema de clasificación binaria. Entre estas podemos mencionar: las redes neuronales artificiales, los arboles de decisión y los Support Vector Machines (SVM) [51]. Para este trabajo, motivados por la efectividad y robustez en

problemas de clasificación reportada en la literatura sobre este tipo de métodos (ver por ejemplo, [48, 50, 51]) seleccionamos USE SVM + PCA.

A continuación se exponen los pasos para desarrollar el modelo propuesto USE SVM + PCA. Estos pasos a seguir han sido extraídos del Capítulo 3, en el cual se ha presentado el modelo en forma general.

1. Escoger el óptimo número 'M'.

   El conjunto de datos de entrenamiento es dividido en 'M' partes iguales, formándose 'M' grupos. Siendo 'M' elegido por un muestreador al azar.

2. PCA para la reducción de la dimensión de los atributos.

   Se tiene en cuenta que el análisis de componentes principales (PCA) es un procedimiento matemático para transformar un conjunto de predictores correlacionados en un conjunto de nuevas variables no correlacionadas llamada componentes principales (PC) que capturan la máxima cantidad de variación en los datos.

   La aplicación del PCA se aplica al dataset de cada grupo generado por 'M'. Esta técnica nos ayudará a reducir la dimensión de atributos, ya que actualmente tenemos 33 atributos, los cuales fueron especificados en el Capítulo 4.

3. Aplicar SVM a los grupos con la data de entrenamiento.

   Una vez que se ha reducido la cantidad de atributos de la data, se aplica SVM a cada grupo de 'M'.

   Teniendo en cuenta el conjunto de datos:

   $$s = \Big((x_1, y_,), (x_2, y_2), \dots, (x_n, y_n)\Big) \in R^N x R$$

   Donde $x_i$es el vector de entrada de 33 dimensiones, y_i es la salida vector que representa la clasificación a la cual pertenece cada cliente considerando que si el valor de y = 1 denota clientes propensos a dejar de utilizar el servicio de tarjetas de crédito en la entidad financiera, caso contrario, si el valor de y=-1, denota clientes fieles al uso de tarjetas de crédito en la entidad financiera.

Dependiendo, si el conjunto S es no linealmente separable, se usará la función Kernel. La función Kernel que se utilizará por ser estándar para este tipo de problema es la del RBF (Radial Basis Function).

$$RBFKernel: k(x, y) = exp(-\frac{\|x_i - x_j\|}{2\sigma^2})$$

Donde:

$$\sigma = \mathbb{R}_+$$

El objetivo es el de entrenar a cada clasificador SVM con la data de entrenamiento, la cual ha sido proporcionada por el usuario. Cada clasificador SVM se entrena, construyendo el modelo con todas las instancias (registros) de entrada y obtener la siguiente función:

$$f(x) = w^T \bullet \phi(x) + b$$

Donde φ (x) es el mapeo no lineal a una alta característica dimensional espacio, y w es el vector de peso, b es un desplazamiento.

En este paso se hace uso de un conjunto de modelos SVM, a lo que se le denomina "Uso combinado de varios modelos", la cual es una técnica muy usada en algoritmos de aprendizaje supervisado. El uso combinado de varios modelos no necesariamente se tiene que utilizar el mismo clasificador. En nuestro caso, hemos utilizado el mismo clasificador por el alto rendimiento del SVM [27]. A continuación en la Figura 4.3mostramos la arquitectura del uso combinado de clasificadores SVM.

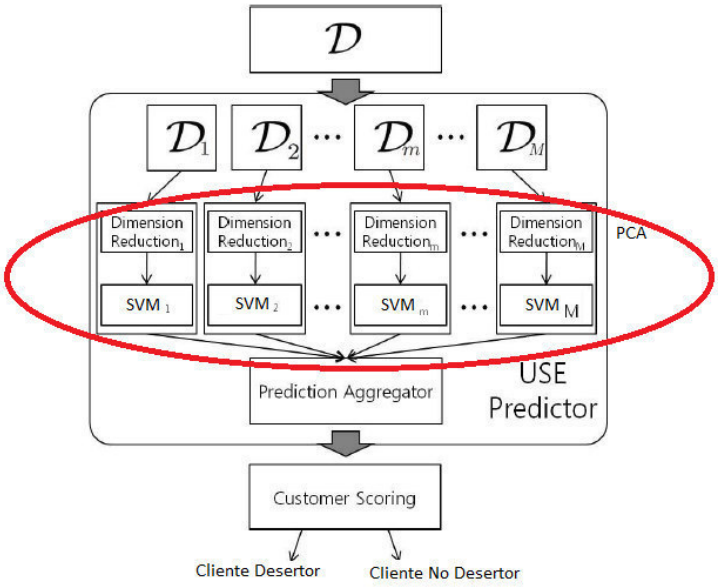


Figura 4.3 Arquitectura del modelo propuesto USE SVM + PCA, donde se utiliza "Uso combinado de modelos SVM"

4. Validar información del test.

Mediante el método de votación por pesos, el cual es un método de pesos y sirve para generar una decisión colectiva entre varios clasificadores, se valida la información del test. Se utilizan todos los modelos entrenados para que validen la información del test por cada instancia. Luego el resultado de la validación (el cual es la clase predicha), se multiplica por el peso de cada clasificador, el cual viene a ser su porcentaje de precisión. El porcentaje de precisión se hallará de la siguiente manera:

$$Presición = \frac{DC + NDC}{DC + NDC + DI + NDI}$$

Donde:

DC : Denota clientes desertores correctamente clasificados.
NDC : Denota clientes no desertores correctamente clasificados.
DI : Denota clientes desertores incorrectamente clasificados.
NDI : Denota clientes no desertores incorrectamente clasificados.

Una vez que tenemos los resultados de cada clasificador, nos disponemos a sumar la cantidad para cada clasificación predicha por cada instancia, y la clasificación con mayor cantidad de votos es la predicción resultante.

Capítulo 5: Ingeniería del Artefacto

En esta sección se presentan los aspectos funcionales de la aplicación y se plasma mediante diagramas, modelos y artefactos del RUP, las consideraciones técnicas y tecnológicas (plataforma) requeridas para el diseño e implementación del sistema de predicción de clientes desertores de tarjetas de crédito, software al cual le hemos puesto el nombre de SISPDTC. El software permitirá la predicción de los clientes desertores, aquellos que están a punto de dejar de usar el servicio de tarjetas de crédito para que se planteen estrategias de marketing al respecto y pueda (la entidad financiera) retener al cliente o los clientes en cuestión.

5.1 Captura de requerimientos

Los requisitos del sistema definen lo que tiene que hacer el sistema y las circunstancias bajo las cuales debe operar.

Los casos de uso proporcionarán un medio sistemático para la captura de requisitos funcionales, y proveerán las bases para la definición de las interfaces de usuario del sistema.

Artefactos creados en este proceso son los que se nombran a continuación:

- *Especificación de Roles*

  Describe los roles del sistema especificando las funcionalidades a las cuales tendrán acceso. A continuación en el Cuadro 5.1 se muestra al detalle la especificación de cada rol identificado para el sistema.

ROL	DESCRIPCIÓN
Analista de Sistema	Se encargará de la preparación y entrenamiento de la data.
Analista de Fidelización	Generará los reportes de identificación de los posibles clientes desertores.
Administrador	Podrá hacer las modificaciones de los parámetros que usa el modelo usado.

Cuadro 5.1 Descripción de roles del sistema

- *Diagrama de caso de uso.*

  Captura los requisitos funcionales de los usuarios y establece la estructura fundamental del sistema. Es el punto de partida para las actividades en análisis, diseño y pruebas. En la Figura 5.1 se presenta el diagrama de caso de uso del sistema SISPDTC.

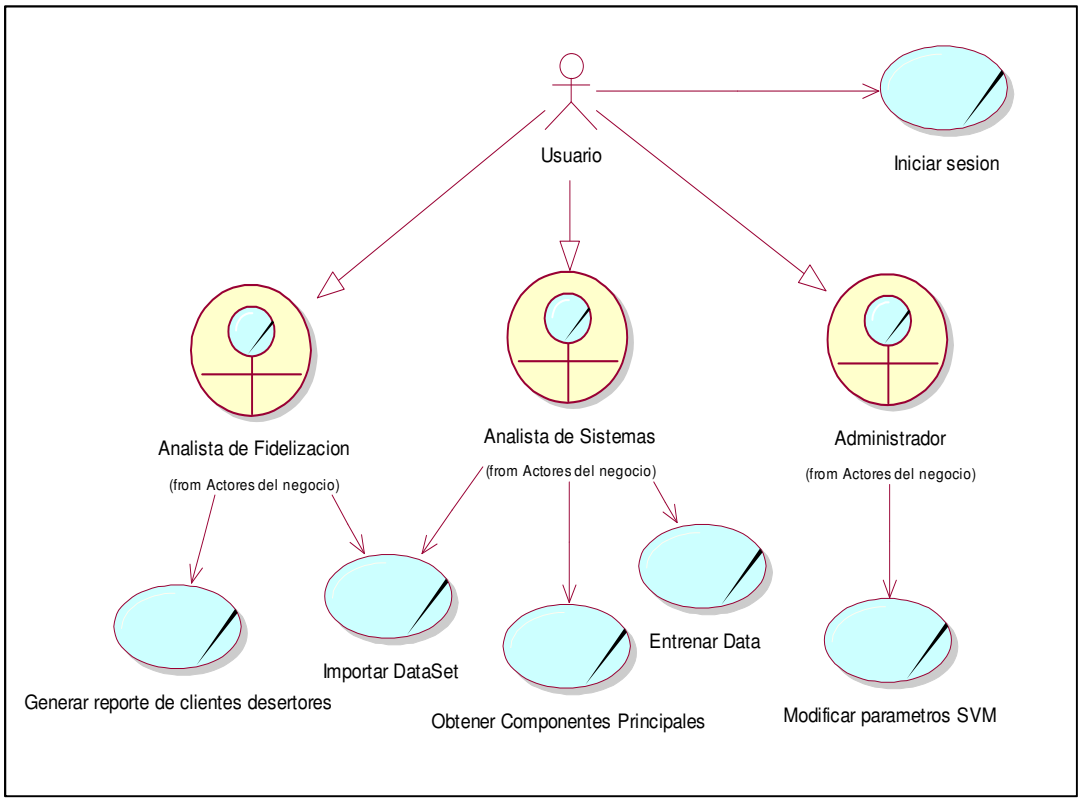


Figura 5.1 Diagrama de Casos de Uso de SISPDTC

A continuación se presenta la descripción de los casos de uso por módulos de diseño arquitectónicamente significativos:

- Módulo de seguridad

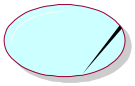
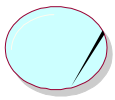
  El presente módulo verifica la autentificación del usuario. Le brinda seguridad al sistema para que el usuario con el rol adecuado pueda hacer las actividades correspondientes a su rol. En el Cuadro 5.2 se muestra el caso de uso que corresponde al presente módulo.

CASOS DE USO	DESCRIPCIÓN
	Iniciar sesión en el sistema a partir de un usuario y contraseña. Cada usuario correspondiente a uno o más roles.

Cuadro 5.2 Descripción de casos de uso del módulo de seguridad

- Módulo de limpieza

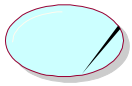
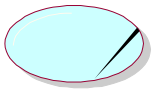
  El presente módulo abarca tanto la carga de archivos como la limpieza del mismo. La limpieza del archivo abarca técnicas planteadas en el Capítulo 3. En el Cuadro 5.3, se muestran los casos de uso que corresponden al presente módulo.

CASOS DE USO	**DESCRIPCIÓN**
	Importar DataSet. Permite la subida de un archivo csv o arff para su posterior procesamiento.
	Obtener Componentes Principales. Aplica el análisis de componentes principales para la normalización y reducción de las variables.

Cuadro 5.3 Descripción de casos de uso del módulo de limpieza

- Módulo de entrenamiento

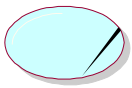
  En el presente módulo se hace referencia a los casos de uso que intervienen en la fase de entrenamiento del clasificador SVM. En el Cuadro 5.4, se muestran los casos de uso que corresponden al presente módulo.

CASOS DE USO	DESCRIPCIÓN
	Modificar Parámetros SVM. Permite la modificación de los parámetros que intervienen en la fórmula del clasificador.
	Entrenar Data. Permite el entrenamiento de la data anteriormente importada.

Cuadro 5.4 Descripción de casos de uso del módulo de entrenamiento

- Módulo de validación

  El presente módulo hace referencia a los casos de uso que intervienen en la fase de validación. En el Cuadro 5.5, se muestran los casos de uso que corresponden al presente módulo.

