ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

"SISTEMA DE APROBACIÓN DE MICROCRÉDITO BASADO EN PATRONES DE POBREZA"

TÉSIS DE GRADO

Previo a la obtención del título de:

INGENIERO EN COMPUTACIÓN
ESPECIALIZACIÓN EN SISTEMAS DE INFORMACIÓN
INGENIERO EN COMPUTACIÓN
ESPECIALIZACIÓN EN SISTEMAS TECNOLÓGICOS
INGENIERO EN COMPUTACIÓN
ESPECIALIZACIÓN EN SISTEMAS TECNOLÓGICOS

Presentado por:

María Eugenia Andrade Ramírez

Enrique Guido Anchundia Aguirre

Néstor Alejandro Uyaguari Espinoza

Guayaquil – Ecuador

2006

AGRADECIMIENTO

A todas las personas que de uno u otro modo colaboraron en la realización de este trabajo y especialmente al MSC. Fabricio Echeverría Director del Tópico, Ing. Carlos Monsalve y al MSC. Federico Raue.

DEDICATORIA

A mis padres, por el esfuerzo, el apoyo y la buena educación de valores que me han sabido inculcar desde el seno de mi hogar.

A mí, por la perseverancia y la dedicación puesta en cada uno de mis objetivos a cumplir.

A mi enamorado por la paciencia, comprensión y apoyo que me brinda en cada momento de mi vida.

Ma. Eugenia Andrade

A Dios por darme la sabiduría que he adquirido en el periodo estudiantil y que seguiré adquiriendo.

A mis padres por enseñarme e inculcarme buenos valores y principios, por apoyarme y aconsejarme a lo largo de la vida estudiantil y laboral, por su preocupación como padres para seguir desarrollándome como ser humano y como profesional.

A mi enamorada por su apoyo incondicional que me ha dado y por la paciencia que ha tenido.

Enrique Anchundia

A Dios quien me ha llenado de bendiciones en mi vida y me ha dado la perseverancia necesaria para terminar mis estudios.

A mi familia, quienes han estado conmigo en buenos y malos momentos en especial mis padres: Rosa María Espinoza y Segundo Diecil Uyaguari.

A mis amigos que han estado apoyándome en todas las decisiones que tomo en mi vida.

Alejandro Uyaguari

TRIBUNAL DE GRADUACIÓN

Ing. Holger Cevallos
SUBDECANO DE LA FIEC
PRESIDENTE

MSC. Fabricio Echeverría
DIRECTOR DE TÓPICO

Ing. Carlos Monsalve
MIEMBRO DEL TRIBUNAL

MSC. Federico Raue
MIEMBRO DEL TRIBUNAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, nos corresponden exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL)

Ma. Eugenia Andrade Ramírez

Enrique Anchundia Aguirre

Néstor Uyaguari Espinoza

RESUMEN

El sistema consistirá aprobar microcrédito a individuos que estén por debajo del umbral de pobreza de una zona dada, para dicho propósito inicialmente se realizara el levantamiento de la información en ciertos sectores marginados realizando un análisis del Índice de la Vivienda de CASPHOR (CHI) para evaluar el nivel de pobreza de las personas. Tomando como base la información obtenida inicialmente, el análisis para la aprobación de microcrédito, incluirá también otros parámetros generados a través de la aplicación de métodos de minería de datos y análisis multivariante , tales como análisis discriminante, análisis de cluster, árboles de decisión y reglas de asociación.

El primer capítulo se enfocará en explicar y describir el micro crédito a nivel mundial y local, así como también el análisis de la problemática en cuanto a la proceso de concesión de micro crédito y la solución al mismo, exponiendo las diferentes metodologías existentes para medir la pobreza y justificando la más adecuada para ser aplicada.

El segundo capítulo comprenderá el análisis y diseño de la solución del problema, es decir del sistema propuesto para la concesión de microcrédito para la clase marginada, para lo cual se establece las especificaciones del problema a través de los casos de uso, escenarios, DIOS, etc.; se modela la base de datos, se especifica las tablas del modelamiento multidimensional, el diseño de las pantallas del sistema y las herramientas a utilizar para su diseño.

El tercer capítulo se enfocará en detallar todo el análisis comprendido en lo referente al círculo virtuoso de la minería de datos involucrada en el sistema a implementar. Es decir, se indicará como se obtendrá los datos iniciales, que técnicas se utilizarán para minar los datos, que información nueva se genera y como se actualizara dicha información en la base de datos.

El cuarto capítulo comprenderá el análisis económico del sistema a implementar, es decir se realizará un análisis costo-beneficio, así como también un análisis comercial del sistema para conocer las debilidades, amenazas, fuerzas y oportunidades del sistema y establecer una mejor distribución de costos y estrategias comerciales del producto final.

Por último se detallarán las conclusiones y recomendaciones, anexos y bibliografía.

ÍNDICE GENERAL

Pág.

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

DIOS de diseño

INTRODUCCIÓN

La pobreza es uno de los grandes males que aquejan a las sociedades del mundo, impidiendo el desarrollo económico y social de sus habitantes. A pesar de que instituciones financieras poseen programas de microcrédito, la realidad muestra que no benefician a los más desfavorecidos. Dando lugar, a la necesidad de establecer un enfoque hacia los grupos considerados vulnerables solucionando el error de excluirlos.

Por tal motivo, a partir de 1997, la cumbre de microcrédito establece la importancia que el microcrédito sea una salida para los sectores de menores ingresos - que generalmente no tienen acceso a ningún tipo de financiamiento bancario - por medio del uso de metodologías que logran indicar el nivel de pobreza de una persona; tales como: el índice de la vivienda de Cashpor[1] y la calificación participativa del patrimonio de Small Enterprise Foundation.[2] [5]

---

[1] Credit and Saving for the Hardcore Poor (CASHPOR) institución microfinanciera cuya misión es identificar y motivar a mujeres pobres en áreas rurales para entregarles crédito.

[2] Institución microfinanciera que trabaja por la erradicación de la pobreza a través de la creación de un ambiente adecuado donde los servicios de crédito y ahorro sean más adaptables.

Dado este conocimiento "El sistema de aprobación de microcrédito basado en patrones de pobreza" (SAMBP), utiliza como herramienta una de las metodologías para medir los niveles de pobreza denominado Índice de Cashpor, la cual se basa en el análisis de la estructura de la vivienda como medio para distinguir los niveles económicos de los hogares y para identificar a los más pobres. Y a su vez, complementada con la utilización de técnicas predictivas y descriptivas de minería de datos tales como "Enhanced K-means" y "Adaptive Bayes Network" (ABN), dan lugar a la generación de patrones de pobreza que proporcionan el conocimiento necesario para establecer si una persona puede ser considerada para la aprobación de microcrédito. Por lo que "SAMBP" logra ser una herramienta útil de análisis de pobreza, y medio referencial de decisión para aprobar microcrédito a los más pobres.

Esta propuesta adicionalmente, sugiere fomentar la iniciativa que en el Ecuador las instituciones financieras orienten sus servicios hacia la clase más desprotegida de la sociedad. Considerándola como una forma de retribuir la confianza que la sociedad les ha otorgado durante años, mostrando que poseen responsabilidad social corporativa.

CAPÍTULO 1

1. CONCESIÓN DE MICROCRÉDITO

1.1. Microcrédito

1.1.1. Definición

El fenómeno del microcrédito nació en los años 70 en Bangladesh, cuando el Banco Grameen[3] comenzó a otorgar pequeños préstamos a personas demasiado pobres como para que los bancos convencionales les otorgaran un préstamo. El microcrédito consiste en entregar pequeños préstamos de dinero a familias pobres, que utilizan como capital de trabajo o para comenzar un pequeño negocio en los países en vías de desarrollo. Siendo por más de seis años la Campaña de la Cumbre del Microcrédito - un proyecto del Fondo Educativo Resultados, organización no gubernamental con sede en

[3] http://www.grameen-info.org/bank/bank2spanish.html

Washington, organizada por: líderes, empresarios y organismos de 137 países- la entidad que viene impulsando este propósito. El préstamo, en general concedido sin garantías, es devuelto habitualmente en un periodo de hasta seis meses o hasta un año. [1]

1.1.2. Concesión de microcrédito en el mundo

"PlaNet Finance" ha sido creada con el objetivo de aumentar la facilidad de acceso al microcrédito en el mundo. Así, provee soporte técnico, reduce los costos del microcrédito y favorece la transparencia de este sector gracias a la promoción de las tecnologías de la información. La organización propone también servicios de evaluación y organiza: talleres y conferencias, a fin de acercar, a los dirigentes y profesionales del microcrédito. En el curso de estos últimos años, tuvo un impacto importante en el sector del microcrédito. [2]

La sede de PlaNet Finance se encuentra en París, Francia. La red comprende una serie de ONGs afiliadas a PlaNet Finance en diferentes países, tal como lo muestra la Figura 1.

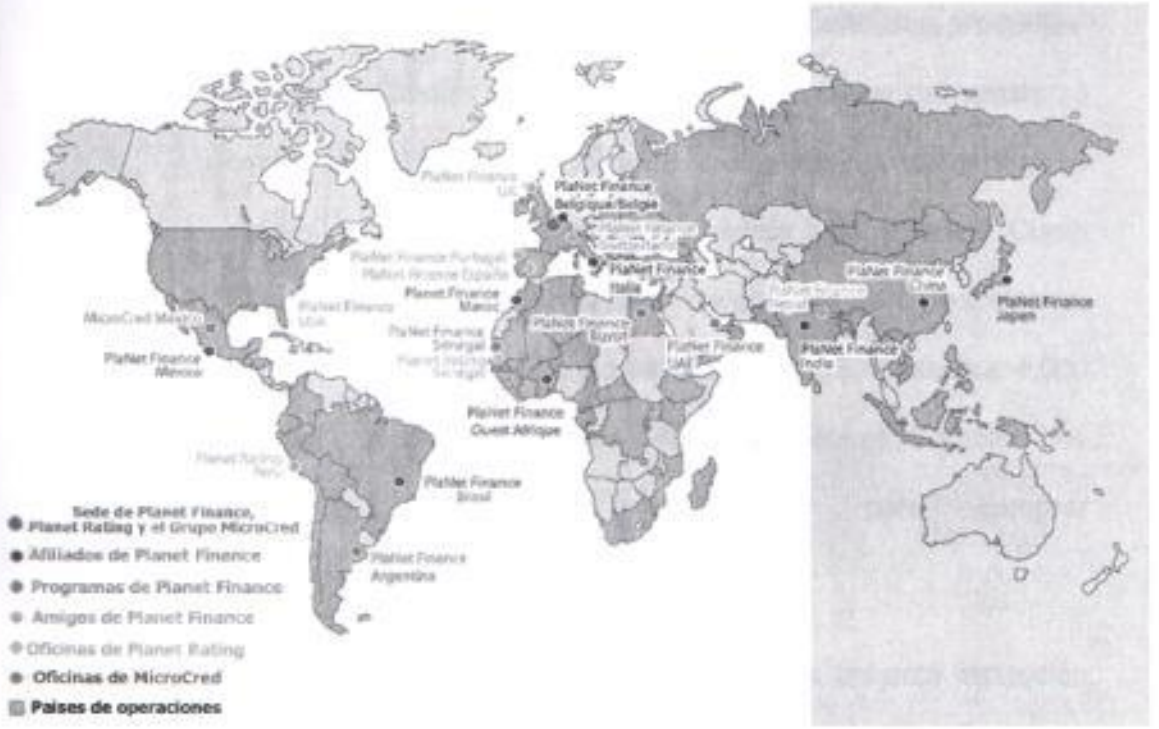


Figura 1. Países miembros de Planet Finance[4]

1.1.3. Concesión de microcrédito en el Ecuador

En el Ecuador, ciertas instituciones financieras brindan el servicio de microcrédito entre ellas tenemos:

UNIBANCO: Unibanco se autodefine como una entidad especializada en microcréditos. Los préstamos están orientados a: empleados, comerciantes y profesionales

[4]Planet Finance: http://www.planetfinance.org/ES/ong-estructura/ong-red-internacional.php

independientes (médicos, etc.). El banco tiene dos productos. "Línea de Crédito por Cuotas", donde el rango del préstamo oscila entre 80 y 1,500 dólares y el plazo va de seis a 18 meses. La tasa de interés es la máxima legal. "Tarjeta Cuota Fácil", es una línea directa de crédito en: papelerías, supermercados, etc. El cupo va desde los 150 hasta los 1,000 dólares. La compra mínima es de cuatro dólares. También hay convenios con casas comerciales para comprar electrodomésticos. [3]

BANCO CENTRO MUNDO: Los productos de esta institución se orientan a quienes disponen de rentas fijas mensuales (trabajadores, comerciantes, etc.). Los montos de los créditos pueden ir desde los 100 hasta los 1,600 dólares. Los plazos van de los seis a los 18 meses y el monto se define de acuerdo a la capacidad de ahorro que dispone el beneficiario. La tasa de interés es del 22.6%. La tasa de interés es fija durante todo el crédito. Para acceder al préstamo no es necesario ser cliente del banco. Los requisitos son: copia de la cédula, papeleta de votación, recibo luz o agua y carnét de afiliación al Instituto Ecuatoriano de Seguridad Social. [3]

BANCO DEL PICHINCHA: El Banco del Pichincha ofrece su "Crédito Preciso", donde presta entre 1,000 y 10,000 dólares. El

dinero se entrega a seis meses plazo o a dos años de acuerdo a los requerimientos del cliente. La tasa de interés anual actual es del 20%.

Los clientes tienen que pagar las cuotas del préstamo mensualmente, pero el cliente puede elegir el día de pago. La cuota es fija. El banco acepta como garantías las siguientes: personal, empresarial, hipotecaria y documento financiero. El interesado debe llenar la solicitud del crédito y presentar los documentos de soporte tales como: copias de la cédula, copia recibo de agua, luz o teléfono del domicilio, rol de pagos o certificado de ingresos si el garante es asalariado y documento de certificación del negocio. [3]

BANCO SOLIDARIO: El Banco Solidario se especializa en créditos a microempresas, hasta por 12,000 dólares y a una tasa de interés del 18% a 18 meses de plazo. Otros productos son la tarjeta La Chauchera, cuyos beneficios son: crédito preaprobado, descuentos en la cadena de proveedores y almacenes, cupo para compra de vivienda, atención médica y red de cajeros Banred. Otro producto es Ahorro Propósito. Es un esquema en que el cliente ahorra hasta el 30% del valor de un bien y el banco le presta el 70% restante. Las condiciones son iguales a las del crédito para microempresa. [3]

BANCO MM JARAMILLO: Esta institución ofrece créditos personales para mayores de 25 años, con montos que van desde 500 hasta 5,000 dólares. El plazo máximo es de tres años con una tasa fija. Para los créditos superiores a los 5001 dólares se puede acceder a un plazo mayor, de hasta cinco años con tasas de interés reajustables cada 90 días. En los créditos de 500 a 10,000 dólares, se requiere solo un garante. Para montos superiores, se requieren garantías reales (activos, inmuebles, etc.). También hay crédito para comprar vehículos, al 17% anual. [3]

1.2. Planteamiento del problema del microcrédito

El grado en que los programas microfinancieros pueden alcanzar a los más pobres entre los pobres, continúa siendo una discusión abierta. No hay un acuerdo generalizado acerca de que, para tener un verdadero impacto en la pobreza, las microfinancieras deban dirigirse expresa y exclusivamente a los más pobres. [4]

Sin embargo, a partir de 1997, cuando surge la Cumbre de Microcrédito, se convierte en un punto de constante discusión, se establece la meta de varios organismos internacionales de lograr que

100 millones de las familias más pobres del planeta tengan acceso a servicios microfinancieros en el 2005. [4]

Los cuatro temas centrales de la declaración y el plan de acción de la cumbre son: [4]

1) servir a los más pobres,

2) servir y fortalecer a la mujer,

3) formar instituciones autosuficientes financieramente y

4) asegurar un impacto positivo y mensurable en las vidas de los clientes y sus familias.

Las instituciones microfinancieras (IMF) al menos han aceptado la recomendación de que son necesarias algunas medidas cuidadosas para determinar a qué sector está atendiéndose y dar preferencia a los más pobres.