CASOS DE USO	DESCRIPCIÓN
	Generar reportes de clientes desertores. Permite la generación del reporte de clientes desertores.

Cuadro 5.5 Descripción de casos de uso del módulo de validación

- *Vistas Arquitectónicas*

  Este tipo de vista se presenta la arquitectura del software, donde destacamos las siguientes vistas: la vista lógica y la vista de casos de uso. Todas estas vistas fueron representadas a través de UML con la herramienta Rational Rose.

Las vistas que vamos a presentar en este apartado nos dan un panorama de la arquitectura del sistema desde varios enfoques como puede ser del negocio, de la aplicación, de la información y de la tecnología.

A continuación presentamos las vistas en cuestión:

- Vista de Casos de Uso

  Los casos de uso aquí mostrados están a un alto nivel de abstracción obviando los detalles de los mismos, son auto-explicativos y permiten apreciar el panorama general de la funcionalidad del proyecto. Los diagramas de casos de uso serán mostrados mediante paquetes. A continuación se presentan los paquetes identificados para nuestro sistema.

  a. Paquete de Seguridad

     En este paquete se muestra el diagrama de casos de uso relacionados a la seguridad del sistema, esto lo podemos ver en la Figura 5.2.

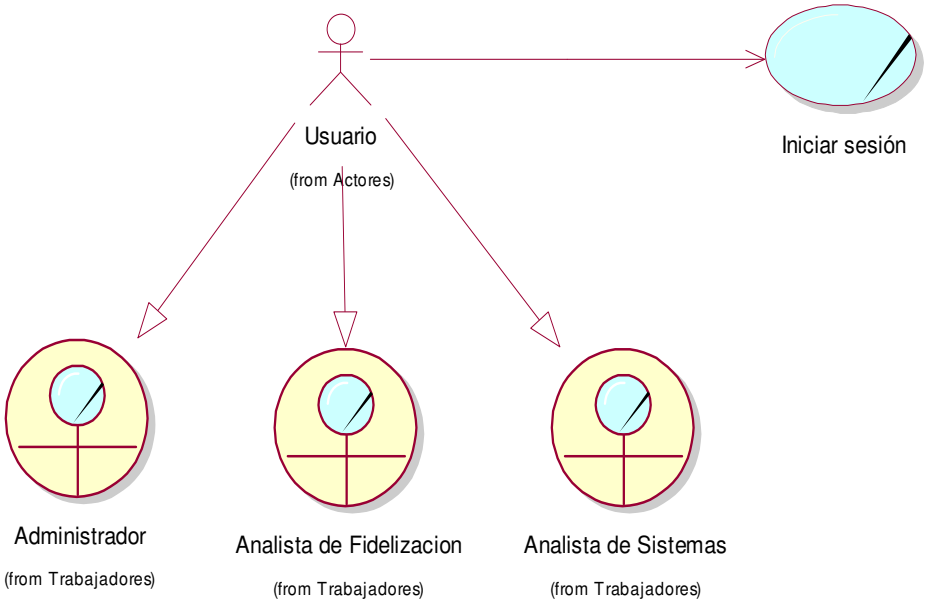


Figura 5.2 Diagrama de Casos de Uso del paquete de Seguridad

  b. Paquete de Administración de variables

     En este paquete se muestra el diagrama de casos de uso relacionados con la administración de variables, esto involucra editar los parámetros de la función kernel RBF con la cual trabaja el sistema actualmente. Este diagrama se presenta en la Figura 5.3.

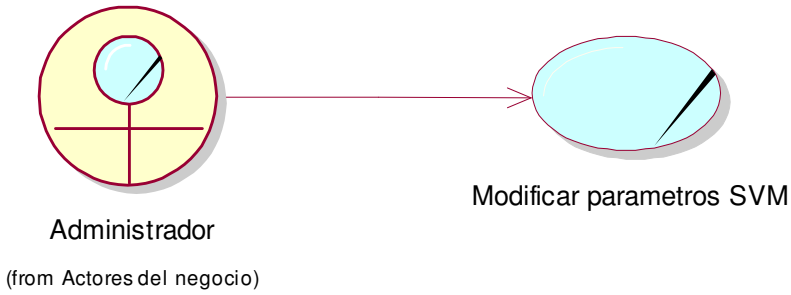


Figura 5.3 Diagrama de Casos de Uso del paquete de Administración de variables

c. Paquete de Entrenamiento de data

En este paquete se presentan los casos de uso que abarca el proceso de entrenamiento de la data (ver la Figura 5.4).

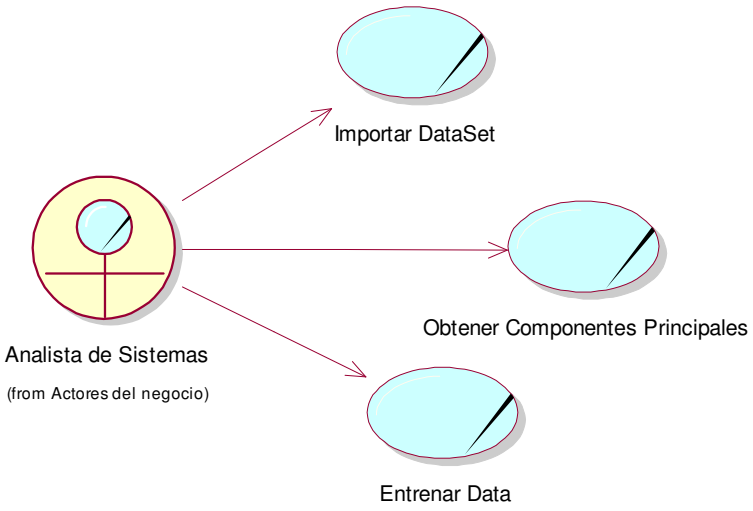


Figura 5.4 Diagrama de Casos de Uso del paquete de Entrenamiento de data

d. Paquete de Validación de data

En este paquete se muestra los casos de uso que abarcan el proceso de validación de data (ver Figura 5.5).

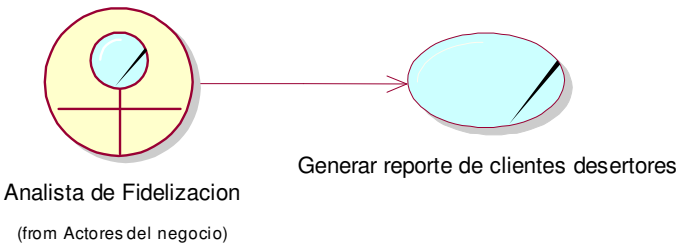


Figura 5.5 Diagrama de Casos de Uso del paquete de Validación de data

- Vista Lógica

  En este tipo de vista se muestra el diagrama de paquetes en la Figura 5.6, en donde se han agrupado los casos de uso de manera cohesiva.

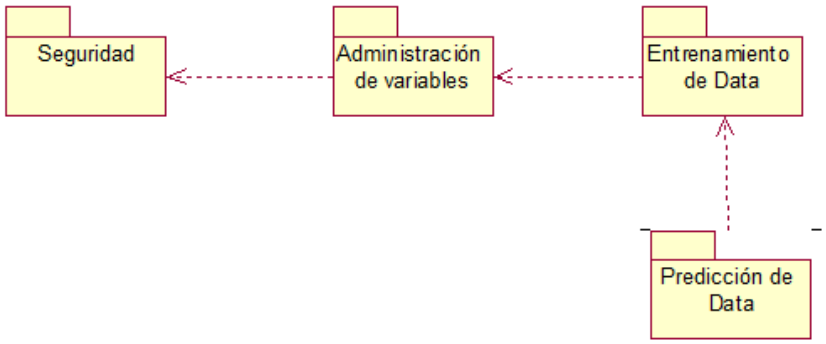


Figura 5.6 Diagrama de Paquetes

5.2 Análisis y diseño

El propósito de esta disciplina es transformar los requerimientos y casos de uso en artefactos que especifiquen el diseño de lo que serán los módulos.

La finalidad de la presente tesis es el desarrollo de los módulos indicados en la sección anterior implementados con la técnica del algoritmo de aprendizaje supervisado Support Vector Machine (SVM), por lo tanto, el interés de las siguientes secciones está enfocado en el análisis y diseño de los artefactos asociados a estos módulos.

Los artefactos creados en este proceso son los siguientes:

- Especificaciones de Casos de Uso.
  Se detallan en el **Anexo A**.
- Diagramas de Actividades

Se detallan en el **Anexo B.**

- Diagrama de Clases.

  El propósito de este artefacto es el de representar los objetos fundamentales del sistema. La Figura 5.7 muestra las clases requeridas para el desarrollo del modelo propuesto USE SVM + PCA.

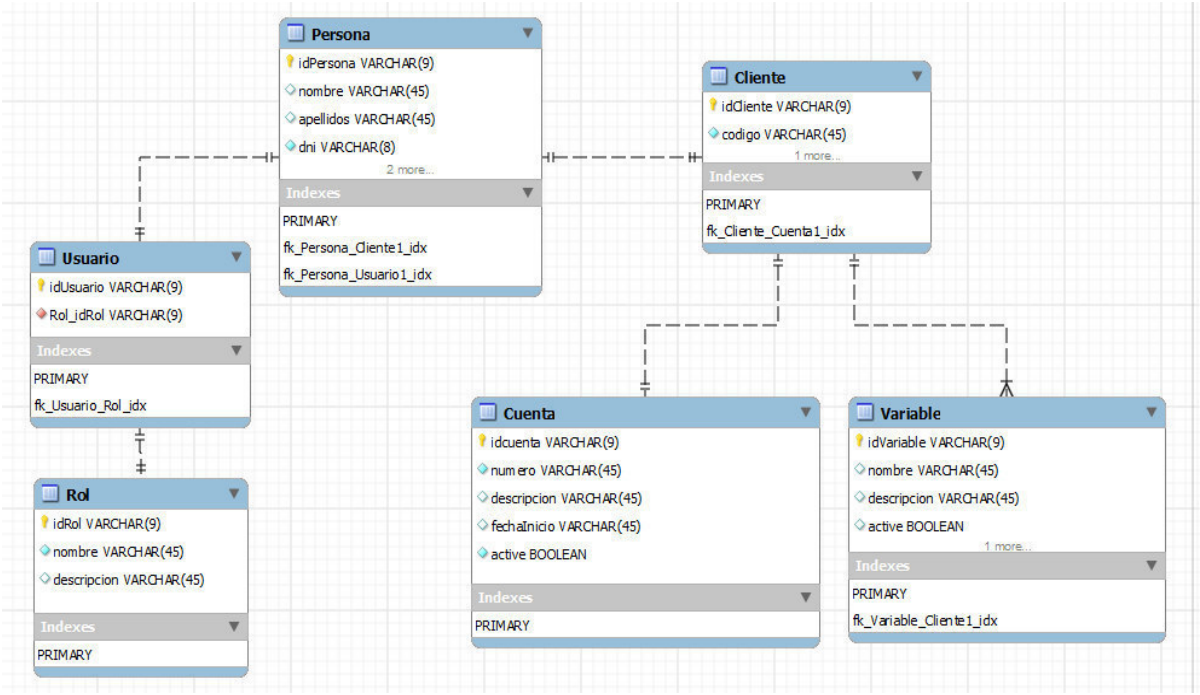


Figura 5.7 Diagrama de clases del sistema SISPDTC

5.3 Implementación

El código fuente de los módulos fue escrito en Java con la versión de JRE 1.7. La arquitectura propuesta es Web, usando el patrón MVC. La herramienta de desarrollo seleccionada es Eclipse Kepler para desarrolladores. Por otro lado, para el proceso de modelamiento se utilizó la herramienta mysql-workbench 5.2.44.

Para el desarrollo del sistema, se utilizó una computadora Intel Core 2 Duo CPU 2.26 GHz, 2 gb DDR2 RAM, 160 GB disponibles de disco duro. Esto es debido a que el uso del SVM requiere un alto rendimiento de la máquina donde se está ejecutando porque gasta muchos recursos en su ejecución.

Arquitectura y Plataforma Tecnológica

A continuación se muestra en la Figura 5.8 la arquitectura del sistema SISPDTC, el cual está dividido en 'n' capas, las cuales son las siguientes:

- Capa de presentación

  Conformada por interfaces de usuario hechas en XHTML.
- Capa de aplicación

  Conformada por aplicaciones Java que se ejecutan en un servidor Apache Tomcat 7.0.
- Capa de datos

  Conformada por tablas de base de datos residentes en el servidor de MySQL 5.5.24.

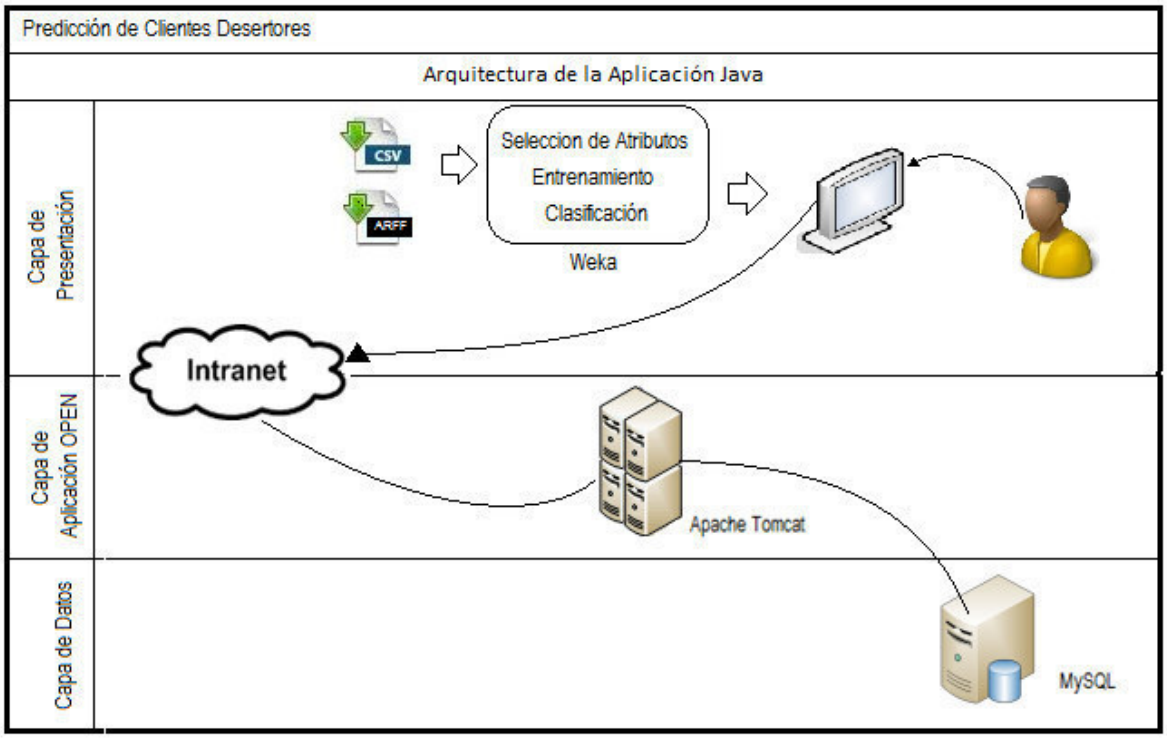


Figura 5.8 Arquitectura del Sistema SISPDTC

5.4 Características del Hardware y Software

A continuación se presentan las características del hardware y software que se han tenido en cuenta para el desarrollo de este sistema de predicción de deserción de clientes de tarjetas de crédito (SISPDTC).

Hardware

HARDWARE	CARACTERÍSTICAS
Servidor de aplicaciones	4 procesador up de 3.0Ghz 10GB RAM, 2x146GB Disco Duro 2 tarjetas de red de 100/1000 Arquitectura Blade y con VMWARE.
Servidor de Base de Datos	4 procesadores up de 3.0 Ghz 10GB RAM, 2x146GB Disco Duro 2 tarjetas de red de 100/1000 Arquitectura Blade y con VMWARE
Estaciones de trabajo	Intel Core 2 Duo, 2 gb DDR2 RAM, 160 GB

Software

SISTEMAS OPERATIVOS
1. Servidores • Sistema Operativo Linux. Red Hat 4 2. Estaciones de Trabajo • Windows XP, Windows 2003
SERVIDORES DE APLICACIONES y WEB
• Apache Tomcat 7.0
SERVIDORES DE BASE DE DATOS
• MySQL 5.5.24

5.5 Planteamiento del modelo de predicción propuesto

En la presente sección nos enfocaremos en el diseño y desarrollo del modelo de predicción. Se realiza una descripción detallada de los pasos necesarios para su implementación en un producto software.

El objetivo primordial es el porcentaje de predicción del comportamiento de compra de un tarjetahabiente como desertor o no de un servicio de la entidad.

Se presenta un esquema general de los pasos a seguir para la aplicación del modelo:

Paso 1. Escoger el número aleatorio 'M'.

Paso 2. Aplicación del Análisis de Componentes Principales (PCA).

Paso 3. Entrenar a los grupos divididos en 'M' partes iguales con SVM.

Paso 4. Validar la información del test (archivo que contiene la información de los clientes que se quiere predecir si son posibles desertores).

A continuación se describe de manera detallada cada paso por el que pasa nuestro modelo, esto acompañado con un extracto de código del sistema que realiza dicho paso.

Paso 1. Escoger el número aleatorio 'M'

El número M será elegido de manera aleatoria. Este número debe estar en el rango entre 1 y el número total de instancias (número total de registros de clientes), y debe ser divisor entre el número total de instancias. El número 'M' servirá para dividir las instancias en 'M' partes iguales.

A continuación se muestra una parte del código donde veremos de qué manera se obtiene el 'M', mediante un método recursivo:

```
public int getRandomDivisorNumber(final int number)
{
int random = (int) Math.random();
while (random == 0){
random = (int) Math.random();
}
if (number / random == 0) {
```

```
return random;
} else {
return getRandomDivisorNumber(number);
      }
   }
```

Paso 2. Aplicación del Análisis de Componentes Principales (PCA)

El análisis de componentes principales (PCA) es una técnica utilizada para reducir la dimensionalidad de un conjunto de datos, una codificación adecuada puede mejorar la tasa de acierto de predicción.

Para la codificación se hizo uso de los métodos implementados en las librerías brindadas por la herramienta Weka 3.6.0:

- libsvm.jar (nos permite utilizar el SVM en un entorno Java)
- wlsvm.jar (complemento de la herramienta weka y libsvm)
- weka.jar (nos permite utilizar la herramienta Weka en un entorno Java)

A continuación mostramos un extracto del código donde veremos de cómo realizar el análisis de componentes principales.

```
BufferedReader reader;
ArffReader arff;
Instances data = null;

//Se obtiene el archivo en formato arff.
reader = new BufferedReader(new FileReader(archivo));
arff = new ArffReader(reader, 1000);
data = arff.getStructure();

// Luego indicamos el índice en el cual se encuentra la clasificación
```

```
data.setClassIndex(data.numAttributes() - 1);
Instance inst;
//Aquí se obtienen las instancias del archivo importado
while ((inst = arff.readInstance(data)) != null) {
                    data.add(inst);
}
//Se invoca a la función PrincipalComponents de la librería de weka.
PrincipalComponents pca = new PrincipalComponents();
//Se ingresan los parametrso para el procesamiento
pca.setCenterData(false);
pca.setMaximumAttributeNames(5);
pca.setTransformBackToOriginal(false);
pca.setVarianceCovered(0.95);
//Se instancia un generedor de ranking de atributos
Ranker ranker = new Ranker();
ranker.setGenerateRanking(true);
ranker.setNumToSelect(-1);
//Se instancia el flitro que unirá el pca con el rankeador.
weka.filters.supervised.attribute.AttributeSelection filter = new
weka.filters.supervised.attribute.AttributeSelection();
filter.setEvaluator(pca);
filter.setSearch(ranker);
//Se genera un Nuevo archivo
filter.setInputFormat(data);
//Se guarda el Nuevo archivo
saveArff(newData, destination+pcaFile)
```

Paso 3. Entrenar a los grupos divididos en 'M' partes iguales con SVM

En este paso nos disponemos a entrenar a los clasificadores SVM que hay en cada grupo de ''M'. Para ello se utilizará la función kernel de tipo RBF con los parámetros que maneja por defecto.

A continuación mostramos un extracto del código donde veremos de qué manera se entrena a los clasificadores SVM (de la librería libsvm) haciendo uso del kernel RBF.

Primero obtenemos el archivo en formato csv o arff, los cuales son los permitidos en weka. Luego, obtenemos las instancias del archivo subido al sistema. Después de ello, obtenemos el número de instancias de la estructura creada para poder hallar 'M' con el método "getRandomDivisorNumber(<número de instancias>)". Una vez obtenido el 'M', calculamos cuántas instancias del total tendrá cada parte de 'M'. Luego, instanciamos la función LibSVM que queremos utilizar para el entrenamiento de nuestro modelo. Finalmente, mediante procesos iterativos, entrenamos cada grupo de 'M' con el método "buildClassiffier(<Instancias>)". De esta manera, entrenamos a los clasificadores SVM con las instancias de cada grupo de 'M'.

```
finalint M = getRandomDivisorNumber(numberInstances);
System.out.println("M =>"+M);
svm = new LibSVM[M];
precisiones = newdouble[M];
finalint cant4M = numberInstances/M;
int inicio = 0;
int intervalo = cant4M;

double accuaryTotal = 0;

double w = Double.parseDouble(1+"")/Double.parseDouble(M+"");

//Mezclala data antesdedividirla
int semilla = 1;
Random rdm= newRandom(semilla);
dataset.randomize(rdm);

for (int i=0;i<M;i++){
	svm[i] = new LibSVM();
	svm[i].setKernelType(new SelectedTag(LibSVM.KERNELTYPE_RBF,
	LibSVM.TAGS_KERNELTYPE));

	train = new Instances(arff.getStructure());
	for (int j=inicio;j<intervalo;j++){
		train.add(dataset.instance(j));
	}
	train.setClassIndex(train.numAttributes()-1);

	int countDesertores=0;
	for (int r = 0; r <train.numInstances(); r++) {
		try {
```

```
            String estado = train.classAttribute().value((int)
            train.instance(r).classValue());

      if(estado.equals("DESERTOR")){
                  countDesertores++;
            }
      } catch (Exception e) {
      // TODO Auto-generated catch block
            e.printStackTrace();
      }
}

try {
      svm[i].buildClassifier(train);
      Evaluation eval = newEvaluation(train);
      eval.evaluateModel(svm[i], train);

      double accuracy = 1-eval.errorRate();
      precisiones[i] = accuracy;
      accuaryTotal = accuaryTotal + w*accuracy;
} catch (Exception e) {
// TODO Auto-generated catch block
      e.printStackTrace();
}
inicio = intervalo;
intervalo = intervalo + cant4M;
}
```

Paso 4. Validar la información del test

Una vez que se ha entrenado a cada clasificador SVM, ahora en este paso mediante un método de pesos por votación se valida la información brindada por el test, archivo que contiene información de los clientes que el usuario desea saber quiénes son identificados por el sistema como posibles desertores. Está técnica de pesos por votación consiste en evaluar cada instancia (registro del archivo test) mediante la votación de cada clasificador SVM que ha sido entrenado en el paso anterior. Cada clasificador tiene un peso (de acuerdo a su porcentaje de acierto en la etapa de entrenamiento) y tiene un resultado de predicción, cada peso de cada clasificador es multiplicado por su resultado (el cual es 'y', resultado de la predicción que puede ser 1 o -1, y estos números representan si son desertores o no). Luego de hacer esta operación, se hace una suma para cada clasificación y la clasificación que tenga un mayor número es el resultado de la predicción para cada instancia.

Con el siguiente extracto de código se colocan todos los clasificadores que fueron entrenados en una instancia de la clase Vote que brinda la herramienta Weka.

```
vote = new Vote();
vote.setClassifiers(svm);
```

```
SelectedTag t1 = newSelectedTag(Vote.AVERAGE_RULE,Vote.TAGS_RULES);
vote.setCombinationRule(t1);
```

Y como podemos ver en el siguiente extracto de código, se da el método de pesos por votación.

```
BufferedReader reader;
ArffReader arff;

reader = new BufferedReader(new FileReader(destination+ nombreArchivo +
".arff"));
arff = new ArffReader(reader, 1000);
atributos = arff.getStructure().toString();
dataset = new Instances(arff.getStructure());
nombreArff = nombreArchivo;

dataset.setClassIndex(dataset.numAttributes() - 1);
Instance inst;

while ((inst = arff.readInstance(dataset)) != null) {
      dataset.add(inst);
}

for (int i = 0; i <dataset.numInstances(); i++) {
      System.out.println(vote.classifyInstance(dataset.instance(i)));
}
```

5.6 Interfaces del sistema

Las interfaces del sistema se presentarán por módulos, los que antes se mencionaron en el presente capítulo.