1.3. Metodologías para evaluar la pobreza

- **Índice de la vivienda de Cashpor**

El Índice de Vivienda de Cashpor (CHI)[5] es una metodología externa y de observación que ofrece una forma rentable de identificar a las personas muy pobres en base al análisis de la condiciones de la vivienda, inicialmente aplicado en toda el Asia rural. Es una adaptación de la propuesta de Grameen[6] para focalizar la pobreza. [4]

Las instituciones que son miembros de la red de Cashpor han encontrado que esta herramienta les permite identificar con rapidez y con un 80% de exactitud a las familias muy pobres en una zona dada. El CHI impone mirar a los rasgos más importantes de una casa tales como: tipo de vivienda, materiales de construcción de paredes, techo, piso, tenencia de la vivienda, forma de abastecimiento del agua para beber, medio de eliminación de aguas servidas, medio de eliminación de basura y tipo de combustible para cocinar. Estos criterios se adaptan según las situaciones locales de cada país. Cada componente recibe puntuaciones o pesos. Se marca una puntuación de corte (que sea localmente relevante) para que las familias cuyas casas puntúen por encima de cierto número no sean elegibles para participar en el programa de crédito, mientras que las

---

[5] Cashpor Housing Index

[6] Organización de microcrédito iniciado en Bangladesh, por el economista Muhammad Yunus con el fin de entregar pequeños prestamos de dinero a personas de bajos recursos, especialmente mujeres. Sin necesidad de garantía, basado simplemente en la confianza mutua, la solidaridad, la responsabilidad y la participación creativa de los beneficiarios.

familias que puntúan por debajo de cierto número de puntos cumplen los requisitos para ser considerados, normalmente usando una prueba administrada por los empleados. Se establece un proceso de apelación para aquellas familias que afirman ser muy pobres pero cuyas casas sacaron una puntuación demasiado alta para ser consideradas para el programa. El CHI no funciona bien en los pueblos con viviendas suministradas por el gobierno; en estos casos la evaluación se puede sustituir por una especie de forma participatoria de clasificación de riqueza. [4]

El primer paso para llevar a cabo el CHI es hablar con oficiales locales informados (en los departamentos de agricultura, salud, educación o bienestar) para localizar en cuál zona geográfica está el mayor número de familias pobres. Segundo, un miembro del personal del programa hace un mapa del camino de las zonas sugeridas. Esto ofrece una lista de barrios con las familias que parecen ser las más elegibles. Una vez que se localizan los barrios, se lleva a cabo una aplicación más intensiva del CHI en estas zonas para que cada casa se dibuje en el mapa y los que cumplen los requisitos son claramente marcados. [4]

Cuando es necesario, una familia puede apelar a su falta de elegibilidad. En este caso, empleados principales llevan a cabo una

entrevista para determinar si la familia está entre la población deseada, a pesar que la evidencia externa muestre lo contrario. La opción de apelación asegura que no se impida la participación en los programas de algunos de los pobres escondidos. [4]

- **Calificación participativa del patrimonio de Small Enterprise Foundation**

La Clasificación de riqueza participativa (PWR)[7] es una modificación de la técnica de la evaluación rural participativa. Es una clasificación de riqueza subjetiva y muy local usada por miembros de una comunidad para averiguar qué miembros son los más necesitados. Éstos generan su propio criterio para clasificar pobreza o riqueza; esto a menudo incluye factores que no son visibles ni fáciles de identificar por alguien de fuera de la comunidad. Al hacer participar a los miembros de la comunidad en el proceso, la gente analiza sus propias situaciones, dando una mayor propiedad a los programas que están establecidos para ayudar a los más pobres en la comunidad. [4]

[7] Participatory Wealth Ranking.

1.4. Solución del problema

- **Estrategias de enfoque a la pobreza**

Las IMF han desarrollado diversas estrategias para enfocarse a la pobreza, que incluyen: [4]

- Formas de identificar a los pobres.
- Formas de atraer a los pobres.
- Formas de excluir al no pobre.
- Formas de desalentar al no pobre.

Para lograr enfocarse eficazmente a los más pobres, necesitan tomarse en cuenta:

- Factores relacionados con los agentes: tipo de IMF, necesidades del cliente, restricciones
- Factores del contexto (marco regulatorio, infraestructura, etc.)
- Resultados de las microfinancieras (a cuántas personas atienden, qué tan pobres son los clientes, en cuáles sectores participan, dónde viven, calidad de los servicios ofrecidos)

• Impacto (metodologías de medición)

Para ello el Comité Ejecutivo de la Campaña de Microcrédito recientemente aprobó la creación del "Conjunto de Herramientas para Medir la Pobreza" (PMTK). El Comité Ejecutivo de la Campaña ha acordado que las primeras dos medidas incluidas en el conjunto de herramientas sean el Índice de vivienda de Cashpor de uso en la Asia rural y la Clasificación de riqueza participativa. Aunque la cumbre no necesariamente propone que estas medidas sean usadas para seleccionar clientes, los programas de microcrédito que deseen servir a las familias más pobres y tengan una metodología que incorpora fácilmente estas herramientas pueden encontrarlas útiles al evaluar el nivel de pobreza de los clientes a los que sirven. Estas herramientas y otras a ser identificadas, serán usadas para evaluar el progreso hacia el objetivo de la cumbre de alcanzar a 100 millones de las familias más pobres para conceder microcrédito. [4]

1.5. Justificación de la metodología a utilizar

La metodología del Índice de la vivienda Cashpor presenta diversas ventajas por las cuales justifica el hecho de su elección para el desarrollo de patrones de pobreza. [4]

Ventajas del CHI:

- Capacidad de distinguir entre las familias pobres y las muy pobres. [4]

- Invierte un bajo costo al eliminar muchas familias de la elegibilidad. La mayor parte de los costos y esfuerzos se concentran en encontrar las familias sumamente pobres. [4]

- Los criterios usados para calificar las casas pueden ser localmente adaptados al mismo tiempo manteniendo la objetividad debido a que son ejecutados por empleados capacitados del programa. [4]

- Los empleados pueden ser capacitados fácilmente para hacer evaluaciones. [4]

- Entrevistas de seguimiento ayudan a prevenir corrupción y aseguran datos más fidedignos. [4]

- La opción de apelación ofrece un método para poder alcanzar a la mayoría de los más pobres escondidos. [4]

CAPÍTULO 2

2. CÍRCULO VIRTUOSO DE LA MINERÍA DE DATOS

¿Qué es la mineria de datos?

La mineria de datos (Data Mining) ayuda a las organizaciones a encontrar información que no es perceptible de forma directa, como por ejemplo patrones de comportamiento, relaciones, asociaciones, etc. que nos permitan tomar mejores decisiones. A través del análisis del pasado, y aplicando algoritmos, se construyen predicciones que nos permiten mejorar nuestra eficiencia y conseguir así una mayor rentabilidad de la actividad de negocio. [6]

A través de este proceso de minería se consigue:

- Sacar perfiles de los clientes y entender su comportamiento.
- Fidelizar a los clientes ofreciendo lo que ellos esperan.
- Mejorar los beneficios y márgenes.

- Aumentar nuestra eficacia incrementando nuestra competitividad.

La minería de datos analiza la información en profundidad, a través de algoritmos de aprendizaje, recorre los datos para descubrir patrones e información encubiertos, y basándose en estas conclusiones construye predicciones de futuro. [6]

Un concepto importante reside en que la generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la definición del problema básico que resolverá el modelo hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los seis pasos básicos siguientes: [7]

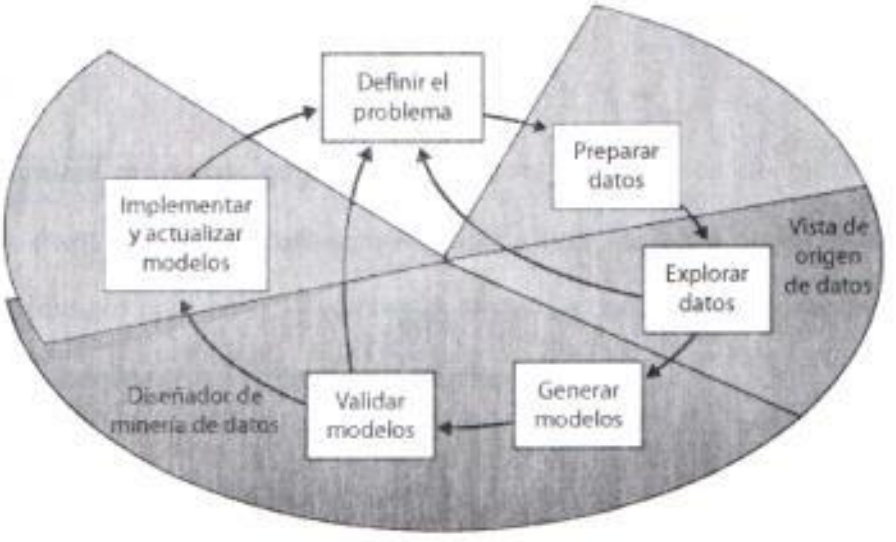


Figura 2. Círculo virtuoso de la minería de datos

1. **Definir el problema:** consiste en definir claramente el problema empresarial, definir el ámbito del problema, definir el atributo a predecir y definir el objetivo final del proyecto de minería de datos.

2. **Preparar los datos:** consiste en consolidar, limpiar y transformar los datos identificados en el paso "Definir el problema". Los datos pueden estar dispersos en la empresa y almacenados en distintos formatos; también pueden contener incoherencias como entradas que faltan o contener errores.

3. **Explorar los datos:** consiste en analizar los datos, para decidir si el conjunto de datos contiene datos con errores; y, a continuación, transformar los datos (normalización, discretización) dependiendo del tipo de valores numéricos y categóricos respectivamente. Es necesario señalar que este paso puede ser incluido en el paso "Preparación de datos".

4. **Generar modelos:** consiste en generar los modelos de minería de datos. Para lo cual, se selecciona y aplica una variedad de técnicas de modelamiento para obtener valores óptimos. La preparación y limpieza de datos es necesario ejecutarlo previamente.

5. **Explorar y validar los modelos:** Consiste en analizar los modelos construidos en el paso anterior, para estar seguros de dichos modelos tienen un alto grado de confianza.

6. **Implementar y actualizar los modelos:** Consiste en aplicar los modelos generados para alcanzar los objetivos trazados en la definición del problema. Como por ejemplo, predecir si una persona es pobre o muy pobre de acuerdo a patrones de pobreza previamente extraídos.

Es necesario señalar que los pasos del círculo virtuoso mencionados anteriormente, se basan en el estándar que rige los procesos de minería de datos: CRISP-DM Project (Cross-Industry Standard Process for Data Mining). Para mayor información ingresar a: http://www.crisp-dm.org/Process/index.htm

Mineria de datos de Oracle

La base de datos Oracle incluye funcionalidad para la mineria de datos en la edición Enterprise. Esta funcionalidad está totalmente integrada y bajo el mismo motor que la parte relacional de la misma. Se puede acceder a toda la funcionalidad Data Mining a través de la API Java que incluye la base de datos, de manera que las aplicaciones puedan sacar el máximo partido de las funciones disponibles. [6]

ODM[3] permite a las compañías extraer información oculta usando una amplia gama de algoritmos. Los algoritmos de minería de datos son técnicas de aprendizaje que sirven para analizar los datos en especificas categorías de problemas. Los algoritmos pueden ser separados dentro de dos técnicas de minería de datos: "aprendizaje supervisado o predictivo" y "aprendizaje no supervisado o descriptivo". [22]

El aprendizaje supervisado, requiere para el análisis de los datos, identificar un atributo clase o variable dependiente, para encontrar patrones y relaciones entre los atributos independientes (predictores) y el atributo dependiente. Un atributo clase es definido para describir, por ejemplo, cuales clientes han comprado recientemente un nuevo carro – por ejemplo, "1" para "SI" y "0" para "NO". Para luego construir un modelo que pueda ser usado para clasificar nuevos datos y hacer predicciones con respecto a dicho atributo clase. Las técnicas supervisadas soportadas por ODM son: clasificación y regresión lineal. De igual manera, los respectivos algoritmos son: "Naive Bayes" (NB), "Decision Tree" y "Adaptive Bayes Network" (ABN) para clasificación; y, "Support Vector Machines" para problemas de regresión lineal. [22]

El aprendizaje no supervisado, sirve para describir las características que poseen los datos. En este aprendizaje no se requiere establecer un atributo clase. Las técnicas soportados por ODM son: reglas de

[3] Oracle Data Mining

clusterización y asociación. De igual manera, los respectivos algoritmos son: "Enhanced K-Means" y "O-Cluster" para clusterización; y, "Association Rules" para encontrar patrones de eventos que ocurren al mismo tiempo. [22]

Para el desarrollo del sistema **"SAMBP"** se utilizaron los siguientes algoritmos de Oracle Data Mining: "Adaptive Bayes Network" (ABN), como algoritmo para la función de clasificación; y "Enhanced K-Means" como algoritmo para la función de agrupamiento o clusterización. [22]
A continuación una breve introducción de cada uno:

"Adaptive Bayes Network" (ABN): Es un algoritmo propietario de Oracle que provee un rápido y escalable mecanismo de extracción de información predictiva (reglas), con respecto a un atributo objetivo. (Mayor detalle Sección 2.4.1)

"Enhanced K-Means": Es un algoritmo de agrupamiento jerárquico basado en distancias, que particiona los datos dentro de números predeterminados de clusters/grupos. "Enhanced K-Means" construye modelos en forma jerárquica, haciendo que el árbol crezca un nodo a la

vez. Es necesario señalar que la versión DBMS_DATA_MINING, soporta atributos numéricos y categóricos. [23](Mayor detalle Sección 2.4.1)

2.1. Definir el problema

Por medio de la aplicación de la minería de datos se pretende extraer patrones de pobreza a partir de un conjunto de datos obtenidos previamente, con el objetivo de proveer pautas para decidir si aprobar o no microcréditos a futuras personas de acuerdo a su nivel de pobreza establecido mediante la Metodología de Cashpor.

El atributo que se pretende predecir es: CONDICIÓN, relacionado al nivel de pobreza de una persona. (Ver Figura 32)

Las relaciones que se desea indagar son aquellas entre los valores de las características de viviendas.

Se pretende solo realizar predicciones y a su vez encontrar patrones interesantes que provean una guía para la aprobación de microcrédito.

2.2. Preparar los datos

- **Obtención de los datos**

Debido a que actualmente en el Ecuador, no se ha realizado una encuesta de niveles de pobreza, aplicando la metodología propuesta del Índice de la vivienda de Cashpor. Se procedió a la solicitud de los datos al INEC[9] del VI Censo de Población y V de Vivienda - 2001, debido a que era la única fuente de información que poseía datos que se asemejaban a los requeridos para el análisis. Para ello, se solicitó información de ciertos barrios de Guayaquil, considerados los más pobres. (La solicitud presentada se detalla en el Anexo C)

El INEC maneja los siguientes términos para organizar los datos de la población a nivel nacional:

SECTOR: conjunto de aproximadamente 150 viviendas.

ZONAS: conjunto de sectores.

BARRIO: conjunto de zonas.

Cabe recalcar, que el INEC solo pudo proveer los datos a nivel de sector y no de vivienda como se solicitó inicialmente, por lo que no se pudo laborar con esta información, debido a que manejan datos resumidos.

[9] Instituto Nacional de Estadísticas y Censos del Ecuador

Por tal motivo se consiguió datos de la Encuesta de Niveles de Vida 2003, realizado en Panamá, en el cual aplican la metodología "**Living Standard Measurement Study**" (**LSMS**), basada en el estudio de los niveles de gastos y consumo de las personas y los hogares, incluyendo una amplia gama de variables relacionadas con las características y condiciones de vida de una muestra seleccionada con representatividad nacional de los países. (Ver detalles Anexo B)

Con motivo de adaptarlo a nuestros requerimientos de datos se procedió a la elección de ciertas variables del Censo de Panamá, que se asemejaban a aquellas que han sido utilizadas para el análisis de Cashpor en otros países tales como: India y Filipinas. [8] Además, las variables seleccionadas guardan relación con las variables usadas por el INEC en el Censo de población y Vivienda - 2001. (Ver Tabla 1)

Se detalla a continuación las variables seleccionadas:

Variables	
V01	TIPO VIVIENDA
V02	PAREDES
V03	TECHO

V04	PISO
V05	TENENCIA
V19	AGUA BEBER
V29	SANITARIO
V33	BASURA
V38	COMBUSTIBLE

Tabla1. Cuadro de variables seleccionadas

Es necesario detallar que a los valores de las variables seleccionadas se le asignaron un valor numérico o peso, para poder obtener una puntuación total por vivienda y establecer la condición de "pobre" y "muy pobre" en base a un puntaje promedio (Ver Anexo B). El puntaje promedio se establece de acuerdo a las suma de lo valores máximos de cada variable divido para dos, y ese dicho resultado sumado uno.