- Módulo de seguridad

  Como vemos en la Figura 5.9, se muestra la pantalla inicial del sistema donde el usuario debe identificarse con su respectivo usuario y contraseña para poder ingresar al sistema a hacer las actividades que tiene a cargo dependiendo del rol del usuario asignado.

Figura 5.9 Interfaz de la pantalla principal del sistema SISPDTC

Una vez que el usuario ha ingresado sus datos correctamente, le aparece una pantalla de bienvenida al sistema como se muestra en la Figura 5.10.

Figura 5.10 Interfaz de la pantalla de Bienvenida al sistema

- Módulo de limpieza

  En el presente módulo se muestra todas las interfaces pertenecientes a este módulo. Este módulo abarca el proceso de limpieza de data donde se aplica la técnica PCA para la reducción de atributos.

Como podemos ver en la Figura 5.11, nos muestra una pantalla en la cual nos muestra las opciones que tenemos en ese módulo.

Figura 5.11 Interfaz de la pantalla inicial del módulo de limpieza

Como podemos ver en la Figura 5.12, nos muestra una pantalla donde nos permite elegir un archivo de tipo csv o arff para hacerle su respectiva limpieza aplicando la técnica PCA, que es la propuesta en este trabajo de investigación.

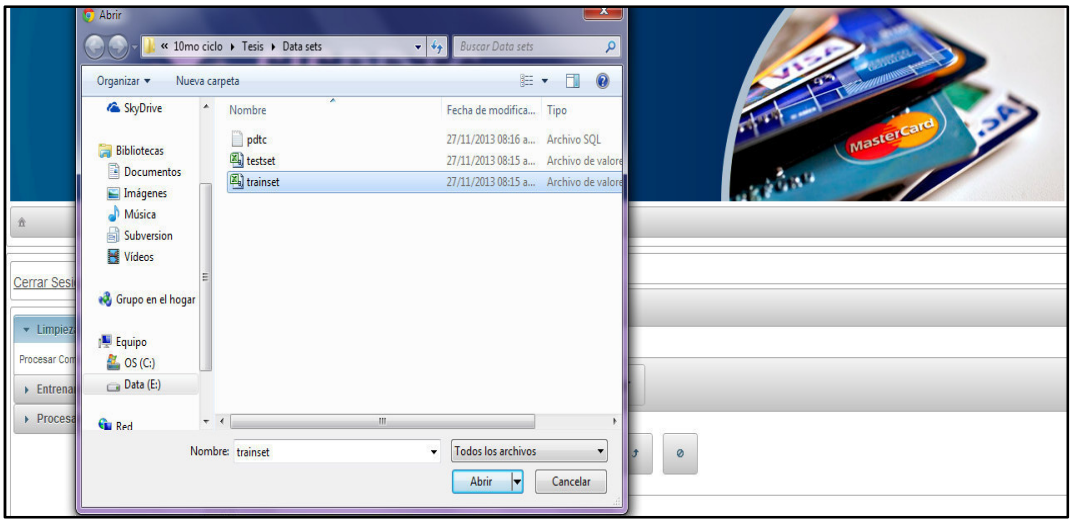


Figura 5.12 Interfaz para cargar un archivo en el módulo de limpieza

Como podemos ver en la Figura 5.13, se muestra una pantalla en la que el archivo ya está seleccionado y ahora debe ser subido al sistema.

Figura 5.13 Interfaz para subir un archivo en el módulo de limpieza

Como podemos ver en la Figura 5.14, se muestra toda la información del archivo subido al sistema. Tanto los atributos como la información del dataset.

Figura 5.14 Interfaz donde se muestra el contenido del archivo antes de la limpieza

Como podemos ver en la Figura 5.15, se muestra la información procesada por el algoritmo PCA.

Figura 5.15 Interfaz del archivo procesado después de aplicar PCA en el módulo de limpieza

- Módulo de entrenamiento

  En este módulo se mostrarán las interfaces que abarcan el proceso de entrenamiento del clasificador SVM.

  Como podemos ver en la Figura 5.16, se muestra la pantalla principal del módulo de entrenamiento donde se muestran todas las opciones que presenta el presente módulo.

Figura 5.16 Interfaz de la pantalla principal del módulo de entrenamiento

Como podemos ver en la Figura 5.17, nos muestra una pantalla donde nos permite elegir un archivo de extensión csv o arff, el que luego servirá para entrenar al clasificador SVM.

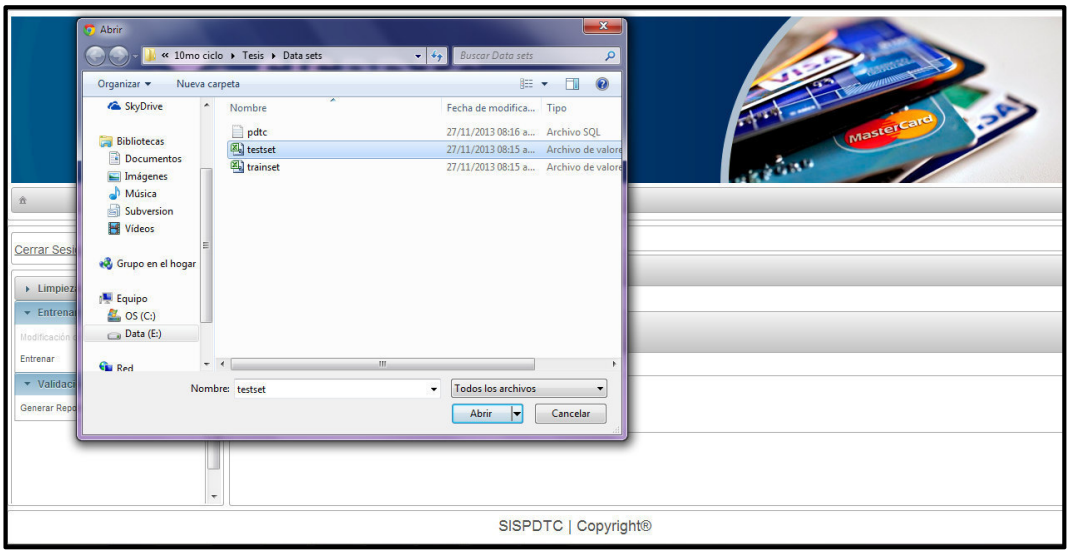


Figura 5.17 Interfaz para cargar un archivo en el módulo de entrenamiento

Como podemos ver en la Figura 5.18, nos muestra que el archivo ya ha sido subido al sistema y ahora está preparado para entrenar al clasificador SVM.

Figura 5.18 Interfaz donde el archivo ya está subido al sistema en el módulo de entrenamiento

Como podemos ver en la Figura 5.19, nos muestra la pantalla luego del entrenamiento, en la que se muestra la tasa de acierto que ha tenido el clasificador SVM al ser entrenado.

Figura 5.19 Interfaz donde se muestra la tasa de acierto en la fase de entrenamiento

- Módulo de validación

En este módulo se muestra todas las acciones que se pueden realizar en la fase de validación, en la que se predice si un cliente es "Desertor" o "No Desertor".

Como podemos ver en la Figura 5.20, nos muestra la pantalla principal del módulo de validación y se muestran las opciones que tenemos en este módulo.

Figura 5.20 Interfaz de la pantalla principal del módulo de validación

Como podemos ver en la Figura 5.21, nos muestra una pantalla en la que ha sido seleccionado un archivo y ha sido subido al sistema. Este archivo es el que contiene la información de los clientes de los cuales se quiere predecir quiénes son identificados como “Desertores”.

Figura 5.21 Interfaz donde el archivo ya ha sido subido al sistema en el módulo de validación

Como podemos ver en la Figura 5.22, nos muestra el listado de clientes que han sido clasificados como posibles “Desertores”. También nos muestra la tasa de acierto que ha tenido el modelo al predecir si un cliente es identificado como “Desertor”.

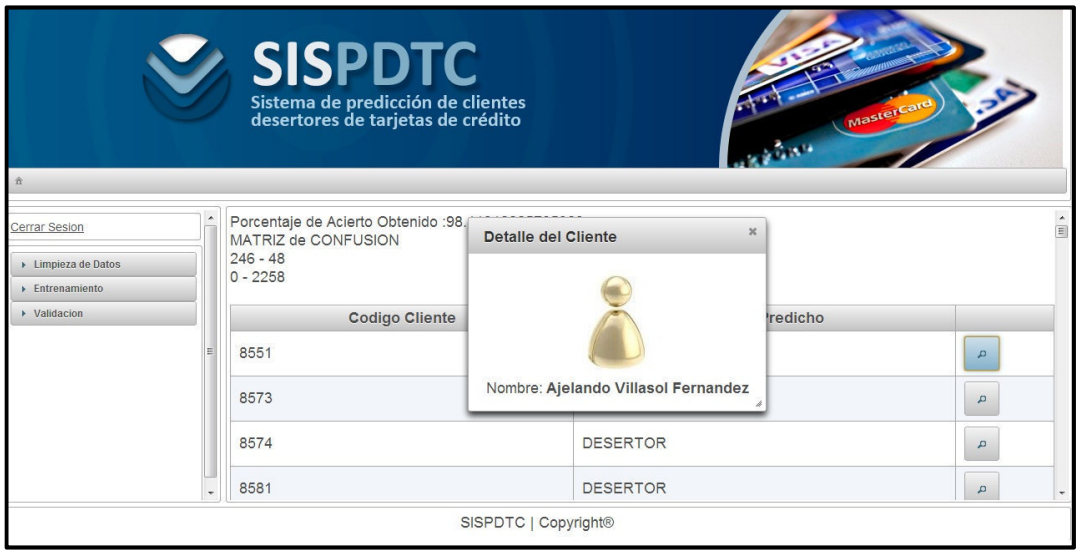


Figura 5.22 Interfaz donde se muestra el resultado de la predicción de clientes desertores

Capítulo 6: EXPERIMENTOS Y RESULTADOS

Este capítulo trata del proceso de entrenamiento del modelo USE SVM, sometido a diferentes escenarios donde se busca el mejor resultado que nos arroje la mayor tasa de predicción. Finalmente, se explica el procedimiento utilizado para la verificación de los porcentajes de acierto obtenidos en la comparación con el modelo SVM.

6.1 Características del Hardware y Software

En este punto se especifica tanto el hardware como software que se han utilizado para realizar las pruebas del sistema.

Hardware

HARDWARE	CARACTERÍSTICAS
Servidor de aplicaciones	4 procesador up de 3.0Ghz 10GB RAM, 2x146GB Disco Duro 2 tarjetas de red de 100/1000 Arquitectura Blade y con VMWARE
Servidor de Base de Datos	4 procesadores up de 3.0 Ghz 10GB RAM, 2x146GB Disco Duro 2 tarjetas de red de 100/1000 Arquitectura Blade y con VMWARE
Estaciones de trabajo	Intel Core 2 Duo, 2 gb DDR2 RAM, 160 GB

Software

SISTEMAS OPERATIVOS
1. Servidores • Sistema Operativo Linux. Red Hat 4 2. Estaciones de Trabajo • Windows XP, Windows 2003
SERVIDORES DE APLICACIONES y WEB
• Apache Tomcat 7.0
SERVIDORES DE BASE DE DATOS
• MySQL 5.5.24

6.2 Obtención de la data

El modelo propuesto USE PCA SVM para la predicción de deserción de clientes de tarjetas de crédito en una entidad financiera ha trabajado con data real de un banco latinoamericano. Esta data se ha obtenido del Business Intelligence Cup 2004, organizado por la Universidad de Chile en 2004.

6.3 Instancias de pruebas

Para este trabajo de investigación se está trabajando con un total de 14694 clientes, siendo 882 clasificados como 'Desertores' y 13812 clasificados como 'No Desertores'. A continuación se detalla, en el Cuadro 6.1, los porcentajes de clientes clasificados como desertores con respecto al total de clientes y lo mismo para los clientes clasificados como no desertores.

	Cantidad	Porcentaje
Desertores	882	6%
No Desertores	13812	94%
Total	14694	100%

Cuadro 6.1 Cuadro de ratios entre desertores, no desertores y el total de ambos

Se ha utilizado el 67% del total de la data para la fase de entrenamiento y el 33% para la fase de validación. Siendo para la fase de entrenamiento un total de 9845 clientes, de los cuales 591 están clasificados como 'Desertores' y 9254 clientes están clasificados como 'No Desertores'. Como podemos observar en el Cuadro 6.2, el ratio entre clientes clasificados como desertores entre el total se mantiene tal cual era en el total inicial, lo mismo con los clientes clasificados como no desertores.

	Cantidad	Porcentaje
Desertores	591	6%
No Desertores	9254	94%
Total	9845	100%

Cuadro 6.2 Cuadro de ratios entre desertores, no desertores y el total de la fase de entrenamiento

Por otro lado, para la fase de validación, se tiene un total de 4849 clientes, de los cuales 291 están clasificados como 'Desertores' y 4558 clientes están clasificados como 'No Desertores'. Como podemos observar en el Cuadro 6.3, el ratio entre clientes clasificados como desertores entre el total se mantiene tal cual era en el total inicial, lo mismo con los clientes clasificados como no desertores.

	Cantidad	Porcentaje
Desertores	291	6%
No Desertores	4558	94%
Total	4849	100%

Cuadro 6.3 Cuadro de ratios entre desertores, no desertores y el total de la fase de validación

Los archivos contenedores de la información de los clientes son de tipo csv, los cuales pueden ser manejados con la herramienta de Microsoft Excel. Cada fila de estos archivos tiene una estructura la cual se muestra en la Figura 6.1. Donde CC representa el código de los clientes, la variable X_i representa los 21 atributos del cliente descrito en el Capítulo 4, y finalmente la variable Y representa la clasificación a la cual pertenece el cliente, el cual es un valor nominal, puede ser 'Desertor' o 'No Desertor'.

CC	X_1	X_2	X_3	X_4	…	X_{20}	X_{21}	Y

Figura 6.1 Estructura de un registro del archivo de entrada

6.4 Pruebas

En este apartado mostraremos los resultados de las pruebas realizadas, tanto en la fase de entrenamiento como en la fase de validación. Se mostrará los porcentajes de acierto de cada fase y su respectiva *matriz de confusión*[7].

6.4.1 Fase de Entrenamiento

Para esta fase obtuvimos un 99.9% de acierto. Obtuvimos el porcentaje de acierto de cada modelo SVM mediante el método de pesos con la siguiente fórmula:

$$\%\ de\ acierto\ de\ cada\ SVM = W \times (1 - E) \times (100\%)$$

Donde:

- W corresponde al peso de cada modelo.

[7] Una matriz de confusión es una herramienta de visualización que se emplea en aprendizaje supervisado.

- E corresponde al ratio de error.

Para obtener el porcentaje total de acierto, se ha sumado cada porcentaje parcial obtenido de cada modelo SVM.

$$\%\ de\ acierto\ Total = \sum W \times (1 - E) \times (100\%)$$

Como podemos observar, ha habido un mínimo porcentaje de equivocación en esta fase de entrenamiento al momento de identificar a los clientes con su respectiva clasificación.

6.4.2 Fase de Validación

Para esta fase obtuvimos un 99.75% de acierto, el cual será mostrado en el Cuadro 6.4 a través de la matriz de confusión. Obtuvimos el porcentaje de acierto mediante la siguiente fórmula:

$$\%\ de\ acierto = \frac{DC + NDC}{DC + NDC + DI + NDI} \times (100\%)$$

Donde:

- DC corresponde al número de clientes desertores clasificados correctamente.
- NDC corresponde al número de clientes no desertores clasificados correctamente.
- DI corresponde al número de clientes desertores clasificados de manera incorrecta.
- NDI corresponde al número de clientes no desertores clasificados de manera incorrecta.

Clasificación	**Desertor**	**No Desertor**
Desertor	286	5
No Desertor	7	4551

Cuadro 6.4 Matriz de confusión de la fase de validación

Con los datos mostrados en la matriz de confusión corroboramos el porcentaje de acierto de esta fase.

$$\%\ de\ acierto = \frac{286 + 4551}{286 + 4551 + 5 + 7} \times (100\%)$$

$$\%\ de\ acierto = \frac{4837}{4849} \times (100\%)$$

$$\%\ de\ acierto = 99.75\%$$

La matriz de confusión mostrada en el Cuadro 6.4 indica que del total de clientes cuya clasificación es la de 'Desertor' (291), 286 han sido clasificados correctamente por el algoritmo de aprendizaje supervisado SVM. También indica que del total de clientes cuya clasificación ha sido la de 'No Desertor', la cual es una cantidad de 4558 clientes, 4551 han sido clasificados correctamente.

Como podemos ver, el resultado es alentador ya que muestra un alto porcentaje de acierto, superando el 90% que nos trazamos en uno de los objetivos de este trabajo de investigación. Este porcentaje de precisión muestra un porcentaje de confiabilidad alto para cualquier entidad financiera que haga uso de este sistema.

Capítulo 7: CONCLUSIONES Y TRABAJOS FUTUROS

En este capítulo, se destacan las conclusiones principales obtenidas en este trabajo de investigación y se resumen los resultados obtenidos del modelo propuesto. Se comentan algunos aspectos relacionados con los trabajos futuros que siguen la propuesta planteada y sobre otros temas de investigación que se pueden derivar.

7.1 Conclusiones

En este estudio se analizó la problemática de la deserción de clientes de tarjetas de crédito en las entidades financieras.

Se realizó la revisión de literaturas para la deserción de clientes en distintos campos, no necesariamente el de tarjetas de crédito, se ven técnicas como: la teoría del Rough Set (RST) y mínimos cuadrados SVM (LS-SVM), sistema de inferencia neuro difuso adaptivo (ANFIS), árboles y redes neuronales, y ninguno de estos muestra un porcentaje de acierto tan alto como el modelo USE PCA SVM.

Se utilizaron perfiles y hábitos de consumo de las personas que poseen una tarjeta de crédito en el Perú que no estaban satisfechas con el servicio brindado por una entidad financiera y, por lo tanto, dejan de utilizar dicho servicio para poder entrenar nuestro modelo.

Se aplicó el modelo USE PCA SVM para predecir a los clientes que piensan dejar de utilizar el servicio de tarjetas de crédito brindado por la entidad financiera.

Se diseñó y desarrolló un software que alcanzó un porcentaje de acierto mayor al 98%, el cual se muestra en el Capítulo 6 Experimentos y Resultados , tal que garantiza su efectividad al ejecutarlo con escenarios reales o simulados de una entidad financiera. La implementación del software desarrollado permitirá el monitoreo de los clientes de una entidad financiera y apliquen estrategias de marketing para los clientes clasificados como Desertores.

Los resultados experimentales de la investigación son alentadores mostrando que el modelo empleado es capaz de alcanzar una buena precisión predicción de clientes desertores.

7.2 Trabajos futuros

En el presente trabajo se ha elegido como técnica el algoritmo SVM y la función kernel RBF, la cual es la más utilizada en estos casos; también se trabaja con variables constantes, las cuales son las que vienen por defecto en la función kernel del SVM. Se propone para un futuro trabajo utilizar optimización de parámetros de la función kernel RBF, para obtener un mejor resultado.

.

Referencias Bibliográficas

[1] Altin y Bajram, “Customers’ Desertion Rate In Microfinance Institutions, Factors And Their Analysis”, Journal of Studies in Economics and Society Vol. 2, No 1, pp. 206-223, 2010.

[2] Ning Wang, Dong-xiao Niu, “Credit Card Customer Churn Prediction Based on the RST and LS-SVM”, 6th International Conference on Service Systems and Service Management pp. 275-279, 2009.

[3] Hee-Su KimyChoong-Han Yoon, “Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market ,”Telecommunications Policy, vol. 28, pp.751-765, 2004.

[4] Hyeonju Seol, Jeewon Choi, and Gwangman Park, “A framework for benchmarking service process using data envelopment analysis and decision tree,” Expert Systems with Applications, vol. 32, pp.432-440, 2007.

[5] Y Bentz, D Merunka, “Neural networks and the multinomial legit for brand choice modeling: a hybrid approach,” Journal of Forecasting, vol. 19, pp.177–200, 2000.

[6] H. Sarimveis and G. Bafas, “Fuzzy model predictive control of non-linear processes using genetic algorithms,” Fuzzy Sets Syst., vol.139, pp.59–80, 2003.

[7] Yang Shulian, “Application of data mining in analysis of customers to leave company in telecom,” Computer and Modernization, vol.2, pp.109-111, 2005.

[8] Neslin, S.A. Gupta, S. Kamakura, W. Lu, J. Mason, C., 2006. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research, 43(2), 204–211.

[9] Coussement, K. F. Benoit, D. Van den Poel, D., 2010. Improved marketing decision making in a customer churn prediction context using generalized additive models, Expert Systems with Applications 37, 2132–2143.

[10] Torkzadeh, G., Chang, J. C.-J., & Hansen, G. W., 2006. Identifying issues in customer relationship management at Merck-Medco. Decision Support Systems, 42(2).

[11] Van den Poel, D., & Larivière, B., 2004. Customer attrition analysis for financial services using proportional hazard models. European Journal of Operational Research,

157(1), 196–217.

[12] Coussement, K. Van den Poel, D., 2008a. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications 34, 313–327.

[13] Xie, Y. Li, X. Ngai, E. Ying, W., 2009. Customer churn prediction using improved balanced random forests, Expert Systems with Applications 36, 5445–5449.

[14] Yu, X., et al. 2010, An extended support vector machine forecasting framework for customer churn in e-commerce.Expert Systems with Application, doi:10.1016/j.eswa.2010.07.049.

[15] [Huang, B. Buckley, B. Kechadi, T., 2010.Multiobjective feature selection by using NSGA-II for customer churnprediction in telecommunications, Expert Systems with Applications 37, 3638–3646.

[16] Tsai, C. Lu, Y., 2009. Customer churn prediction by hybrid neural networks, Expert Sys-tems with Applications, 36, 12547–12553.

[17] Pendharkar, P., 2009,Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services, Expert Systems with Applications 36, 6714–6720.

[18] Lemmens,A., and Croux,C., 2006. "Bagging and boosting classification trees to predict churn," Journal of Marketing Research, vol. 43, no. 2, pp. 276-286, 2006.

[19] Coussement, K. Van den Poel, D., 2008b. Integrating the voice of customers through call center emails into a decision support system for churn prediction- Information & Management, 45, 164–174.

[20] Gary Cokins, Ken King, "Managing Customer Profitability and Economic Value in the Telecommunication Indutry", SAS Institute White paper.

[21] Hangxia Ma, Min Qin, Jianxia Wang. (2009), "Analysis of the Business Customer Churn Based on Decision Tree Method", The Ninth International Conference on Control and Automation, Guangzhou, China.

[22] MO Zan, ZHOA Shan, LI Li, LIU Ai-Jun, 2007, "A predictive Model of Churn in Telecommunications Base on Data Mining"., IEEE International Conference on Control and Automation", Guangzhou, China.

[23] V. Umayaparvathi, K. Iyakutti, March 2012, “Applications of Data Mining Techniques in TelecomChurn Prediction”, International Journal of Computer Applications (0975 – 8887), Volume 42– No.20.

[24] Hossein Abbasimehr, Mostafa Setak y M. J. Tarokh, April 2011, “A Neuro-Fuzzy Classifier for Customer Churn Prediction”, International Journal of Computer Applications (0975 – 8887), Volume 19– No.8.

[25] S. Chaudhuri and U. Dayal.An overview of data warehousing and olap technology. SIGMOD Rec., 26:65–74, March 1997.

[26] L. Wright. The crm imperative practice vs theory in the telecommunications industry. The Journal of Database Marketing, 9:339–349(11), 1 July 2002.

[27] Jaewook Lee Namhyoung Kim, Kyu-Hwan Jung, Yong Seog Kim, 2012, “A New Ensemble Model for Efficient Churn Prediction in Mobile Telecommunication”, 45th Hawaii International Conference on System Sciences.

[28] P. E. Rossi, R. McCulloch, and G. Allenby.The Value of Household Information in Target Marketing.Marketing Science, 15(3):321–340, 1996.

[29] L. Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

[30] Y. Freund and R. Schapire.Experiments with a new Boosting algorithm. In Proc. of 13th Int’l Conf. on Machine Learning, pages 148–156, Bari, Italy, 1996.

[31] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, Boosting, and variants", Machine Learning, 36(1–2):105–139, 1999.

[32] L. Breiman. Stacked regression. Machine Learning, 24(1):49–64, 1996.

[33] H. Kaiser. The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20:141–151.

[34] R. B. Cattell. The scree test for the number of factors. Multivariate Behavioral Research, 1:245–276.

[35] D. Lee and J. Lee. Domain described support vector classifier for multi-class classification problems. Pattern Recognition, 40:41–51, 2007.

[36] D. Lee and J. Lee. Equilibrium-based support vector machine for semisupervised classification. IEEE Trans. On Neural Networks, 18(2):578–583, 2007.