Dado: V1, V2, V3...Vn

Siendo P= promedio

P= ((max (V1) + max (V2) +..... max (Vn)) / 2) + 1

Dado este análisis, el puntaje promedio de nuestro sistema es: 20

Finalmente, dichos pesos varían de acuerdo al contexto y dependiendo de la región o país de análisis. Para el caso del

desarrollo de "SAMBP", se basó en el conocimiento de los niveles de vida de Ecuador.

- **Limpieza y transformación de datos**

Los datos que se obtuvieron se los procedió a abrir desde el programa estadísticos SPSS, exportándolos a un archivo excel y siguiendo el procedimiento abajo detallado para la obtención final de los datos listos para ser usados.

1. Reemplazo de campos faltantes por el valor 0.
2. Importación del archivo Excel depurado a una tabla Access.
3. Exportación de la tabla Access a un archivo de texto.
4. Cambio de extensión del archivo de texto (*.txt) a un archivo delimitado por comas (*.csv).

- **Extracción de datos**

Para la extracción de los datos del archivo *.CSV a la base de datos (Oracle) se procedió a crear un archivo *.CTL para automatizar la población de los datos en las respectivas tablas.

A continuación se detalla un ejemplo de un archivo (*.ctl):

Ejemplo archivo vivienda.ctl :

```
load data
infile 'c:\vivienda.csv'
append
preserve blanks
  into table VIVIENDA
fields terminated by ',' optionally enclosed by '"'
TRAILING NULLCOLS
(ID_VIVIENDA, DIRECCION, PROVINCIA, CIUDAD, SECTOR,
BARRIO,
ID_PROPIETARIO,
ID_TIPO_VIVIENDA,
ID_PAREDES,
ID_TECHO,
ID_PISO,
ID_TENENCIA,
ID_AGUA_BEBER,
ID_SANITARIO,
ID_BASURA,
ID_COMBUSTIBLE,
TOTAL_CASHPOR,
CONDICION)
```

Preparación de datos para la aplicación de los algoritmos de minería de datos

La preparación de datos se refiere a ciertas transformaciones requeridas por el algoritmo de minería de datos, antes de que sea invocado. La preparación de datos puede tomar muchas formas, tales como: unir dos o más tablas que son necesarios en una sola tabla o vista, transformar atributos numéricos a categóricos, etc. [8]

Oracle Data Mining ejecuta las siguientes transformaciones:

Discretización: Se aplica a atributos categóricos, para reducir la cantidad de valores distintos de una variable o atributo, por medio de la agrupación de valores relativos o cercanos. Es aconsejable usar discretización para aquellos atributos categóricos que poseen gran cantidad de posibles valores, por ejemplo: "País".

Algoritmos como "Adaptive Bayes Network" y "Naive Bayes" se benefician de esta transformación. [8]

Normalización: Se aplica a atributos numéricos, convierte los valores numéricos de los atributos a un rango común de (0-1) - (1-0). Es aconsejable utilizar normalización cuando se tiene valores continuos como por ejemplo: las calificaciones.

Algoritmos como "K-Means" y "Support Vector Machine" se benefician de esta transformación. [8]

Se detalla a continuación el proceso de preparación de datos para cada uno de los algoritmos usados en el sistema "SAMBP":

Algoritmo "Adaptive Bayes Network" (ABN)

Se procede a realizar un análisis de cada atributo para establecer que tipo de discretización se aplica. En este caso se concluyó, no discretizar ninguno de los 9 atributos de la vivienda, debido a que a pesar que son de tipo categóricos, todos poseen poca cantidad de posibles valores. (Ver tabla 2)

Adicionalmente, los atributos "ID_VIVIENDA" Y "CONDICION", son también excluidos de la discretización; el primero, por ser la clave principal de la tabla donde se encuentra los valores ("VIVIENDA") y el segundo, por ser el atributo a predecir.

NOMBRE COLUMNA	TIPO DE DATO	DISTINCT	EXCLUIDA	TIPO DISCRETIZACIÓN
ID_VIVIENDA	NUMBER	6363	SI	---
ID_TIPO_VIVIENDA	VARCHAR2	7	SI	---
ID_PAREDES	VARCHAR2	8	SI	---
ID_TECHO	VARCHAR2	8	SI	---
ID_PISO	VARCHAR2	6	SI	---
ID_TENENCIA	VARCHAR2	6	SI	---
ID_AGUA_BEBER	VARCHAR2	8	SI	---
ID_SANITARIO	VARCHAR2	5	SI	---
ID_BASURA	VARCHAR2	9	SI	---
ID_COMBUSTIBLE	VARCHAR2	6	SI	---
CONDICION	VARCHAR2	2	SI	---

Tabla 2. Análisis de atributos para algoritmo ABN

Algoritmo "Enhanced K-Means"

En la preparación de los datos para el algoritmo "Enhanced K-Means", no se procede a normalizar los datos debido a que todos son atributos categóricos. Y normalmente en este algoritmo se normalizan preferiblemente los atributos numéricos.

Por lo tanto, se excluyen de normalización a los siguientes atributos:

- "ID_VIVIENDA",
- "ID_TIPO_VIVIENDA",
- "ID_PAREDES",
- "ID_TECHO",
- "ID_PISO",
- "ID_TENENCIA",
- "ID_AGUA_BEBER",
- "ID_SANITARIO",
- "ID_BASURA",
- "ID_COMBUSTIBLE",
- "CONDICION"

Debido a que el algoritmo "Enhanced K-means" utiliza distancias numéricas como referencia de agrupamiento, si los datos iniciales son categóricos, como en este caso, procede a convertir los atributos categóricos a una colección de atributos binados. El número de los atributos binados es igual al número de valores distintos de cada

atributo categórico. Después que todos los atributos son manejados como números, en los cálculos de las iteraciones de k-means. Cuando se reporta el centroide de un atributo categórico. El valor categórico asociado con el atributo binado y con el valor más grande con respeto al centroide, es mostrado. Este es el valor de la moda para atributos categóricos de registros asignados a clusters (Ver detalles en Anexos B)

2.3. Explorar los datos

Para el sistema "SAMBP" la fase de "exploración de los datos" fue incluida en el paso anterior, concluyéndose que para "ABN" y para "Enahnced K-Means", no se aplicaron método de discretización y normalización, dado el tipo de datos que la muestra inicial contiene.

2.4 Generar modelos

2.4.1. Técnicas de mineria de datos

Técnica de mineria de datos supervisada o predictiva usada en el sistema "SAMBP"

- **Clasificación:** La clasificación de una colección de elementos consiste en dividirlos dentro de categorías o clases. En el contexto de la minería de datos, la clasificación es hecha usando un modelo que es construido sobre datos históricos. [8]

La meta de la clasificación predictiva o supervisada es exactamente predecir la clase objetivo para cada registro de datos nuevos, no históricos. [8]

Los algoritmos de ODM utilizados para la técnica de clasificación son:

- **Algoritmo "Decision Tree":** Las reglas de árbol de decisión proporcionan un modelo transparente para que un usuario o analista comercial pueda entender las bases de las predicciones del modelo. [8]

Además de la transparencia, el algoritmo de Árbol de Decisión provee velocidad y escalabilidad. Es decir, escala linealmente con el número de atributos predictores, en el orden de nlog (n) con el número de filas, n. [8]

Los áboles de decisión son útiles para detallar perfiles, como por ejemplo: el mejor cliente, factores asociados con fraudes, etc.

- **Algoritmo "Naive Bayes" (NB):** Naive Bayes es una técnica de clasificación y predicción que construye modelos que predicen la probabilidad de posibles resultados. Este algoritmo predice resultados binarios o multiclase. En los problemas binarios, cada registro cumplirá o no el comportamiento modelado. Por ejemplo, se puede construir un modelo para averiguar si un cliente será fiel o cambiará de proveedor. Naive Bayes puede hacer predicciones para problemas multiclase, en los cuales hay varios resultados posibles. Por ejemplo, se puede construir un modelo para predecir qué clase de servicio prefiere cada cliente. [8]

  NB hace predicciones usando el teorema de Bayes, el cual deriva la probabilidad de una predicción de la evidencia subyacente. El Teorema de Bayes indica que la probabilidad de ocurrir el acontecimiento 'A' dado que ha ocurrido el acontecimiento 'B' (P (A|B)) es proporcional a la probabilidad del acontecimiento 'B' dado que ha ocurrido el acontecimiento 'A' multiplicado por la probabilidad de ocurrir del acontecimiento 'A' ((P (B|A) P (A)). [8]

- **Algoritmo "Support Vector Machine" (SVM):** El algoritmo "Support Vector Machine" es un método general para la resolución de problemas de clasificación, regresión y estimación. Problemas del tipo: Encontrar una función y=f(x), dado un conjunto de patrones entrada-salida, (X1, Y1),... (Xn, Yn). [8]

La idea del método consiste en que un problema pueda solucionarse linealmente; trabajando, por ejemplo, con un núcleo gaussiano, un SVM proporciona el número de funciones, sus centroides y sus pesos de forma simultánea. [8]

Los SVMs son particularmente aplicados para descubrir patrones ocultos en problemas que tienen gran cantidad de atributos independientes. Ej. Predecir tratamientos de enfermedades basados en perfiles genéticos.

Algoritmo de clasificación usado en el sistema "SAMBP"

- **Algoritmo "Adaptive Bayes Network" (ABN):** Es un algoritmo propietario de Oracle, que provee una rápida y

escalable extracción de información predictiva de los datos con respecto a la clase o atributo a predecir. [8]

"Adaptive Bayes Network", construye clasificadores bayesianos usando MDL (Minimun Description Lenght) para minimizar el error del valor de predicción y para alinear los predictores. El algoritmo ABN posee diferentes modos, por medio de los cuales puede describir su modelo: Pruned Naive Bayes (Naive Bayes Build), Boosted (Multi Feature Build) y Simplified decisión tree (Single Feature Build); este último se utilizó en el sistema, es un equivalente simplificado del algoritmo "Árbol de Decisión C4.5" (Ver Fig.3). [8]

```
Función C4.5
        (R: conjunto de atributos no clasificadores,
        C: atributo clasificador,
        S: conjunto de entrenamiento) devuelve un árbol de decisión;
Comienzo
        Si S está vacío,
                Devolver un único nodo con Valor Falla;
        Si todos los registros de S tienen el mismo valor para el atributo clasificador,
                Devolver un único nodo con dicho valor,
        Si R está vacío,
                Devolver un único nodo con el valor más frecuente del atributo
                clasificador en los registros de S [Nota: habrá errores, es decir,
                registros que no estarán bien clasificados en este caso];
        Si R no está vacío,
                D <- atributo con mayor Proporción de Ganancia(D,S) entre los
                atributos de R;
                Sean {dj | j=1,2, ..., m} los valores del atributo D;
                Sean {Sj | j=1,2, ..., m} los subconjuntos de S correspondientes a los
                valores de dj
                respectivamente;
                Devolver un árbol con la raíz nombrada como D y con los arcos
                nombrados d1, d2,...,dm, que van respectivamente a los árboles
                C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);
Fin
```

Figura 3. Algoritmo C4.5

El C4.5 es una extensión del algoritmo ID3 (Ver Figura 4). El C4.5 construye un árbol de decisión mediante el algoritmo "divide y reinarás" y evalúa la información en cada caso utilizando los criterios de entropía - la entropía se utiliza para encontrar el parámetro más significativo en la caracterización de un clasificador - y ganancia o proporción de ganancia - se selecciona al atributo con mayor ganancia de entropía en cada iteración- según sea el caso. (Ver Figura 5) [13] [14]

```
Función ID3
    (R: conjunto de atributos no clasificadores,
    C: atributo clasificador,
    S: conjunto de entrenamiento) devuelve un árbol de decisión;
Comienzo
    Si S esta vacio,
        Devolver un único nodo con Valor Falla;
    Si todos los registros de S tienen el mismo valor para el atributo clasificador,
        Devolver un único nodo con dicho valor;
    Si R está vacio,
        Devolver un único nodo con el valor más frecuente del atributo
        clasificador en los registros de S [Nota: habrá errores, es decir, registros
        que no estarán bien clasificados en este caso];
    Si R no esta vacio,
        D ← atributo con mayor Ganancia (D,S) entre los atributos de R;
        Sean {dj | j=1,2,..., m} los valores del atributo D;
        Sean {Sj | j=1,2,..., m} los subconjuntos de S correspondientes a los
        valores de dj respectivamente;
        Devolver un árbol con la raiz nombrada como D y con los arcos
        nombrados d1, d2, ..., dm que van respectivamente a los árboles
        ID3(R-{D}, C, S1), ID3(R-{D}, C, S2), ... ID3(R-{D}, C, Sm);
Fin
```

Figura 4. Algoritmo ID3

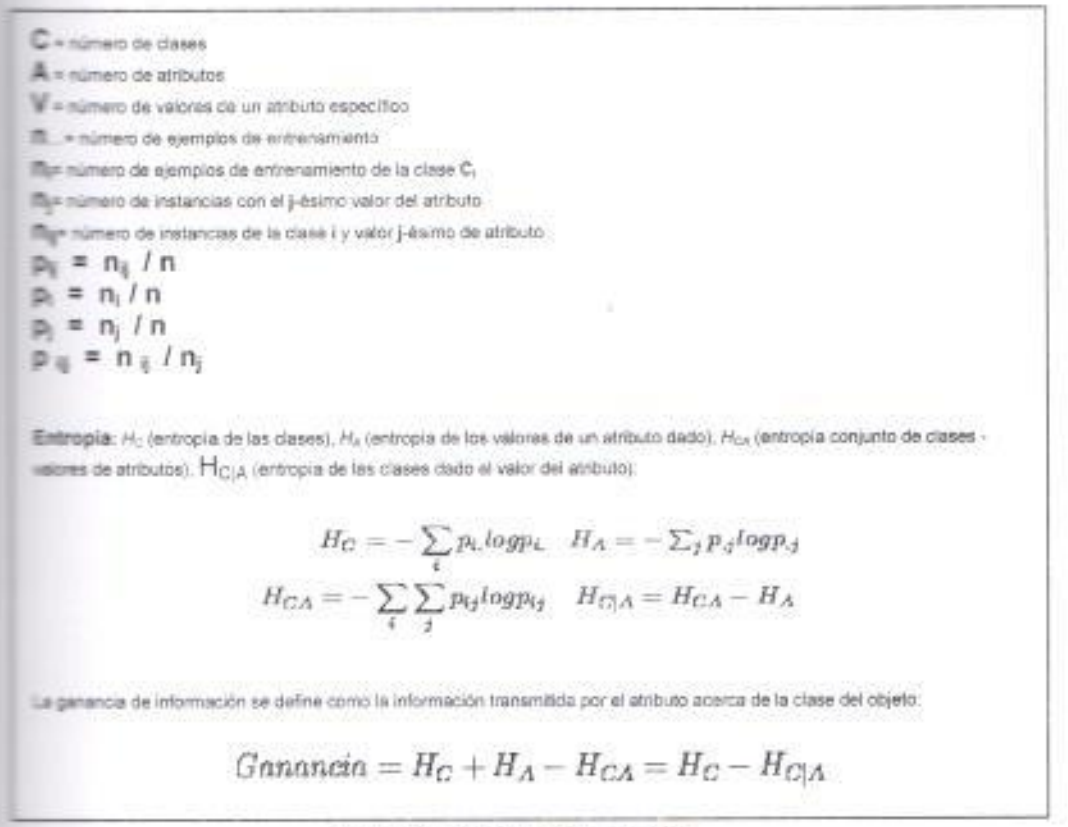


Figura 5. Fórmula de Entropía

El algoritmo C4.5, varía del algoritmo ID3, en la manera en que realiza las pruebas sobre las variables. En cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos. [13]

Los tres tipos de pruebas posibles propuestas por el C4.5 son

1. La prueba "estándar" para las variables discretas, con un resultado y una rama para cada valor posible de la variable. [13]

2. Una prueba más compleja, basada en una variable discreta, en donde los valores posibles son asignados a un número variable de

grupos con un resultado posible para cada grupo, en lugar de para cada valor. [13]

3. Si una variable A tiene valores numéricos continuos, se realiza una prueba binaria con resultados A <= Z y A > Z, para lo cual debe determinarse el valor límite Z. [13]

Todas estas pruebas se evalúan de la misma manera, mirando el resultado de la proporción de ganancia, o alternativamente, el de la ganancia resultante de la división que producen. Ha sido útil agregar una restricción adicional: para cualquier división, al menos dos de los subconjuntos Ti deben contener un número razonable de casos. Esta restricción, que evita las subdivisiones casi triviales, es tenida en cuenta solamente cuando el conjunto T es pequeño. [13]

Solamente en este modo, el modelo ABN extrae reglas humanamente legibles.[10] Por lo tanto, las reglas producidas por ABN es una de sus principales ventajas sobre "Naive Bayes". Ya que provee de un modelo transparente de reglas que pueden ser entendidas por analistas de mercado o de negocio. ABN predice valores binarios[11] y multiclases[12], así mismo usa costos y probabilidades para construir y clasificar. [8]

---

[10] Reglas lógicas (IF / THEN), que proveen un rápido entendimiento e interpretación al observador.

[11] Valores binarios indican decisiones de SI/NO (comprar/no comprar, cobrar/no cobrar, etc.)

[12] Valores multiclase indican una alternativa preferida como: color de suéter, rango de sueldo)

Aplicación del algoritmo ABN en "SAMBP"

La aplicación del algoritmo ABN en el sistema "SAMBP" construye el modelo llamado: "abnmodel". Este modelo genera reglas/patrones de predicción con su respectivo porcentaje de confianza y de cobertura como por ejemplo:

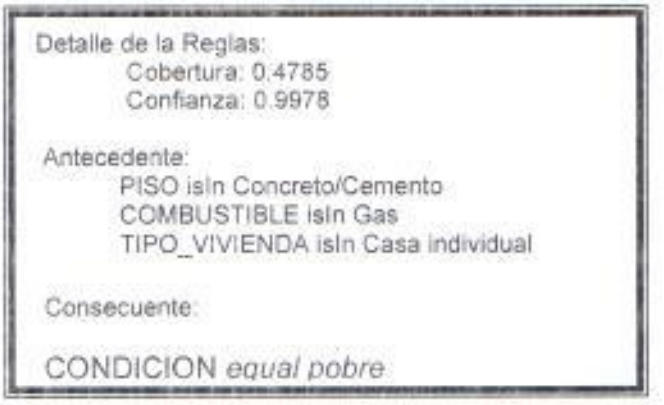


Figura 6. Patrones obtenidos algoritmo "ABN"

Técnica de minería de datos no supervisada o descriptiva usada en sistema "SAMBP"

- **Clusterización:** Es útil para la exploración de datos. Si hay muchos casos y agrupaciones naturales no obvias, el algoritmo de clusterización puede ser usado para encontrar agrupamientos naturales. El análisis de clusterización identifica grupos embebidos en los datos. Un cluster es una colección de datos que son similares en algún sentido para otro. La

clusterización puede también servir como un paso útil de pre-procesamiento de datos para identificar grupos homogéneos sobre los cuales construir modelos supervisados. [8]

En Oracle Data Mining, un cluster es caracterizado por su centroide, histogramas de atributos y la colocación de los clusters en el modelo de árboles jerárquicos.

Oracle Data Mining ejecuta clusterización jerárquica usando una versión mejorada del algoritmo K-Means. Los cluster encontrados por este algoritmo son usados para crear reglas que capturan las principales características de los datos asignados a cada cluster. [8]

Los algoritmos de ODM utilizados para la técnica de clusterización son:

- **Algoritmo O-Cluster:** Es un algoritmo de particionamiento basado en cuadrículas, es decir, divide el espacio en un número finito de celdas, formando una cuadrícula. El algoritmo crea particiones ortogonales paralelas al eje en el espacio. Los cluster resultantes son descritos por intervalos a lo largo de los ejes atributo y el correspondiente centroide e histograma. Solo las áreas con mayor densidad sobre el nivel de sensibilidad puede ser identificado como

cluster. O-cluster determina automáticamente el numero de clusters. [8]

Algoritmo de clusterización usado en el sistema "SAMBP"

- **Algoritmo "Enhanced K-Means"**

El algoritmo de K-Means es un algoritmo de agrupamiento/clusterización basado en distancias, que particiona los datos dentro de números predeterminados de grupos/clusters. El algoritmo de K-Means trabaja sobre distancias métricas tales como: Euclidiana, Mahalanobis, y Coseno, para medir la similaridad entre puntos de datos. Luego los puntos de datos son asignados al grupo/clúster más cercano de acuerdo a la distancia métrica usada. [8]

Oracle Data Mining implementa una versión mejorada del algoritmo K-Means con las siguientes características:

- El algoritmo Enhanced K-Means es un algoritmo jerárquico de tipo divisivo o también llamado "top-down", es decir, crea una descomposición jerárquica de un conjunto de datos formando un dendograma – árbol que divide la base de datos recursivamente en conjuntos cada vez más pequeños, usando divisiones binarias y refinamiento de todos los nodos

después que convergen los hijos del nodo dividido. En este sentido, el algoritmo es similar al algoritmo "Bisecting K-Means" [8]

> "Bisecting K-Means" es un algoritmo jerárquico de tipo divisivo. Es decir, se parte de la raiz, que es un solo grupo conteniendo a todos los elementos y se va haciendo divisiones paulatinas hasta llegar a las hojas que representa a la situación que en cada ejemplo es un grupo. Usa distancias Euclideana como métrica. (Ver Figura 7). [15]

C_L = elemento cualquiera

w = promedio de la muestra

C_R = distancia calculado con C_L y w

M= muestra

Bisecting K-Means

Paso 1. (Inicialización). Aleatoriamente seleccionar un punto, decir $C_L \in R^*$; luego calcular el centroide w de M, ver (1), y calcular $C_R \in R^*$ donde $C_R = w - (C_L - w)$.

Paso 2. Dividir $M = [X_1, X_2, \ldots\ldots Xn]$ dentro de dos subgrupos M_L y M_R respectivamente, de acuerdo a la siguiente regla:

$Xn \in M_L$ si $\| Xn - C_L \| \leq \| Xn - C_R \|$

$Xn \in M_R$ si $\| Xn - C_L \| > \| Xn - C_R \|$

Paso 3. Calcular los centroides de M_L y M_R, w_L y w_R, como en (2).

Paso 4. Si $w_L = C_L$ y $w_R = C_R$, parar, caso contrario $C_L = w_L$, $C_R = w_R$, ir al Paso 2.

(1) **Centroide de M.**

$$w = \frac{1}{N}\sum_{j=1}^{N} M_j$$

Donde Mj es la j-esima columna de M.

(2) **Centroides de los subclusters M_L y M_R**

$$w_L = \frac{1}{N_L}\sum_{j=1}^{N_L} M_{L,j} \qquad w_R = \frac{1}{N_R}\sum_{j=1}^{N_R} M_{R,j} =$$

Donde, $M_{L,j}$ y $M_{R,j}$ son las J-esima columnas de M_L y M_R, respectivamente.

Figura 7. Algoritmo Bisecting K-Means

- El algoritmo puede hacer crecer el árbol un nivel a la vez o un nodo a la vez, el nodo con mayor varianza (suma de las distancias euclidianas de cada elemento al centro del nodo) es dividido para incrementar el tamaño del árbol hasta que el número de grupos/clusters óptimo sea alcanzado. "Enhanced K-Means" retorna para cada grupo/cluster, un centroide y una

regla describiendo las características que encierran la mayoría de los datos asignados al cluster. [8]

- Finalmente se recalca que la implementación del algoritmo usado en el sistema es: ODM-DBMS_DATA_MINING, para efectos de manejar atributos categóricos. [23]

Aplicación del algoritmo "Enhanced K-Means" en "SAMBP"

La aplicación del algoritmo "Enhanced K-Means" en el sistema "SAMBP" construye el modelo llamado: "kmModel_jdm". El cual genera grupos descriptivos, de acuerdo a la cantidad de clusters deseados, en este caso, se fijó en 5, los clusters deseados de acuerdo al análisis de "Elbow" (Detalle Anexos B). Así como también se generan, reglas descriptivas y porcentaje de confianza[13] y soporte[14] de cada grupo que dan lugar a la formación de un árbol jerárquico de información. Por ejemplo:

Detalles del modelo de clusterización:
Número de clusters: 9
Número de niveles del árbol: 5
Casos: 6363

[13] Confianza o Precisión: porcentaje de veces que la regla se cumple cuando se puede aplicar.
[14] Soporte o Cobertura: porcentaje de instancias que la regla predice correctamente.

Cluster ID	Soporte	Confianza
Cluster 2	0.8711060302943698	0.87110603029437
Cluster 4	0.834029227557411	0.834029227557411
Cluster 7	0.9171122994652406	0.9171122994652406
Cluster 8	0.779141104294786	0.779141104294479
Cluster 9	0.8346774193548387	0.834677419354839

Cluster 2	
Antecedente	TENENCIA isin Alquilada AND TENENCIA isin Hipotecada AND TENENCIA isin Propia PAREDES isin Bloque/Ladrillo COMBUSTIBLE isin Gas AGUA_BEBER isin Acueducto publico PISO isin Concreto/Cemento AND PISO isin Ladrillo SANITARIO isin Alcantarillado AND SANITARIO isin Letrina AND SANITARIO isin Tanque Septico TECHO isin Concreto/Cemento AND TECHO isin Metal BASURA isin Servicio municipal AND BASURA isin Servicio particular TIPO_VIVIENDA isin Apartamento AND TIPO_VIVIENDA isin Casa individual CONDICION isin pobre
Consecuente	Cluster equal 2.0

Tabla 3. Cluster generados "Enhanced K-Means"

2.4.1.1 Justificación de la técnica aplicada

Análisis descriptivo

Característica	Enhanced K-Means	O-Cluster
Método de Clusterización	Basado en distancia	Basado en partición ortogonal (grillas)
Tipo de Atributos	Categóricos y Numéricos	Numéricos
Número de casos	Maneja colección de datos de cualquier tamaño	Más apropiado para colección de datos que tienen mas de 500 casos
Número de atributos	Más apropiado para colección de datos con poco número de atributos	Más apropiado para colección de datos con alto número de atributos
Número de cluster	Especificado por el usuario	Automáticamente generado

Cluster jerárquico	SÍ	SÍ
Probabilidades	SÍ	SÍ
Preparación de data	Normalización	Normalización

Tabla 4. Justificación uso de algoritmo "Enhanced K-Means"

Análisis predictivo

Característica	Naïve Bayes	Adaptive Bayes Network	Support Vector Machine	Decision Tree
Velocidad	Muy rápido	Rápido	Rápido con aprendizaje activo	Rápido
Certeza	Bueno en muchos dominios	Bueno en muchos dominios	Significante	Bueno en muchos dominios
Transparencia	No reglas	Reglas solo para construcción de características simples	No reglas	Reglas
Interpretación de valores nulos	SÍ	SÍ	SÍ	SÍ

Tabla 5. Justificación de uso algoritmo ABN

Nota: Ambos cuadros comparativos fueron obtenido del documento: http://download-east.oracle.com/docs/cd/B19306_01/datamine.102/b14339.pdf.

2.5. Explorar y validar los modelos

Los modelos supervisados son analizados/validados para evaluar la certeza y el error de sus predicciones. Para ellos existen métricas tales como: matriz de confusión y opcionalmente el ROC.

- **Matriz de confusión**: Mide la certeza de las predicciones hechas por un modelo supervisado, así como también el porcentaje de error que el modelo produce cuando clasifica registros. Las filas indexadas de una matriz de confusión corresponden al *Valor Predicho*, los cuales son los valores de la variable clase. Las columnas indexadas corresponden al *Valor Actual*. Los números en cada celda corresponden al número de registros que fueron clasificados en esa coordenada, representados con la siguiente nomenclatura: TP, TN, FP, FN. Los valores TP y TN forman la diagonal de aciertos, tal como se muestra en la siguiente tabla:

		actual	
		pobre	muy pobre
predicho	pobre	TP	FP
	muy pobre	FN	TN

Diagonal de los aciertos

Tabla 6. Estructura Matriz de Confusión

Donde:

TP (True Positive): Casos positivos que han sido correctamente clasificados para el primer valor de la variable clase.

TN (True Negative): Casos positivos que han sido correctamente clasificados con respecto al segundo valor de la variable clase.

FP (False Positive): Casos negativos que han sido clasificados incorrectamente con respecto al primer valor de la variable clase.

FN (False Negative): Casos negativos que han sido clasificados incorrectamente con respecto al segundo valor.

- **ROC (Receiver Operating Characteristics)**: Analiza un modelo predictivo calculando y graficando la certeza positiva (TPR) y el error negativo (FPR) del modelo usando valores de la matriz de confusión. Para posteriormente graficar el espacio ROC y determinar si un modelo es Buen Clasificador o Mal Clasificador de acuerdo a los siguientes gráficos:

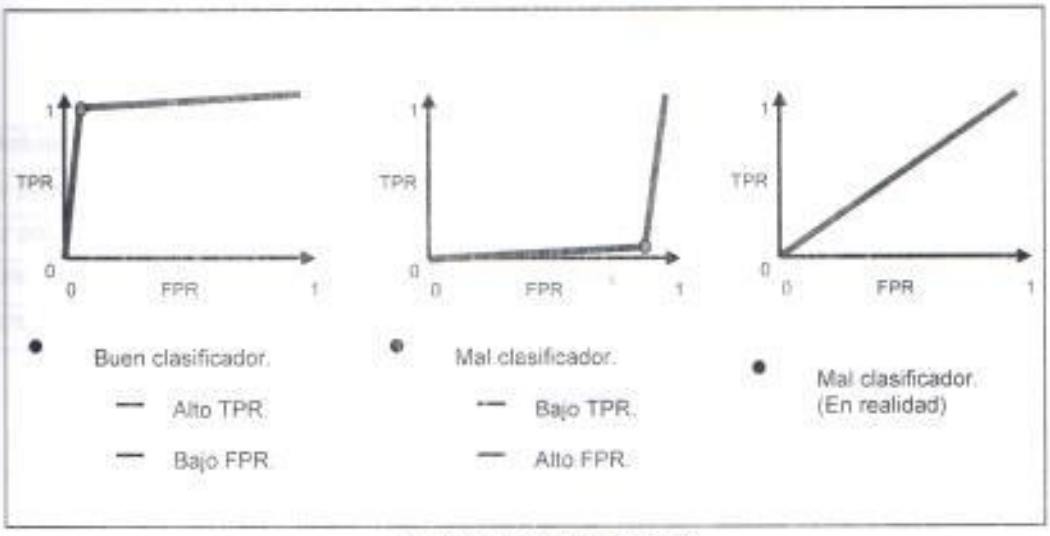


Figura 8. Gráficos ROC

Aplicación de métricas para la evaluación del modelo ABN en el sistema "SAMBP":

Se procedió a evaluar el modelo construido por el algoritmo ABN usando matriz de confusión y el ROC, dando los resultados que se muestran a continuación:

- **MATRIZ DE CONFUSION:**

Actual	Predicho	Valor
muy pobre	muy pobre	841
muy pobre	pobre	124
pobre	muy pobre	63
pobre	pobre	5335

		actual	
		pobre	muy pobre
predicho	pobre	5335	124
	muy pobre	63	841

Tabla 7. Matriz de confusión algoritmo ABN

Dichos resultados se interpretan de la siguiente forma:

- El valor de 63 para un valor actual indexado de "pobre" y un valor predictor de "muy pobre", indica que el modelo clasificó incorrectamente un "pobre" como "muy pobre" 63 veces. Y un valor de 5335 para un valor actual y un valor predictor de "pobre", indica que el modelo clasificó correctamente un "pobre" 5335 veces.