[37] K.-H. Jung, D. Lee, and J. Lee.Fast support-based clustering method for large-scale problems. Pattern Recognition, 43:1975–1983, 2010.

[38] D. Lee and J. Lee. Dynamic dissimilarity measure for supportbased clustering. IEEE Trans. on Knowledge and Data Engineering, 22(6):900–905, 2010.

[39] J. Lee and D. Lee. An improved cluster labeling method for support vector clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(3):461 –464, march 2005.

[40] J. Lee and D. Lee. Dynamic characterization of cluster structures for robust and inductive support vector clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 28(11):1869 –1874, nov. 2006.

[41] Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM (Basado en la Tesis: "Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM)" de José Alberto Gallardo Arancibia).

[42] C. Burges B. Schölkopf and A. Smola. Advances in kernel methods: Support vector machines. Cambridge, MA: MIT Press, 1999.

[43] C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, vol. 2, no. 2, 1998.

[44] V.Ñ. Vapnik. The nature of statistical learning theory. New York: Springer-Verlag, 1995.

[45] C. Cortes and V.Ñ. Vapnik.Support vector networks. Machine Learning, vol. 20, pp 273-297, 1995.

[46] V.Ñ. Vapnik. Statistical learning theory. New York: Wiley, 1998.

[47] N. Cristianini and J. Shawe-Taylor.Support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridbe MA, ISBN 0-

521-78019-5, 2000.

[48] Guangli Nie, Wei Rowe, Lingling Zhang, Yingjie Tian y Yong Shi, Credit card churn forecasting by logisttic regression and decision tree [Publicación]. Science Direct, an international journal, Expert System with Applications volume 38, Issue 12 (November/December 2011).

[49] C. Burges. A tutorial on support vector machines for pattern recognition.Data Mining and Knowledge Discovery, 2(2):121–167, 1998.

[50] N. Cristianini and R. Holloway. Support Vector and Kernel Methods.Springer-Verlag, Berlin, Heidelberg, 2003.

[51] N. Cristianini and J. Shawe-Taylor.An Introduction to Support VectorMachines.Cambridge, UK, Cambridge University, 2000.

[52] A. Athanassopoulos. Customer satisfaction cues to support market segmentation and explain switching behavior. Journal of Business Research, 47(3):197–207, 2000.

[53] C.B. Bhattacharya. When customers are members: Customer retention in paid membership context. Journal of the Academy of Marketing Science, 26(1):31–44, 1998.

[54] J. Ganesh, M.J. Arnold, and K.E. Reynolds.Understanding the customer base of service providers: An examination of the differences between switchers and stayers. Journal of Marketing, 64(3):65–87, 2000.

[55] M. Paulin, J. Perrien, R.J. Ferguson, and A.M.A. Salazar. Relational norms and client retention: external effectiveness of commercial banking in canada and mexico. International Journal of Bank Marketing, 16(1):24–31, 1998.

[56] E. Ramusson. Complaints can build relationships. Sales and MarketingManagement, 151(9):89–90, 1999.

[57] F. Reichheld and E. Sasser. Zero defections: Quality comes to services.

Harvard Business Review, 1990:105–111, September–October, 2000.

[58] D. Van den Poel and Bart Larivi`ere. Customer attrition analysis for financial services using proportional hazard models. Technical Report B–9000, Department of Marketing,Ghent University, Hoveniersberg 24, 2003.

[59] Iván Aquino M., 23 de abril del 2009. Publicación “Tecnologías de Bases de Datos en Entidades Financieras Peruanas” en el Colegio de Matemáticos del Perú.

[60] Gurrala Jagadish, B.Ravi Kiran, H.P.Divya Krishna, C.Pallavi. Publication “Applications of Data Mining Techniques in Customer Churn Prediction - A comprehensive Survey”

[61] Jaime Miranda, Pablo Rey, Richard Weber. Publicación “Predicción de Deserción de Clientes para una Institución Financiera mediante Support Vector Machines”. Revista Ingeniería de Sistemas Volumen XIX, Octubre 2005.

[62] Nekuri Naveen, Vadlamani Ravi, Dudyala Anil Kumar. Journal Article “Application of fuzzyARTMAP for churn prediction in bank credit cards”. Volume 1, Number 4/2009, August 2009.

ANEXO A

SISPDTC – SISTEMA DE PREDICCIÓN DE CLIENTES DESERTORES DE TARJETAS DE CRÉDITO

Especificación de Caso de Uso (ECU)

CU – Entrenar Data

Versión 1.0

1. **Breve Descripción**

   El sistema permitirá entrenar al modelo de minería de datos con un archivo de extensión csv (archivo Excel). El cual tendrá como contenido registros de los clientes en un determinado rango de fecha. Cada registro de un cliente deberá tener un identificador del cliente, sus respectivos valores por cada atributo y finalmente la clasificación a la cual pertenece.

2. **Flujo de Eventos**

   Flujo Básico <<Entrenar Data>>

   1. El caso de uso inicia cuando el usuario se identifica con su código y contraseña respectiva en el sistema y presiona el botón “Entrar”.
   2. El sistema valida los datos ingresados por el usuario. Valida que el usuario tenga asignado el rol correspondiente para realizar esta operación.
   3. El sistema muestra la pantalla principal.
   4. El usuario selecciona la opción “Subir archivo”.
   5. El sistema despliega una ventana donde le permite al usuario buscar el archivo que desea subir.

6. El usuario selecciona el archivo que desea entrenar (con extensión csv) y selecciona la opción “Aceptar”.
7. El sistema verifica que se haya seleccionado un archivo con la extensión csv.
8. El usuario selecciona la opción “Entrenar”.
9. El sistema entrena al modelo de minería de datos con la data cargada.
10. El sistema muestra un mensaje de éxito en el entrenamiento con el tiempo de demora que le ha tomado entrenar la data.
11. El usuario selecciona la opción “Salir”.
12. El sistema muestra la pantalla inicial de login.
13. Finaliza el caso de uso.

Flujos Alternativos << Entrenar Data>>

2.1 Los datos del usuario fueron ingresados incorrectamente. Si uno de los datos ingresados por el usuario es incorrecto, entonces se mostrará en la pantalla un mensaje de error “Datos incorrectos”. El caso de uso finaliza.
7.1 El sistema reconoce que el archivo seleccionado no es csv. Si el sistema reconoce que el archivo seleccionado no es csv (el único permitido por el sistema), entonces se mostrará un mensaje de error “Extensión inválida”. 7.2 El sistema despliega una ventana de archivos, donde le permite al usuario volver a seleccionar un archivo pero esta vez de extensión csv. Continua la secuencia.

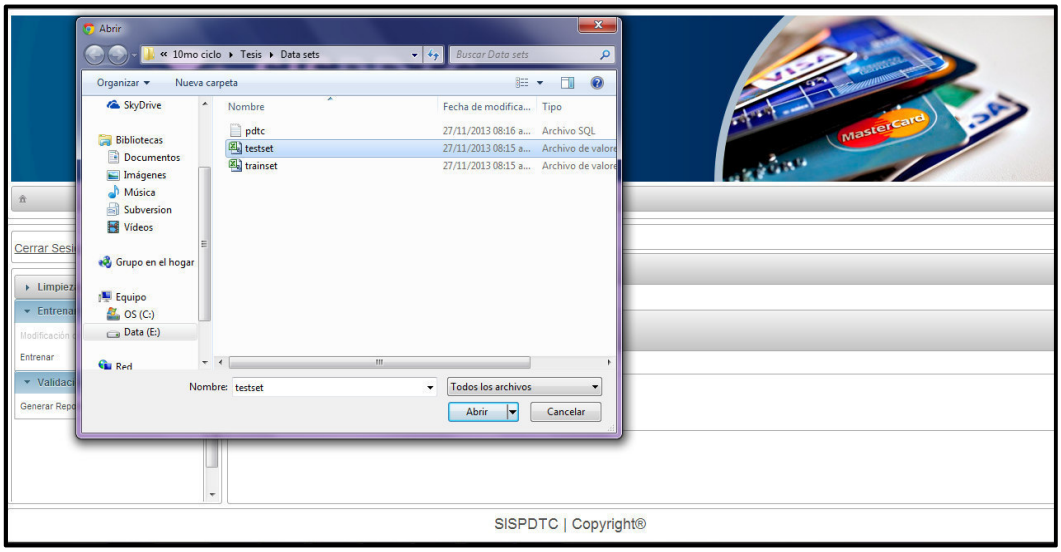
3. Pre-Condiciones

Al usuario se le ha debido asignar el rol correspondiente para hacer esta tarea.

4. Post-Condiciones

No aplica.

5. Prototipos




SISPDTC
Sistema de predicción de clientes desertores de tarjetas de crédito
Cerrar Sesion
Limpieza de Datos
Entrenamiento
Modificación de parametros del SVM
Entrenar
Validacion
Generar Reporte de Clientes Desertores
Entrenar
Elija el archivo de tipo ARFF:
Elegir Archivo
Entrenar
Esperando a localhost...
SISPDTC | Copyright®


SISPDTC
Sistema de predicción de clientes desertores de tarjetas de crédito
Cerrar Sesion
Limpieza de Datos
Entrenamiento
Validacion
Entrenamiento =>Tasa de Acierto Obtenido :1.0
SISPDTC | Copyright®

SISPDTC – SISTEMA DE PREDICCIÓN DE CLIENTES DESERTORES DE TARJETAS DE CRÉDITO

Especificación de Caso de Uso (ECU)

CU – Generar reporte de clientes desertores

Versión 1.0

1. Breve Descripción

El sistema permitirá generar un reporte de los clientes identificados como posibles desertores.

2. Flujo de Eventos

Flujo Básico << Generar reporte de clientes desertores>>

1. El caso de uso inicia cuando el usuario se identifica con su código y contraseña respectiva en el sistema y presiona el botón “Entrar”.
2. El sistema valida los datos ingresados por el usuario. Valida que el usuario tenga asignado el rol correspondiente para realizar esta operación.
3. El sistema muestra la pantalla principal.
4. El usuario selecciona la opción “Subir archivo”.
5. El sistema despliega una ventana donde le permite al usuario buscar el archivo que desea subir.
6. El usuario selecciona el archivo que desea entrenar (con extensión csv) y selecciona la opción “Aceptar”.

7. El sistema verifica que se haya seleccionado un archivo con la extensión csv.
8. El usuario selecciona la opción “Generar Reporte”.
9. El sistema muestra en pantalla una lista de los clientes identificados como posibles desertores.
10. El usuario selecciona la opción “Guardar”.
11. El sistema despliega una ventana donde le permite al usuario identificar en qué lugar de la PC desea guardar el archivo.
12. El usuario selecciona la ubicación donde desea que se guarde el reporte y selecciona la opción “Aceptar”.
13. El sistema guarda el reporte en la ubicación seleccionada por el usuario.
14. El usuario selecciona la opción “Salir”.
15. El sistema muestra la pantalla inicial de login.
16. Finaliza el caso de uso.

Flujos Alternativos << Generar reporte de clientes desertores>>

2.1 Los datos del usuario fueron ingresados incorrectamente. Si uno de los datos ingresados por el usuario es incorrecto, entonces se mostrará en la pantalla un mensaje de error “Datos incorrectos”. El caso de uso finaliza.
7.1 El sistema reconoce que el archivo seleccionado no es csv. Si el sistema reconoce que el archivo seleccionado no es csv (el único permitido por el sistema), entonces se mostrará un mensaje de error “Extensión inválida”. 7.2 El sistema despliega una ventana de archivos, donde le permite al usuario volver a seleccionar un archivo pero esta vez de extensión csv. Continua la secuencia.