Los cálculos de porcentajes de certeza, error y precisión de acuerdo a los resultados de la Matriz de Confusión son los siguientes:

Total registros clasificados = TP + TN + FP + FN = 6363

Certeza = TP + TN / total registros clasificados

Certeza = 5335 + 841 / 6363 = 0.9706 = 97.06%

Error = FN + FP / total registros clasificados =
Error = 124 + 63 / 6363 = 0.0294 = 2.94%

Precisión = TP / (TP + FP) =
Precisión = 5335 / (5335 + 124) = 0.9773 = 97.73%

- **ROC:**

TPR = TP / TP + FN = 5335 / 5335 + 63 = 0.9883

FPR = FP / TN + FP = 124 / 841 + 124 = 0.1284

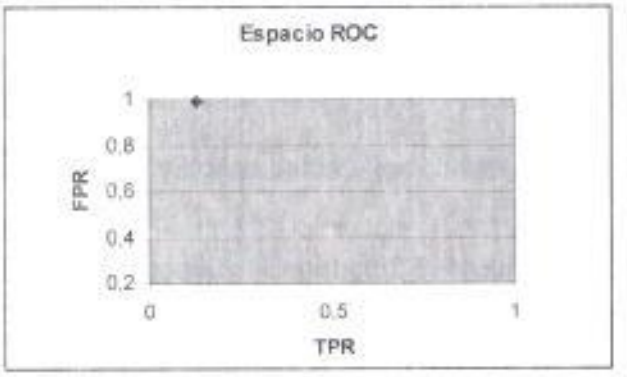


Figura 9. Resultados ROC

De acuerdo al grafico y comparando con el grafico de buen clasificador, anteriormente citado, nos podemos dar cuenta que nuestro modelo puede ser considerado un buen modelo predictivo, ya que el valor TPR es alto y el valor FPR es bajo, con lo cual se forma un punto que se acerca a uno,

denominándolo como Buen Clasificador, de acuerdo a la métrica del ROC.

Para el caso del **Algoritmo "Enhanced K-Means"** no se requiere analizar el modelo, debido a que es un algoritmo descriptivo.

2.6. Implementar y actualizar los modelos

Aplicación de modelo "Enhanced K-Means":

En los cluster descubiertos por "Enhanced K-Means" se generan probabilidades bayesianas; que luego son usadas durante la aplicación del modelo para asignar datos a los cluster.

Para la aplicación del modelo "Enhanced K-Means", se procede evaluar un nuevo registro según el modelo construido. Dando como resultado, información del siguiente tipo:

Grupo	Probabilidad
8	9160287210707583 9

Tabla 8. Análisis del modelo "Enhanced K-Means"

El resultado es interpretado de la siguiente forma:

El registro aplicado al modelo es asignado al Cluster 8 con una probabilidad del 91.6%. El cluster 8 (Ver Figura 10), luego de la construcción del modelo K-Means, detalla lo siguiente:

Figura 10: Detalle cluster 8 "Enhanced K-Means"

Aplicación de modelo "Adaptive Bayes Network" (ABN):

Para la aplicación del modelo ABN, se procede evaluar un nuevo registro:

tipo_ vivienda	paredes	techo	piso	tenencia	agua _ beber	sanitario	basura	combustible

choza o rancho	cana/paja/palos	paja/penca	otros	propia	Acueducto comunidad	letrina	La queman	leña

Dando como resultado, información del siguiente tipo:

PREDICCION	PROBABILIDAD
muy pobre	0.9998
pobre	0.0002

Tabla 9. Resultados aplicación modelo ABN

Lo cual es interpretado de la siguiente forma: El nuevo registro evaluado en el modelo posee una probabilidad del 0.02% de pobreza, y un 99.98% de pobreza extrema. Indicando que puede ser considerado como un candidato idóneo para la aprobación de microcrédito.

CAPÍTULO 3

3. ANÁLISIS Y DISEÑO DEL SISTEMA

3.1. Análisis del sistema

3.1.1. Casos de uso

Caso de uso 1:

Nombre: Extracción, transformación y limpieza de datos

Descripción: Permite cargar datos al sistema desde un archivo, con el fin de poblar la base de datos del sistema.

Notas: Solo permite cargar archivos con extensión *.csv (Excel). Por medio de un archivo de carga (*.ctl) donde se definen la tabla donde se van a cargar los datos así como los campos.

Caso de uso 2:

Nombre: Generar patrones

Descripción: Se refiere a aplicar clusterización (K-Means) y redes bayesianas (ABN) para generar patrones o reglas en donde se detallen características propias de cada grupo y luego, permitan la clasificación/aprobación de las personas para un microcrédito.

Notas: Se generan una sola vez, cuando se extrae los datos por primera vez.

Caso de uso 3:

Nombre: Clasificación de datos

Descripción: Se refiere a clasificar a las personas según los patrones/reglas extraídos, por medio del algoritmo ABN (Adaptive Bayes Network).

Notas: El ingreso de un nuevo registro para clasificarlo se lo realiza a través de un formulario.

Caso de uso 4:

Nombre: Consultas históricas

Descripción: Permite realizar consultas sobre el estado de las personas (aprobación o no aprobación de microcrédito).

Notas: Sólo se permitirá realizar consultas por nombre y apellido.

3.1.2. Diagrama de contexto

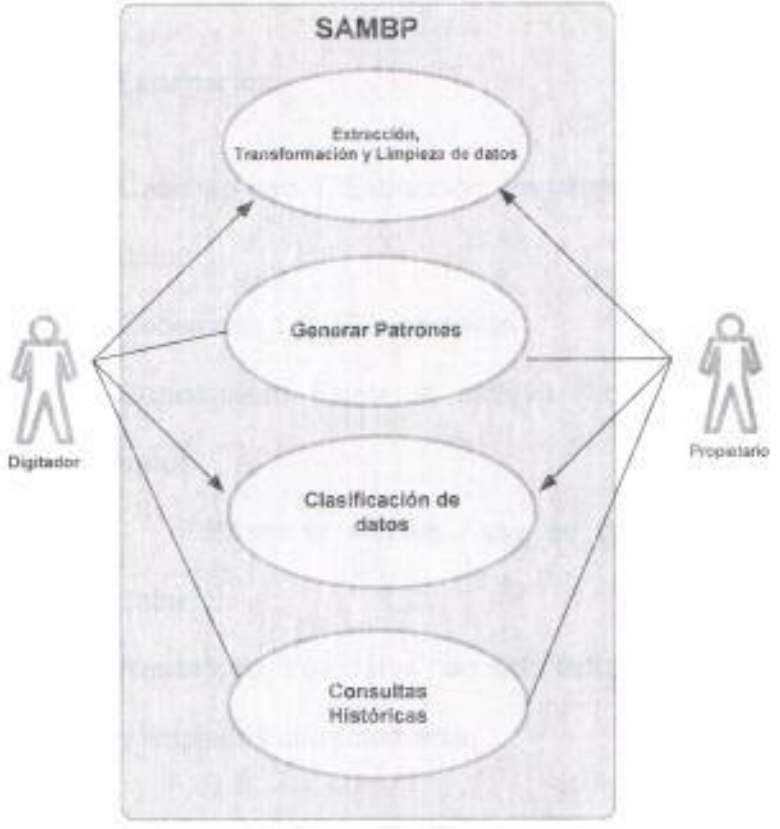


Figura 11. Diagrama de contexto del sistema "SAMBP"

Actores

- **Digitador:** Rol de la persona que interactúa directamente con el sistema, al ejecutar todas la operaciones.

- **Propietario:** Rol de la persona a quien se le va a realizar un análisis de las características de su vivienda para la aprobación de microcrédito.

3.1.3. Escenarios

Caso de uso 1: Extracción, transformación y limpieza de datos

Escenario 1.1: ETL correcto

Suposición: Existe el archivo (*.ctl) para cargar los datos.

- Existe el archivo *.csv de donde se extrae los datos.

Resultado: Los datos han sido extraídos, transformados y limpiados correctamente.

Escenario 1.2: ETL incorrecto

Suposición: Existe el archivo (*.ctl) para cargar los datos.

- No Existe el archivo *.csv de donde se extrae los datos.

Resultado: Cargada de datos no exitoso.

Caso de uso 2: Generar patrones

Escenario 2.1: Generación de patrones correctamente

Suposición: - El usuario realiza la acción de generar patrones por medio de un algoritmo predictivo (ABN) y descriptivo (K-Means).

- El algoritmo escogido realiza su procedimiento.

Resultado: Se generan los patrones de manera satisfactoria.

Escenario 2.2: Generación de patrones incorrectamente.

Suposición: - El usuario realiza la acción de generar patrones por medio de un algoritmo predictivo (ABN) y descriptivo (K-Means).

- El algoritmo escogido no realiza su procedimiento correctamente.

Resultado: Generación de patrones incorrectamente.

Caso de uso 3: Clasificación de datos

Escenario 3.1: Datos clasificados correctamente

Suposición: El usuario ingresa los datos del nuevo registro a clasificar.

- El usuario escoge aplicar el modelo generado por el algoritmo ABN.

Resultado: Dependiendo del resultado del modelo se establece el estado de aprobación de microcrédito.

Escenario 3.2: Datos clasificados incorrectamente

Suposición: El usuario ingresa los datos del nuevo registro a clasificar.

- El usuario escoge aplicar el modelo generado por el algoritmo ABN.

Resultado: "No se puede aplicar el modelo escogido a dichos datos".

Caso de uso 4: Consultas históricas

Escenario 4.1: Consulta de información exitosa

Suposición: - Se ingresa el nombre y apellido de la persona que se desea consultar correctamente.

Resultado: Se muestra información acerca de la persona (aprobación/no aprobación de microcrédito).

Escenario 4.2: Consulta de información no exitosa

Suposición: - Se ingresa el nombre y apellido de la persona que se desea consultar incorrectamente.

Resultado: - Mensaje del sistema "No existe dicha persona en el sistema".

3.1.4. Diagramas de análisis de interacción de objetos

Escenario 1.1: ETL correcto

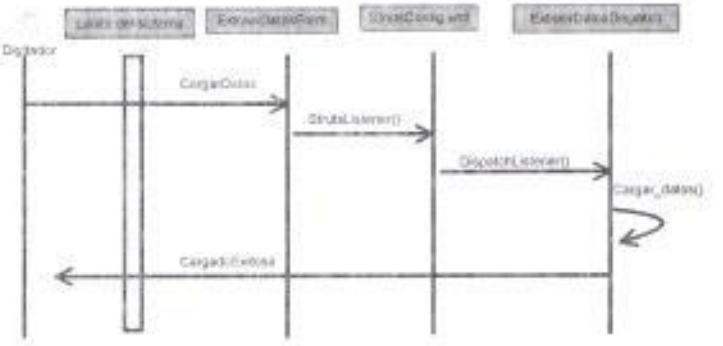


Figura 12. ETL correcto

Escenario 1.2: ETL incorrecto

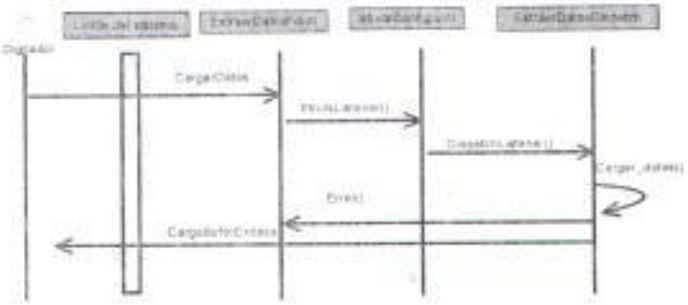


Figura 13. ETL incorrecto

Escenario 2.1: Generación de patrones correctamente

ABN

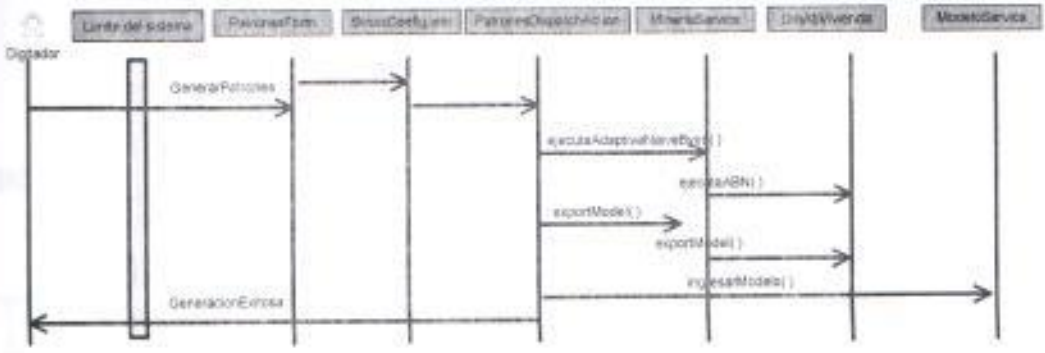


Figura 14. Generación de patrones predictivos correctamente ABN

K-means

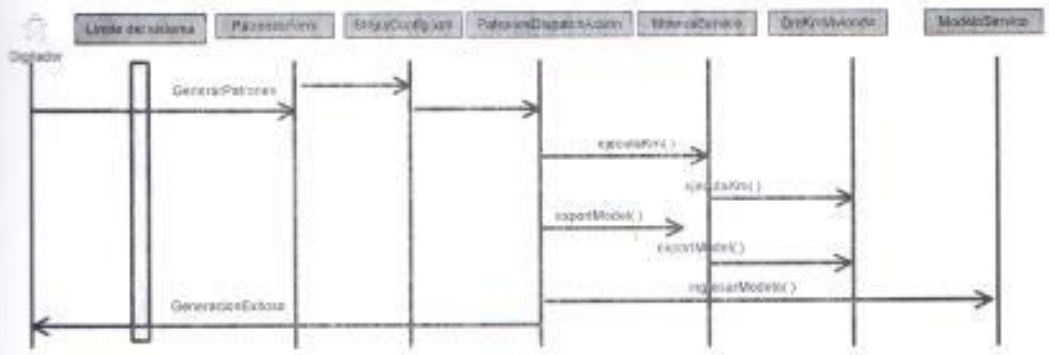


Figura 15. Generación de patrones descriptivos correctamente K-Means

Escenario 2.2: Generación de patrones incorrectamente

ABN

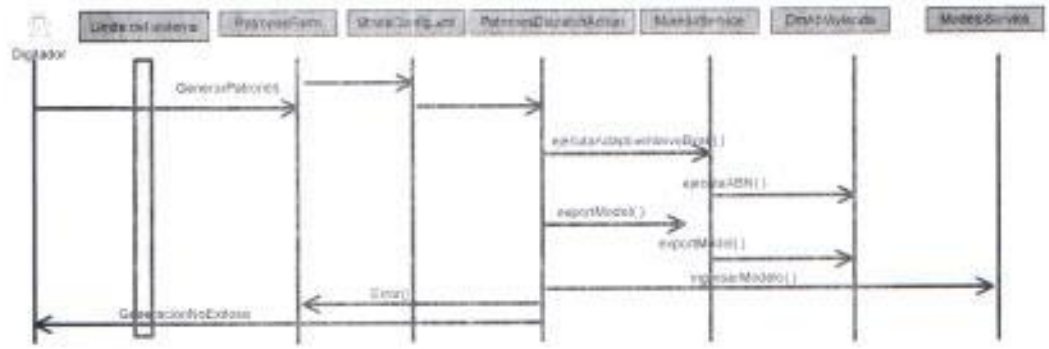


Figura 16. Generación de patrones predictivos incorrectamente ABN

K-means

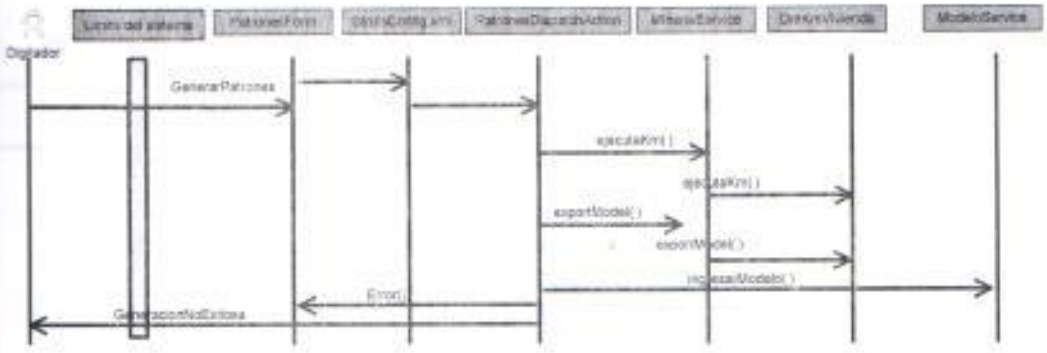


Figura 17. Generación de patrones descriptivos incorrectamente K-Means

Escenario 3.1: Datos clasificados correctamente

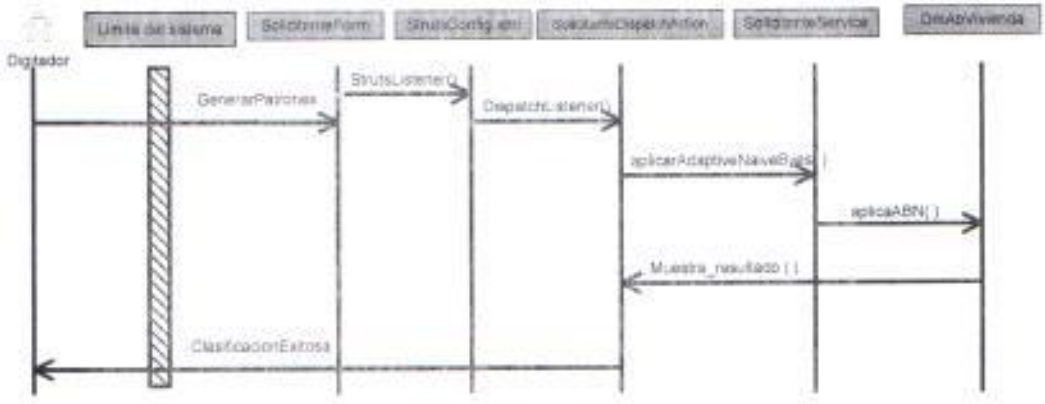


Figura 18. Clasificación correcta

Escenario 3.2: Datos clasificados incorrectamente

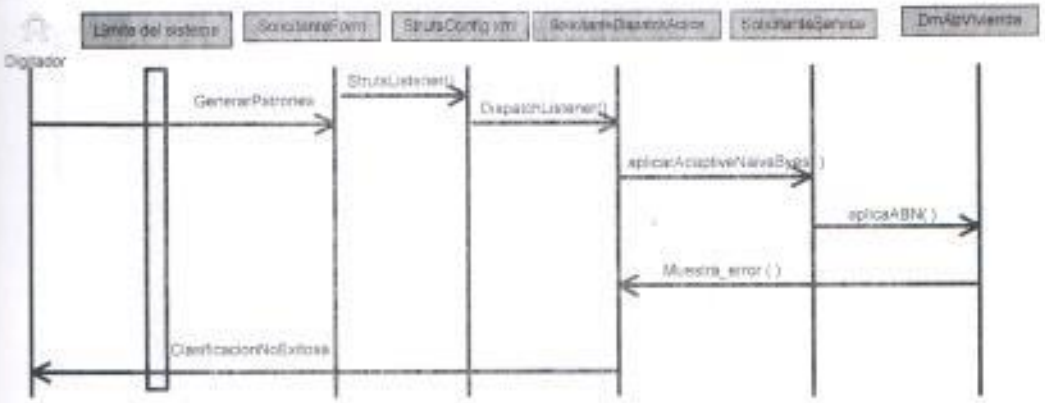


Figura 19. Clasificación incorrecta

Escenario 4.1: Consulta de información exitosa

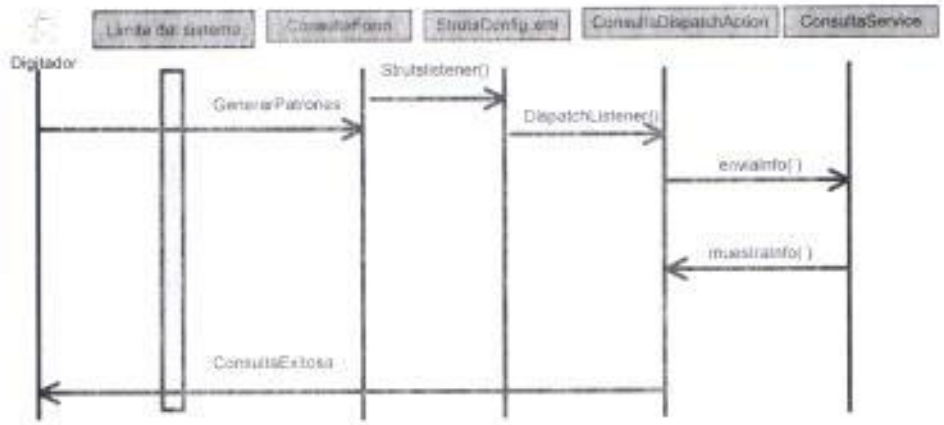


Figura 20. Consulta exitosa

Escenario 4.2: Consulta de información no exitosa

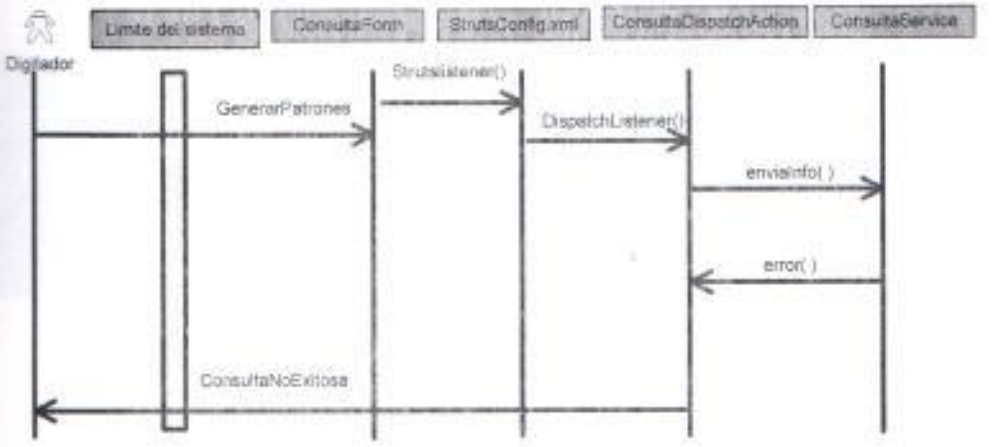


Figura 21. Consulta no exitosa

3.1.5. Modelo conceptual de la base de datos

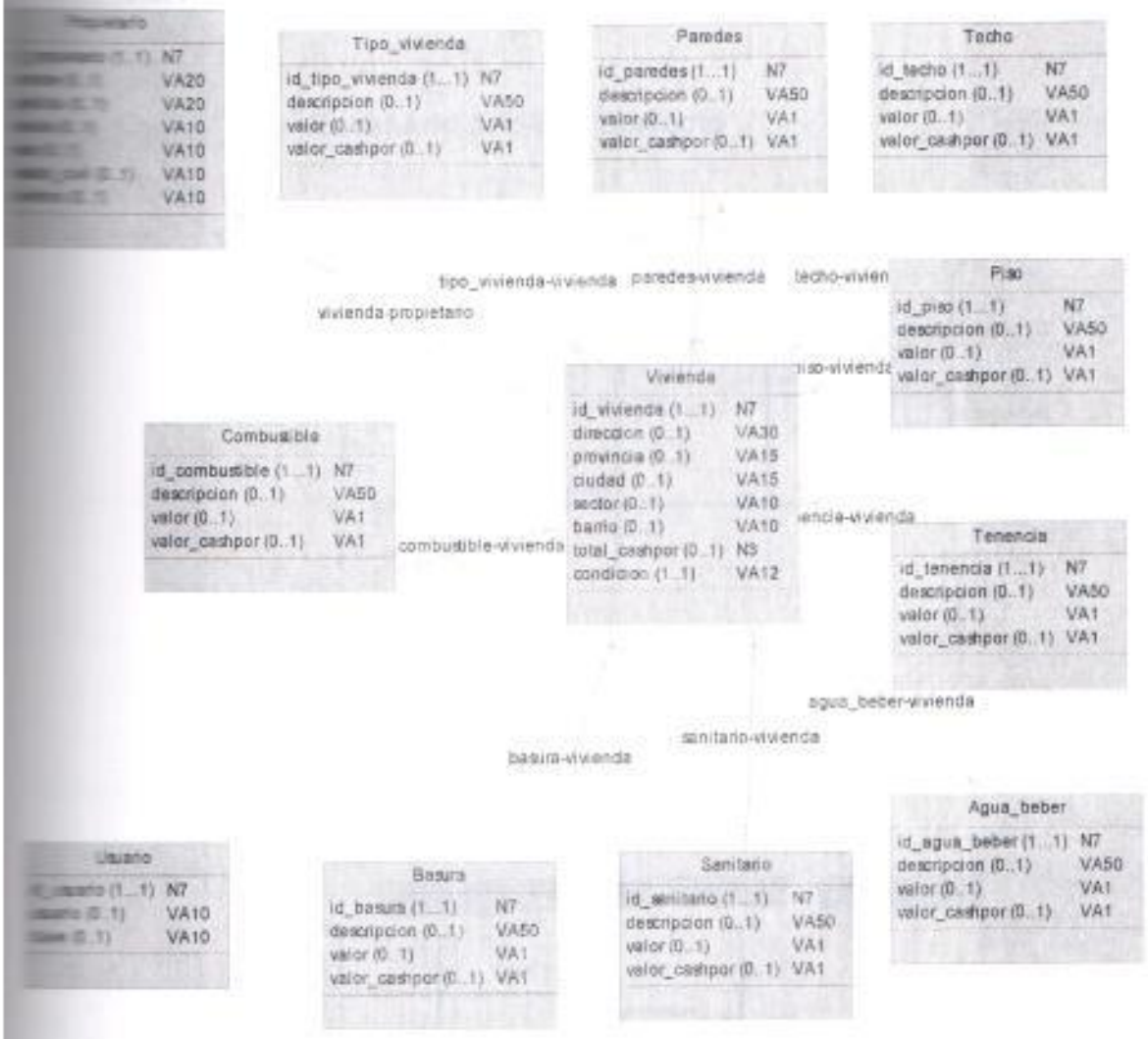


Figura 22. Modelo conceptual

3.2. Diseño del sistema

3.2.1. Diagramas de diseño de Interacción de Objetos

Escenario 1.1: ETL correcto

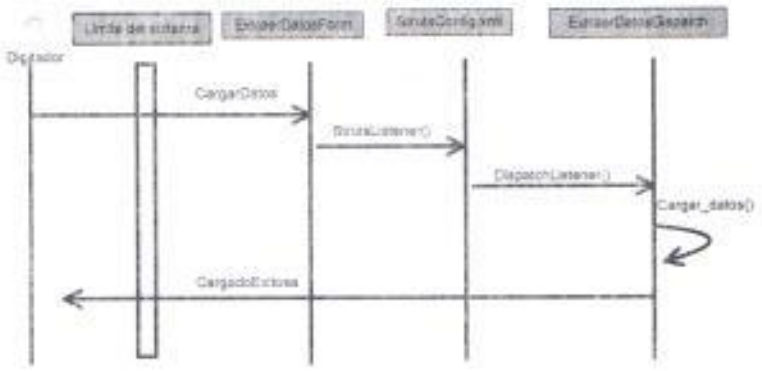


Figura 23. ETL correcto

Escenario 1.2: ETL incorrecto

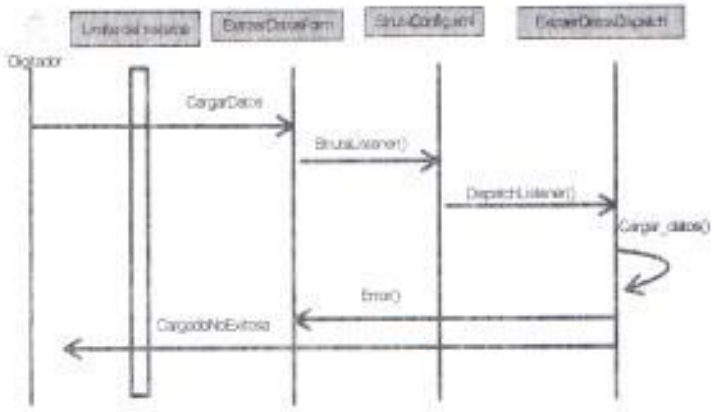


Figura 24. ETL incorrecto

Escenario 2.1: Generación de patrones correctamente

ABN

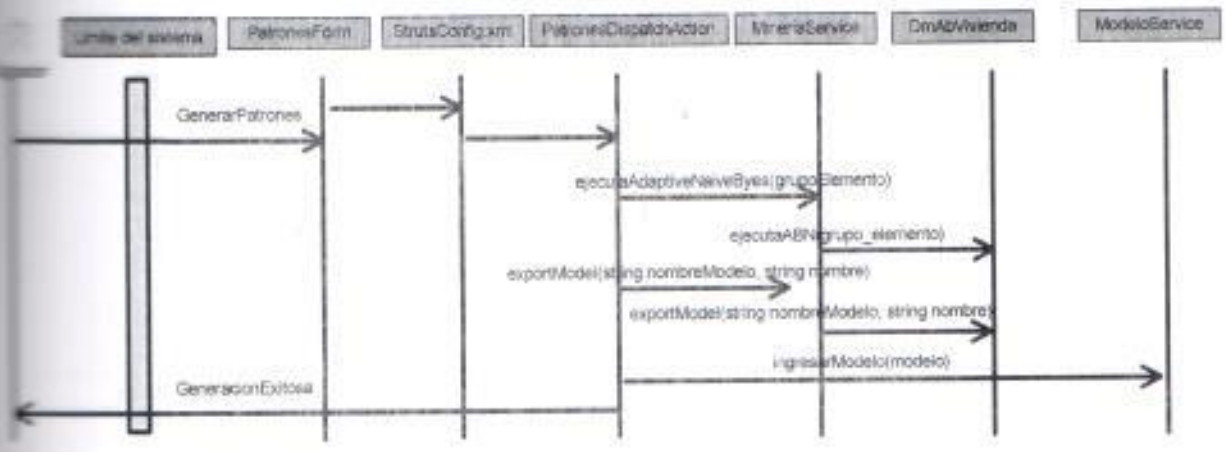


Figura 25. Generación de patrones predictivos correctamente ABN

K-means

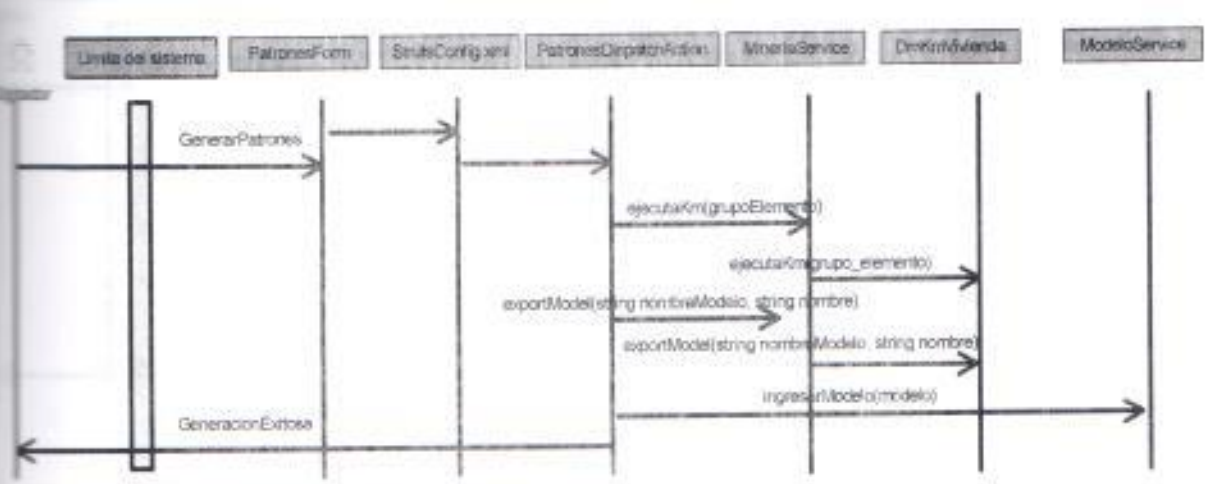


Figura 26. Generación de patrones descriptivos correctamente K-Means

Escenario 2.2: Generación de patrones incorrectamente

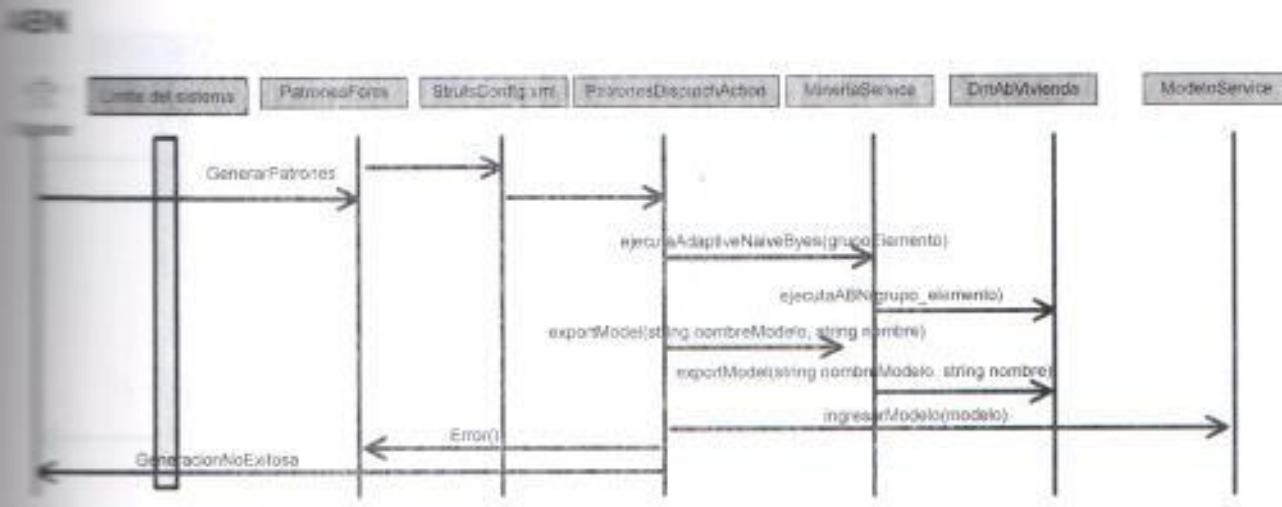


Figura 27. Generación de patrones predictivos incorrectamente ABN

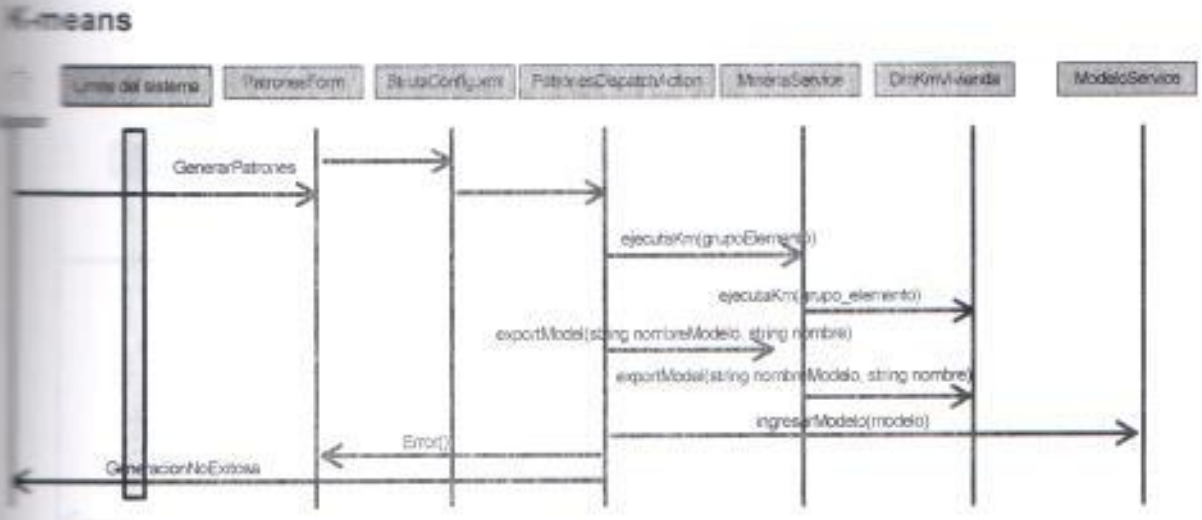


Figura 28. Generación de patrones descriptivos incorrectamente K-Means

Escenario 3.1: Datos clasificados correctamente

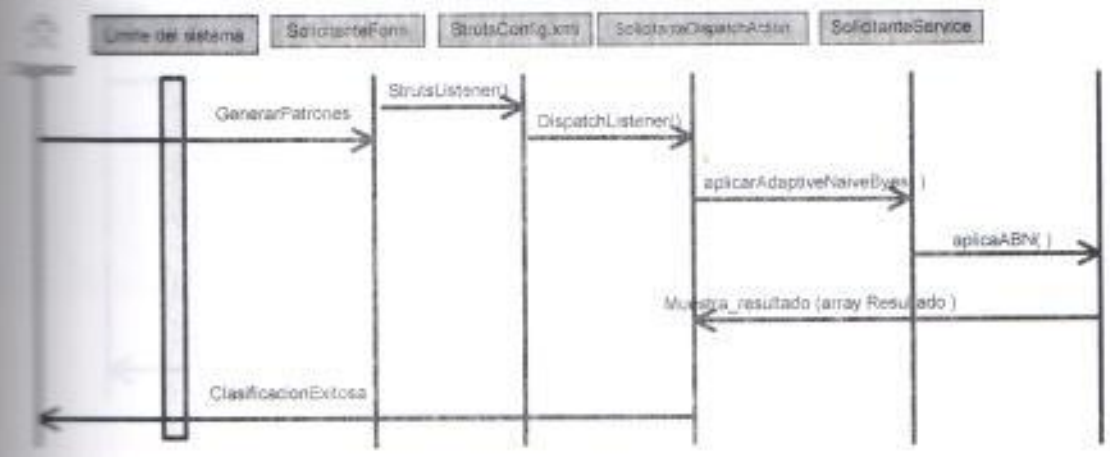


Figura 29. Clasificación correcta

Escenario 3.2: Datos clasificados incorrectamente

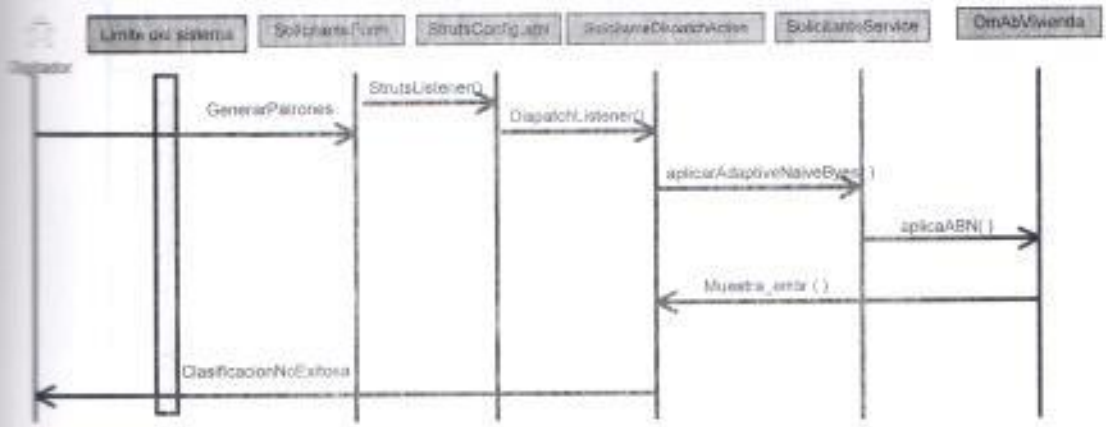


Figura 30. Clasificación incorrecta

Escenario 4.1: Consulta de información exitosa

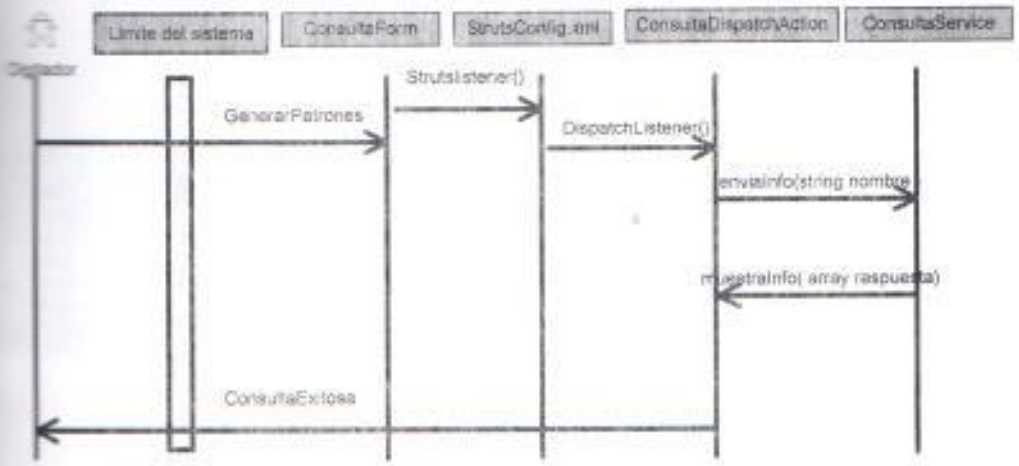


Figura 31. Consulta exitosa

Escenario 4.2: Consulta de información no exitosa

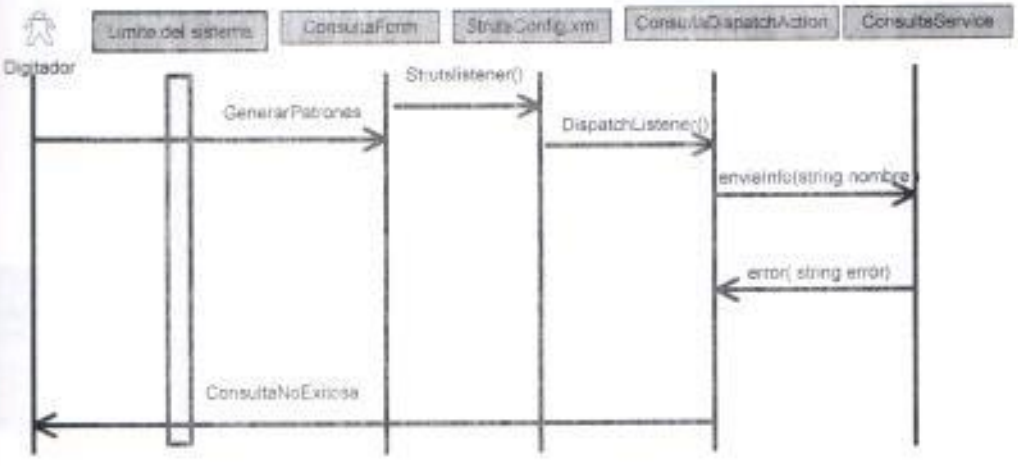


Figura 32. Consulta no exitosa

3.2.2. Modelo lógico de la base de datos

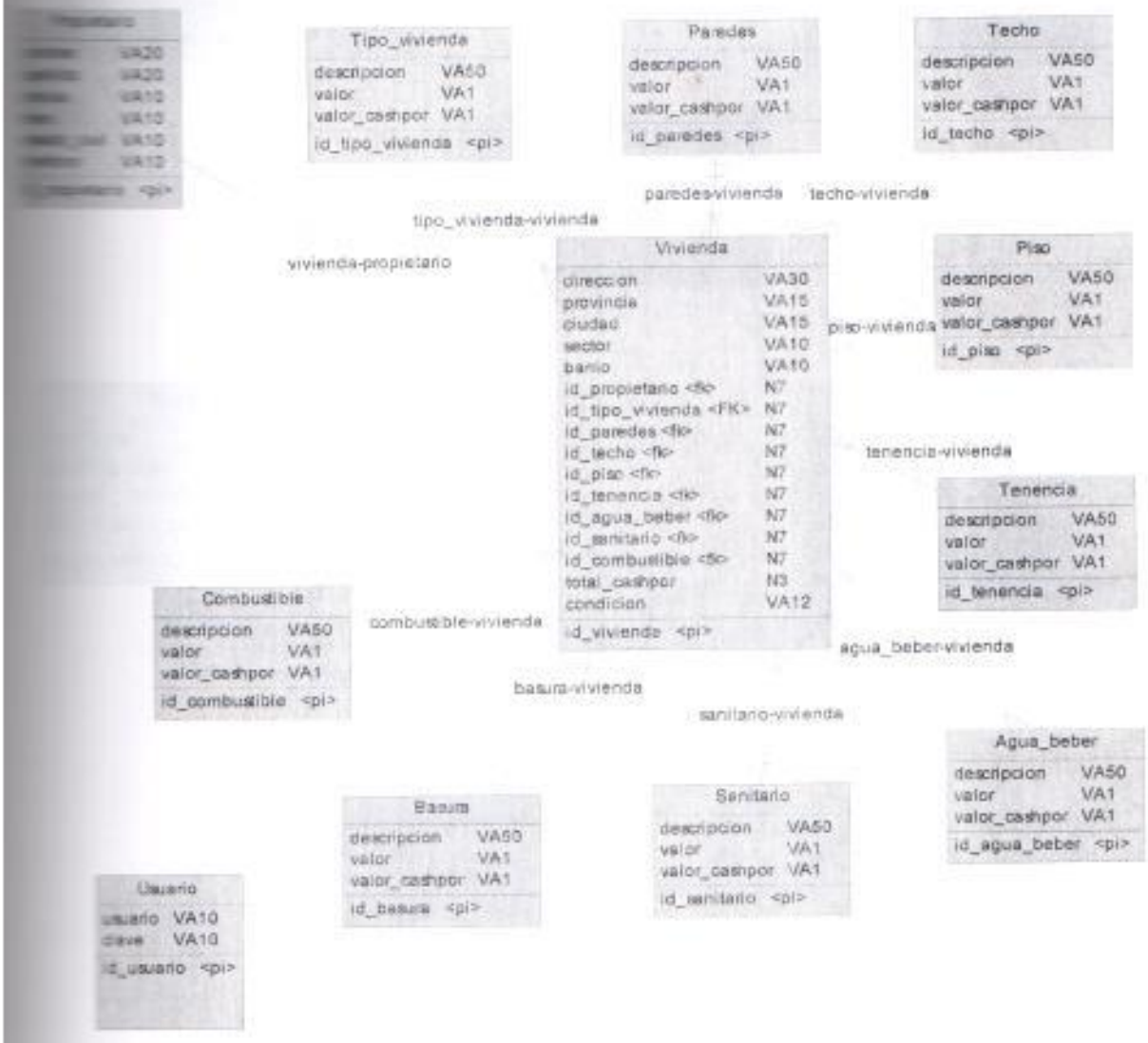


Figura 33. Modelo lógico

3.2.3. Modelo multidimensional

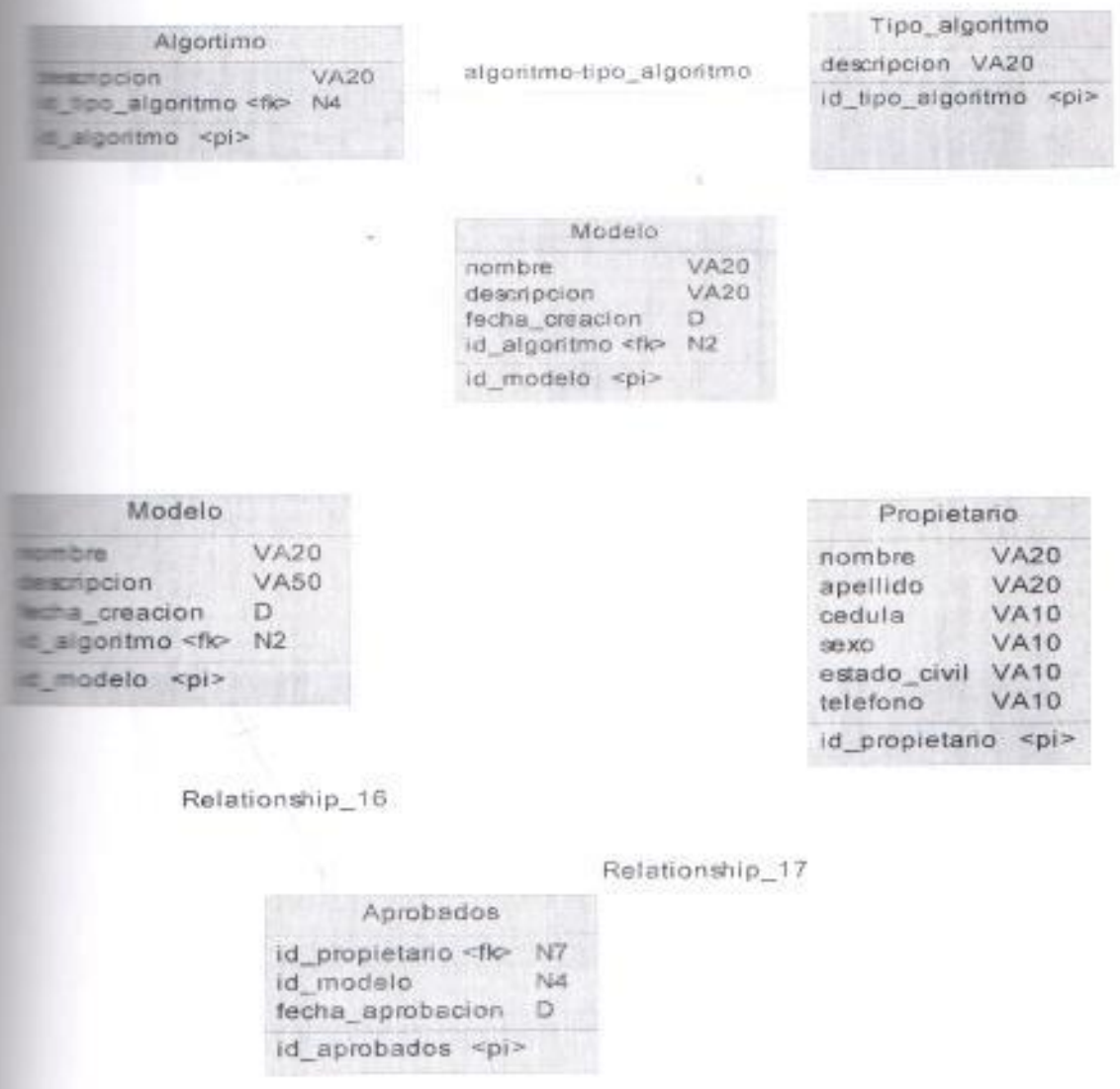


Figura 34. Modelo multidimensional de la base de da