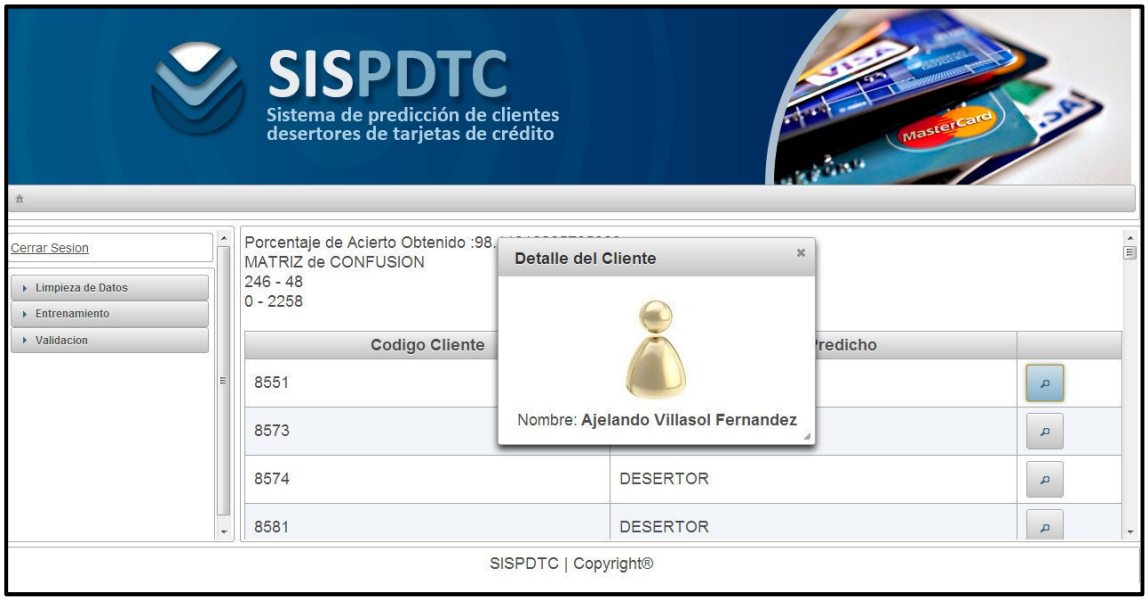
3. Pre-Condiciones

Al usuario se le ha debido asignar el rol correspondiente para hacer esta tarea.

4. Post-Condiciones

El reporte de clientes identificados como desertores.

5. Prototipos



SISPDTC – SISTEMA DE PREDICCIÓN DE CLIENTES DESERTORES DE TARJETAS DE CRÉDITO

Especificación de Caso de Uso (ECU)

CU – Modificar parámetros SVM

Versión 1.0

1. Breve Descripción

El sistema permitirá modificar los parámetros de la función kernel RBF (Radial Basis Function Kernel) del SVM (Support Vector Machine), el cual se está utilizando para entrenar el modelo y predecir la data.

2. Flujo de Eventos

Flujo Básico <<Modificar parámetros SVM>>

1. El caso de uso inicia cuando el usuario se identifica con su código y contraseña respectiva en el sistema y presiona el botón “Entrar”.
2. El sistema valida los datos ingresados por el usuario. Valida que el usuario tenga asignado el rol correspondiente para realizar esta operación.
3. El sistema muestra la pantalla principal, donde muestra una lista de todas las variables utilizadas en la función kernel RBF. Muestra los valores que actualmente utiliza el sistema en un campo editable.
4. El usuario modifica los valores que desee de cada variable.
5. El usuario selecciona el botón “Modificar”.
6. El sistema valida los datos ingresados por el usuario, que pertenezcan al rango

correspondiente de cada variable.

7. El sistema muestra una ventana de éxito la cual muestra el siguiente texto “Los valores de las variables han sido correctamente actualizados”.
8. El usuario selecciona la opción “Salir”.
9. El sistema muestra la pantalla inicial de login.
10. Finaliza el caso de uso.

Flujos Alternativos <<Modificar parámetros SVM>>

2.1 Los datos del usuario fueron ingresados incorrectamente. Si uno de los datos ingresados por el usuario es incorrecto, entonces se mostrará en la pantalla un mensaje de error “Datos incorrectos”. El caso de uso finaliza.
6.1 El sistema reconoce un error en uno de los valores ingresados por el usuario. Si el sistema reconoce un error en uno de los valores ingresados por el usuario le mostrará en la pantalla un mensaje de error “Los valores ingresados para la variable <nombre de la variable> fueron incorrectamente ingresados”. 6.2 El sistema le vuelve a mostrar la pantalla inicial. Regresa al punto 4.

3. Pre-Condiciones

Al usuario se le ha debido asignar el rol correspondiente para hacer esta tarea.

4. Post-Condiciones

Las variables de la función kernel RBF modificadas.

5. Prototipos

Aún no implementado. Opción deshabilitada.

SISPDTC – SISTEMA DE PREDICCIÓN DE CLIENTES DESERTORES DE TARJETAS DE CRÉDITO

Especificación de Caso de Uso (ECU)

CU – Obtener componentes principales

Versión 1.0

1. Breve Descripción

El sistema permitirá leer un archivo con extensión arff el cual tendrá las variables usadas así como el contenido de los registros de los clientes en un determinado rango de fecha para poder aplicarle el análisis de componentes principales.

2. Flujo de Eventos

Flujo Básico <<Obtener componentes principales>>

1. El caso de uso inicia cuando el usuario se identifica con su código y contraseña respectiva en el sistema y presiona el botón “Entrar”.
2. El sistema valida los datos ingresados por el usuario. Valida que el usuario tenga asignado el rol correspondiente para realizar esta operación.
3. El sistema muestra la pantalla principal.
4. El usuario selecciona la opción “Obtener Componentes Principales”.
5. El sistema despliega una ventana donde le permite al usuario buscar el archivo que desea procesar.
6. El usuario selecciona el archivo y selecciona la opción “Aceptar”.

7. El usuario selecciona la opción “Procesar”.
8. El sistema procesa el archivo seleccionado.
9. El sistema muestra un mensaje de éxito en el entrenamiento y los nuevos atributos.
10. Finaliza el caso de uso.

Flujos Alternativos <<Obtener componentes principales>>

2.1 Los datos del usuario fueron ingresados incorrectamente. Si uno de los datos ingresados por el usuario es incorrecto, entonces se mostrará en la pantalla un mensaje de error “Datos incorrectos”. El caso de uso finaliza.

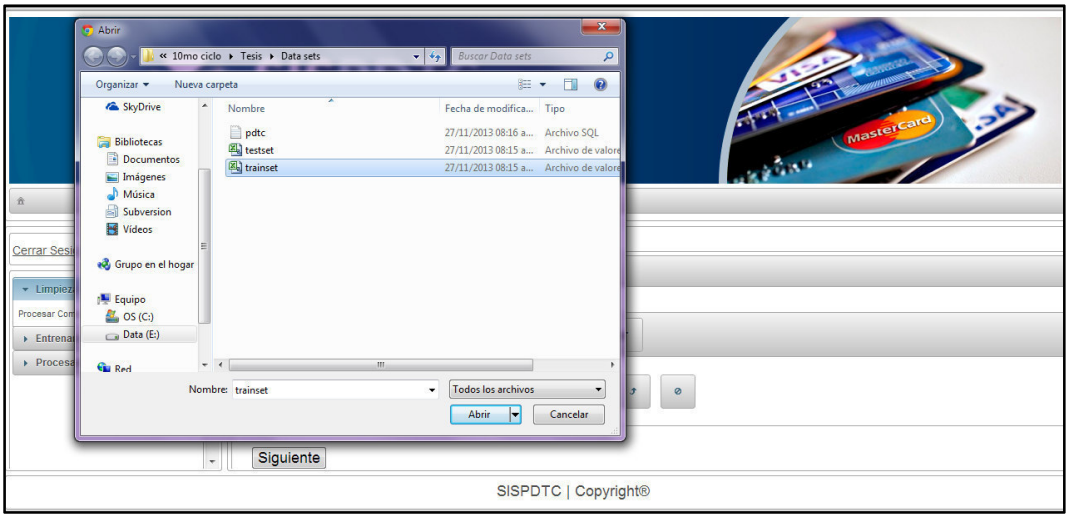
3. Pre-Condiciones

Al usuario se le ha debido asignar el rol correspondiente para hacer esta tarea.

4. Post-Condiciones

Un nuevo archivo creado con limpieza de datos.

5. Prototipos


Abrir
10mo ciclo ▸ Tesis ▸ Data sets
Buscar Data sets
Organizar
Nueva carpeta
SkyDrive
Bibliotecas
Documentos
Imágenes
Música
Subversion
Videos
Grupo en el hogar
Equipo
OS (C:)
Data (E:)
Red
Nombre
Fecha de modifica...
Tipo
pdtc
27/11/2013 08:16 a...
Archivo SQL
testset
27/11/2013 08:15 a...
trainset
27/11/2013 08:15 a...
Nombre: trainset
Todos los archivos
Abrir
Cancelar
Siguiente
SISPDTC | Copyright®


SISPDTC
Sistema de predicción de clientes desertores de tarjetas de crédito
Cerrar Sesion
Limpieza de Datos
Procesar Components Principales
Entrenamiento
Procesamiento
Importacion
Elija el archivo de tipo ARFF o CSV a usar:
Buscar Archivo
Importar
Cancelar
trainset.csv 880.93 KB
Siguiente
SISPDTC | Copyright®


SISPDTC
Sistema de predicción de clientes desertores de tarjetas de crédito
Cerrar Sesion
Limpieza de Datos
Procesar Components Principales
Entrenamiento
Validacion
Nombre de Archivo:
trainset.arff
Procesar
Atributos:
@relation trainset @attribute Codigo numeric @attribute X1 numeric @attribute X2 numeric @attribute X3 numeric @attribute X
numeric @attribute X5 numeric @attribute X6 numeric @attribute X7 numeric @attribute X8 numeric @attribute X9 numeric @a
X10 numeric @attribute X11 numeric @attribute X12 numeric @attribute X13 numeric @attribute X14 numeric @attribute X15 n
@attribute X16 numeric @attribute X17 numeric @attribute X18 numeric @attribute X19 numeric @attribute X20 numeric @attr
X21 numeric @attribute X22 numeric @attribute X23 numeric @attribute X24 numeric @attribute X25 numeric @attribute X26 n
@attribute X27 numeric @attribute X28 numeric @attribute X29 numeric @attribute X30 numeric @attribute X31 numeric @attr
X32 numeric @attribute X33 numeric @attribute Class {NODESERTOR} @data
Dataset:
SISPDTC | Copyright®