3.2.4. Flujo de ventanas y layouts

Figura 35. Ingreso al sistema "SAMBP"

Figura 36. Menú Principal del sistema "SAMBP"

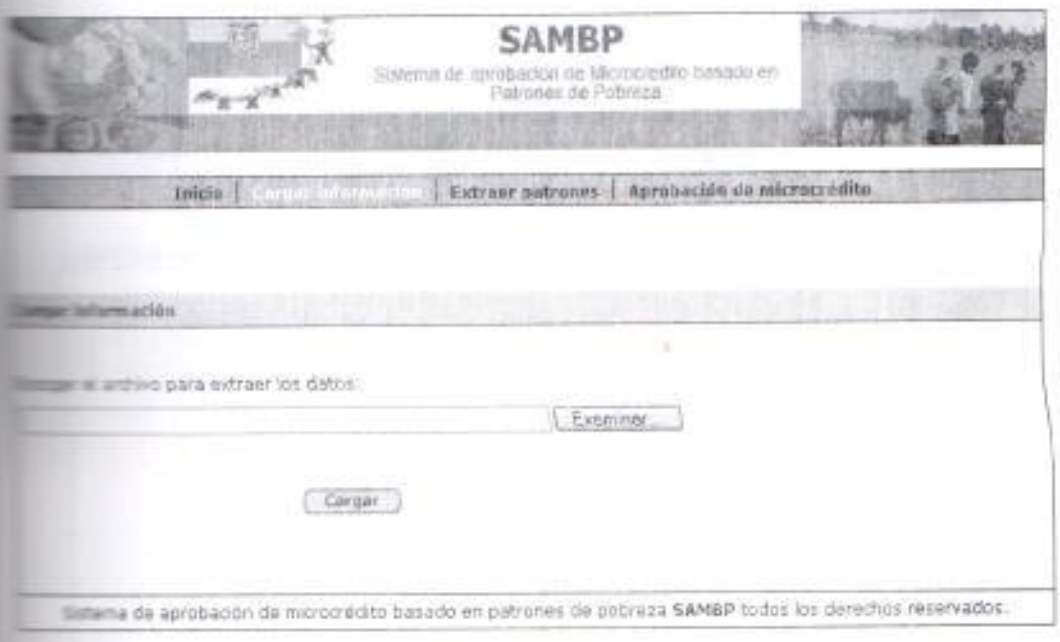


Figura 37. Cargar datos

Figura 38. Extracción de patrones

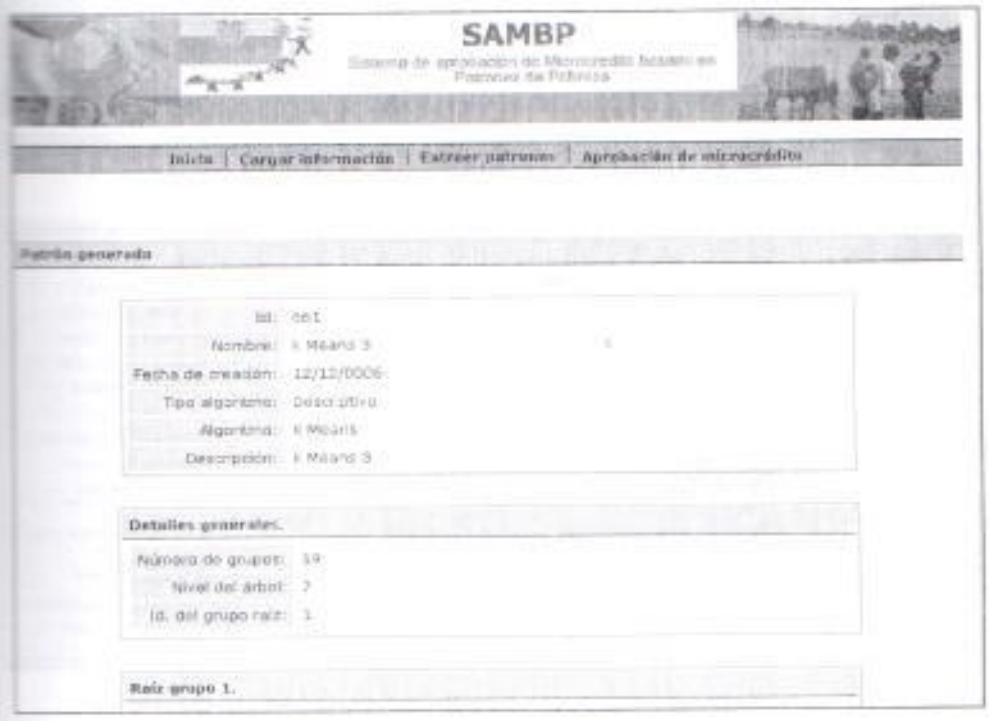


Figura 39. Patrones Generados

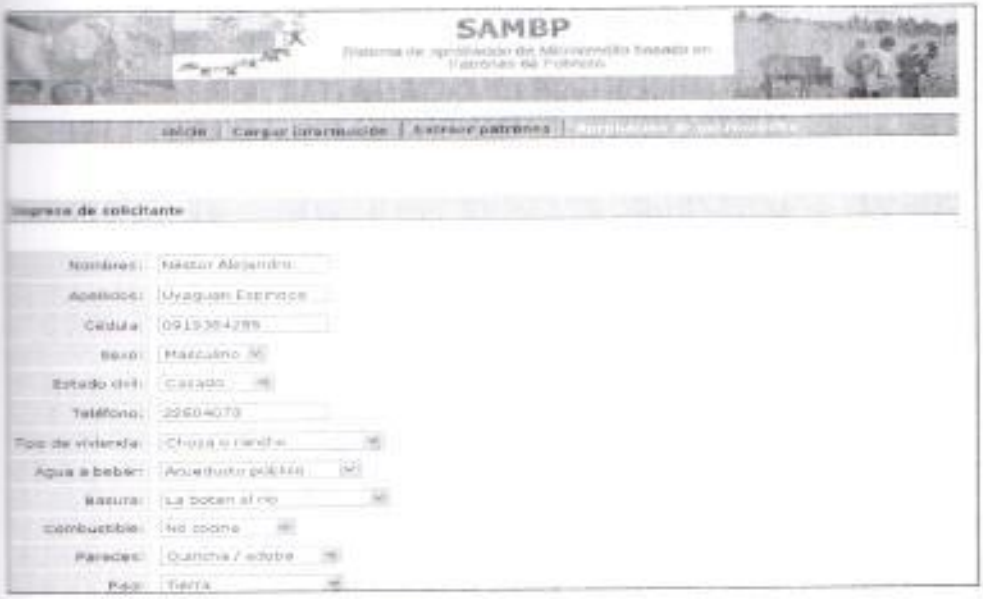


Figura 40. Ingreso de datos

Figura 41. Resultados aprobación microcrédito

3.2.5. Herramientas para el desarrollo del sistema

Las herramientas que se uso para el desarrollo y la implementación del "sistema de aprobación de microcrédito basado en patrones de pobreza" son:

- Oracle 10g release 10.2.1
- Exadel/Eclipse release 3.1.2
- PLSQL Developer versión 7.0.2
- Fireworks MX 2004

3.2.6. Implantación, evaluación y pruebas

Mapa conceptual de las pruebas de unidades

El sistema de aprobación de microcrédito llevó cabo las pruebas de las unidades basándonos en el mapa conceptual que detallamos a continuación. [16]

Figura 41. Mapa conceptual de las pruebas de unidades

El objetivo específico de las pruebas de unidades es encontrar y corregir la mayor cantidad de defectos, en donde se prueba cada módulo por separado, a continuación explicamos los pasos:

- Plan para prueba de unidades, es la identificación de los puntos en donde se encuentran los mayores problemas potenciales; es decir son aquellas partes de los módulos donde se pueden generar errores o resultados erróneos.

- Adquisición de conjunto de pruebas, parte en donde se adquiere resultados anteriores que hemos obtenido realizando pruebas con dicho módulo.
- Ejecución de las pruebas de unidades, ejecución del módulo para analizar los resultados que este nos muestra y poderlos comparar con resultados obtenidos anteriormente.

Tipo de pruebas usadas

Los tipos de pruebas utilizadas en las unidades fueron: caja negra, para ir observando que los procesos sean los correctos y probar los casos mas representativos, posteriormente usamos:

- La caja gris, debido a que no nos despreocupamos de que es lo que está pasando dentro de nuestros procesos.
- La prueba de la caja blanca ya cuando el producto se encuentra casi listo sin errores ni fallas.

En cada fase de prueba de los módulos se probó con datos de frontera, los cuales son los posibles valores que

vamos a estar ingresando al sistema y a su vez para que *no exista ningún error de validación*.

Los módulos que fueron probados cuidadosamente debido a que son los mas importante del sistema fueron: carga de datos, extracción de patrones (descriptivo K-Means, predictivo ABN), aprobación de microcrédito, debido a que un dato erróneo seria fatal en el sistema.

Para las pruebas de unidades usamos los datos de la Encuesta de Niveles de Vida 2003, realizado en Panamá, en el cual aplican la metodología "Living Standard Measurement Study" (LSMS) las cuales fueron adaptadas a la metodología de Cashpor.

Pruebas a usar en los métodos

Para realizar las pruebas de los métodos, lo realizaremos de esta manera:

1. Verificar la operación de valores normales de los parámetros.