ANEXO B

DIAGRAMA DE ACTIVIDADES DEL CASO DE USO <INICIAR SESIÓN>

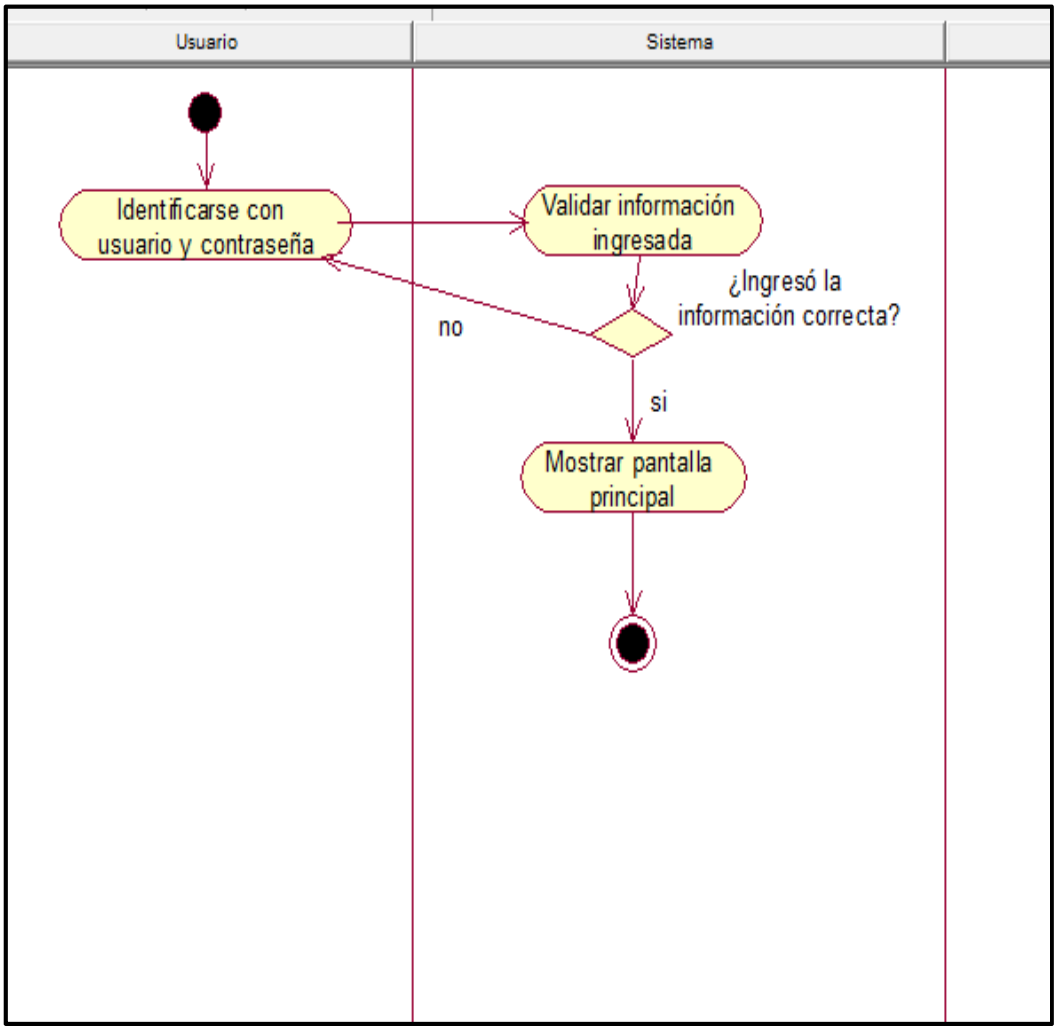


DIAGRAMA DE ACTIVIDADES DEL CASO DE USO <ENTRENAR DATA>

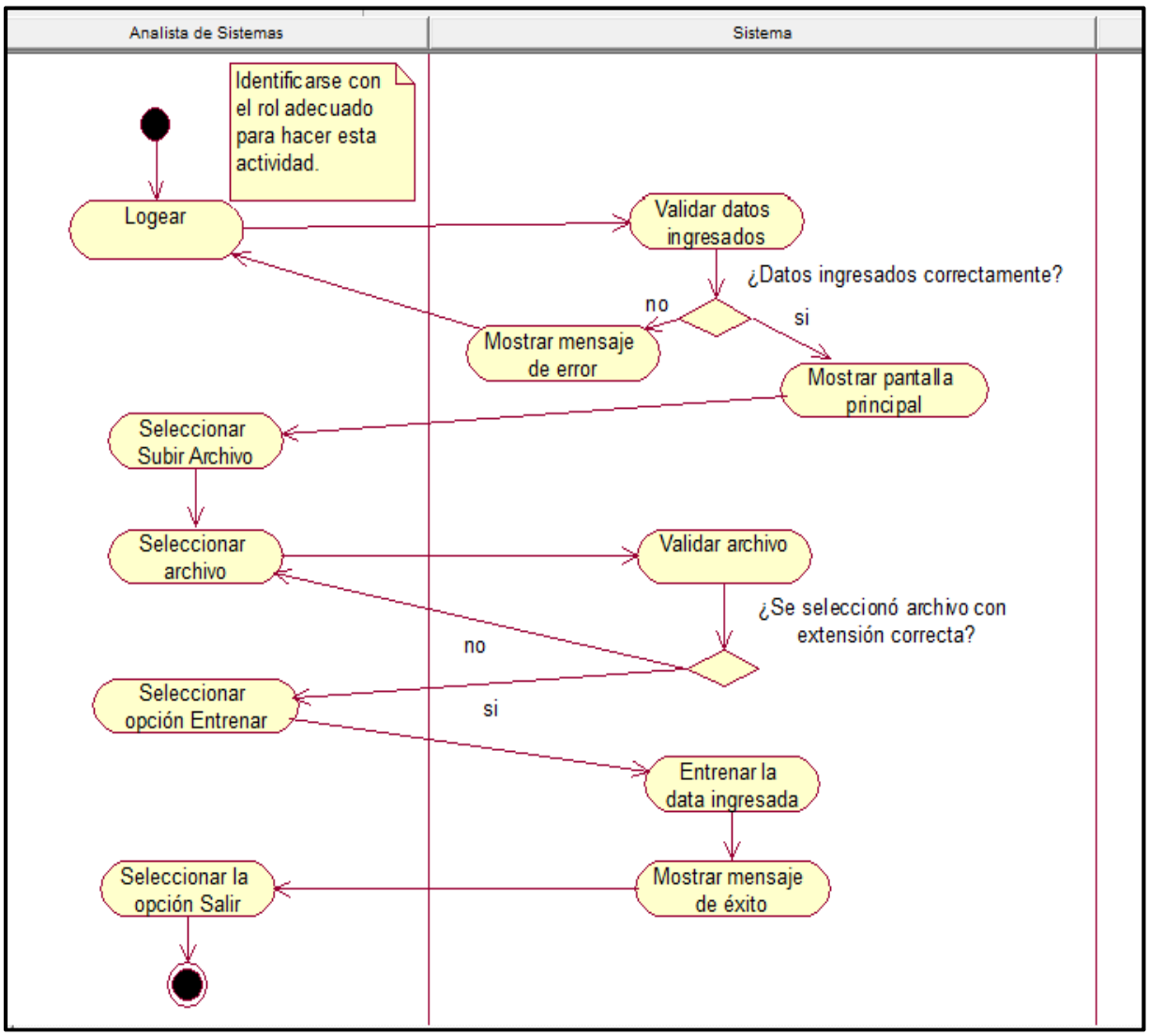


DIAGRAMA DE ACTIVIDADES DEL CASO DE USO <GENERAR REPORTE DE CLIENTES DESERTORES>

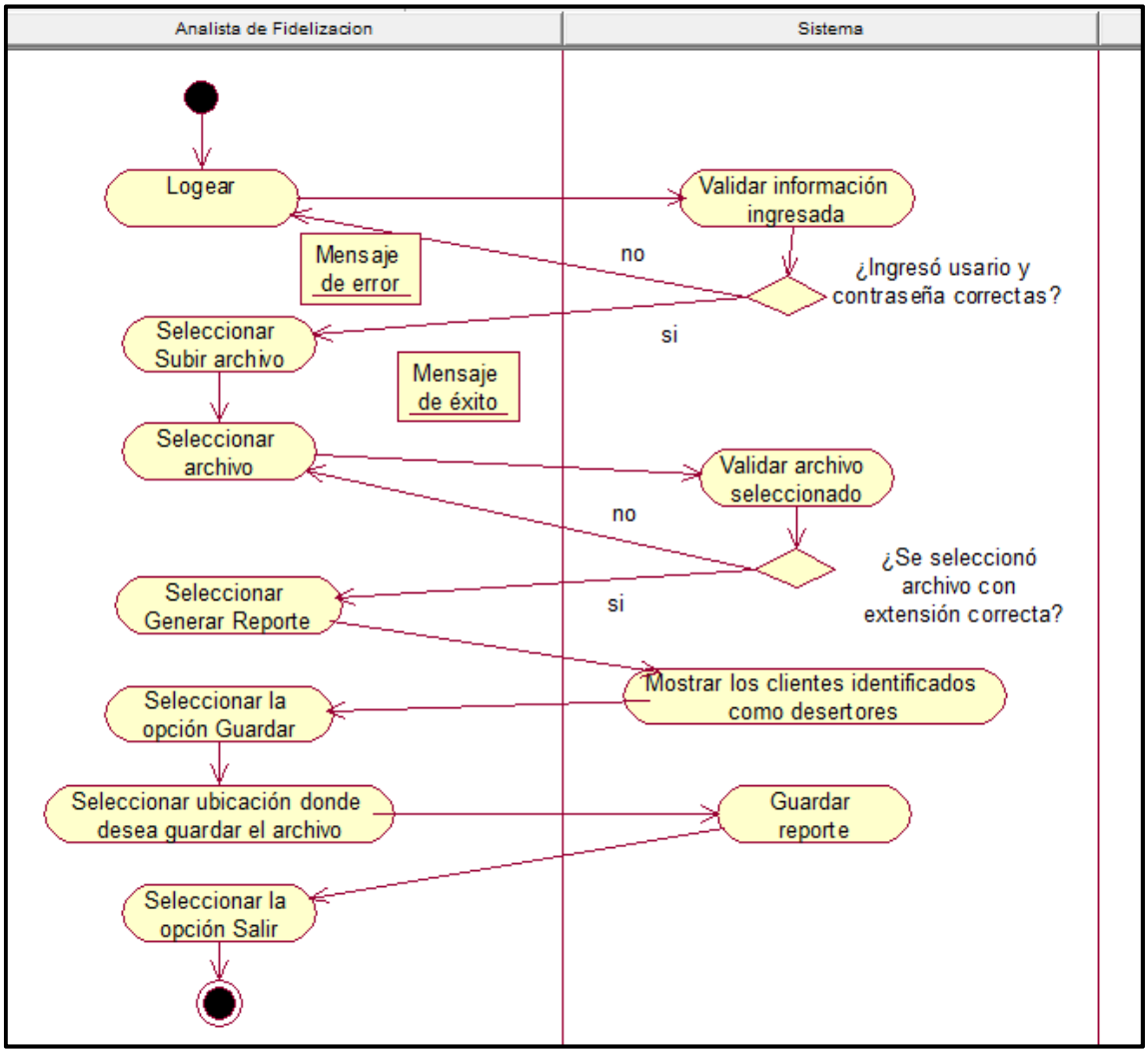


DIAGRAMA DE ACTIVIDADES DEL CASO DE USO <MODIFICAR PARÁMETROS SVM>

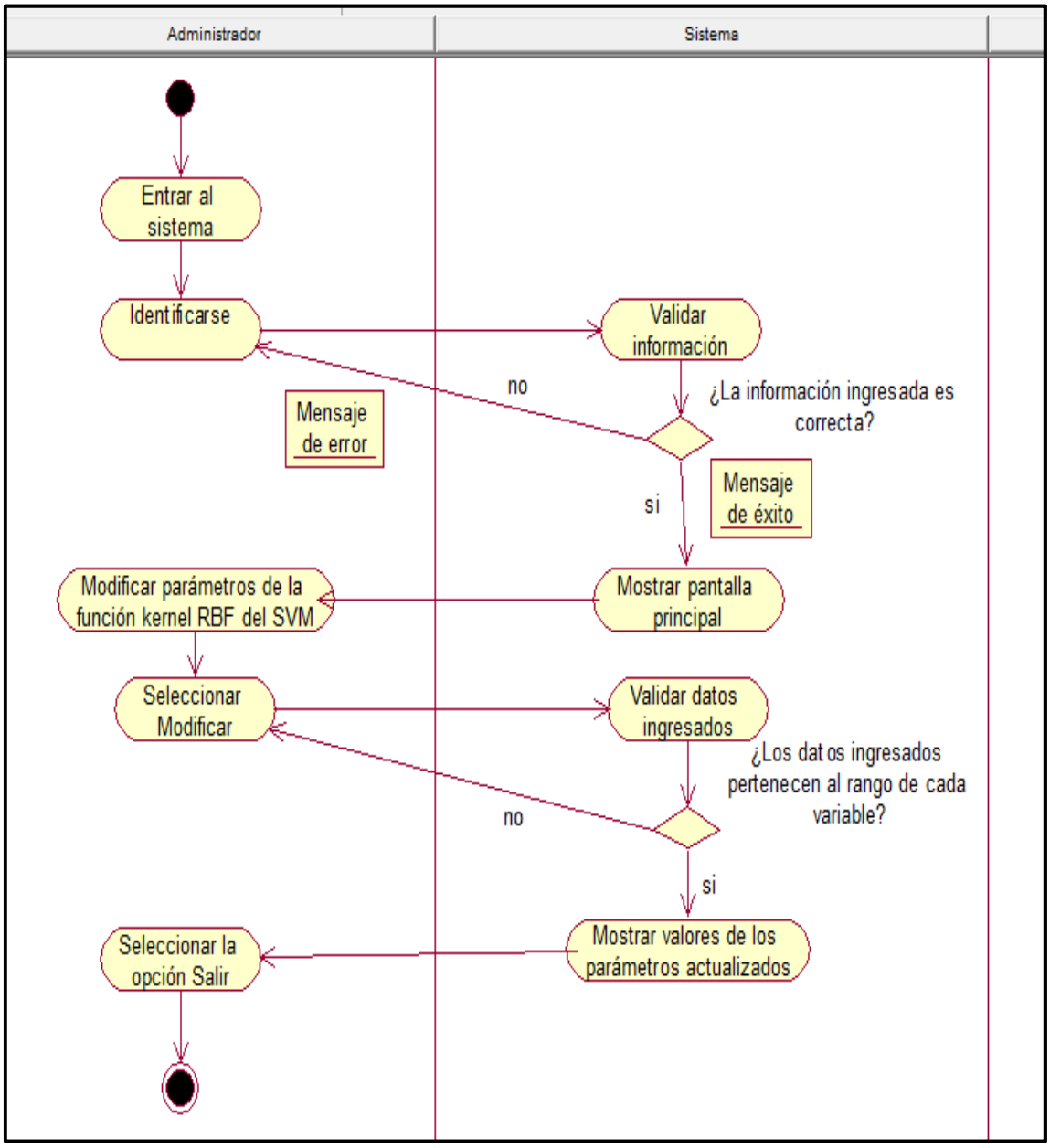


DIAGRAMA DE ACTIVIDADES DEL CASO DE USO <OBTENER COMPONENTES PRINCIPALES>