2. Verificar la operación con valores límites de los parámetros.

3. Verificar la operación para los valores de parámetros fuera de los límites.

4. Verificar que se están ejecutando todas las instrucciones.

5. Verificar la conexión abierta hacia la base de datos.

6. Verificar todas las trayectorias que pueden seguir.

7. Verificar el uso de todos los objetos llamados.

8. Verificar el manejo de todas las estructuras de datos.

9. Verificar el manejo de la base de datos.

10. Verificar la terminación de todos los ciclos.

11. Verificar cierre de conexión con la base de datos.

CAPÍTULO 4

4. ANÁLISIS ECONÓMICO DEL SISTEMA

4.1. Análisis de costo

4.1.1. Estimaciones de costo-beneficio

Costo: El costo de desarrollo del prototipo incluye los siguientes rubros:

- Costo de equipo de desarrollo

CANTIDAD	EQUIPO	DESCRIPCIÓN	MARCA	PRECIO
2	CPU	Memoria 1 GB DDR, Mainboard BIOSTAR, Procesador: Intel Celeron 2.66Ghz, Disco duro: Samsung 80GB	ELECLON	$ 600.00
2	MONITOR	15"	LG	$ 200.00
2	MOUSE	OPTICO	BDI	$ 20.00
2	TECLADO		SAT	$ 20.00
SUBTOTAL				$ 840.00
IVA(12%)				$ 100.8
TOTAL				$ 940.80

Tabla 10. Costo de equipos

Nota: Los valores detallados en el costo de equipos fueron obtenidos de: Tekocsa S.A "Dr. PC".

- *Costo de infraestructura de desarrollo*

CANTIDAD	EQUIPO	DESCRIPCIÓN	MARCA	PRECIO
2	ESCRITORIO	-------	-------	$ 80,00
2	SILLA RECLINABLE	-------	-------	$ 80,00
SUBTOTAL				$ 160,00
IVA(12%)				$ 19,2
TOTAL				**$ 179,20**

Tabla 11. Costo de infraestructura

Nota: Los valores detallados en el costo de infraestructuras fueron obtenidos de: Ferrisariato

- Costo de implantación

Tipo de Gasto	No.	Costo	Costo Final
Equipo Servidor	1	$ 1.300,00	$ 1.300,00
Windows XP Professional	1	$ 300,00	$ 300,00
Unidades de Respaldo (HD)	2	$ 160	$ 320
Licencia Base de Datos (Oracle 10g Estándar Edition) Servidor [17]	1	$ 4.995,00	$ 4.995,00
Valor adicional por conexión	1	$ 149	$ 149
Web Content Tomcat	1	Open source	$ 0
Total			**$ 7.064,00**

Tabla 12. Costo de implantación

Nota: Los costos de implantación no se incluirán en el valor total de la inversión inicial, debido a que en dichos gastos incurrirá el cliente.

- Costo de personal de desarrollo

PERSONAL	FUNCION	TIEMPO TRABAJO (MES)	SUELDO
1	Desarrollo de interfaz	1	$ 700.00
1	Desarrollo procesos OLAP y mineria de datos	2	$ 2,000.00
1	Desarrollo procesos entre interfaz y cubo de datos	3	$ 2,000.00
TOTAL			**$ 4,700.00**

Tabla 13. Costo de personal

Nota: Los valores detallados en el costo de personal de desarrollo fueron obtenidos de: Grupo Isaías.

- Costo de materiales de desarrollo

SERVICIOS	TIEMPO (MESES)	VALOR
ENERGIA ELECTRICA	3	$ 100.00
INTERNET CABLE MODEM	3	$ 280.00
TOTAL		$ 380.00

Tabla 14. Costo de materiales

TOTAL COSTOS INCURRIDOS:		$ 6.200,00

Beneficios:

- Disminuir margen de tiempo para aprobar un microcrédito, ya que se orienta para personas muy pobres en las cuales solo interesa sus caracteristicas de vivienda.
- Automatizar una de las herramientas para medir niveles de pobreza como CASHPOR para ser aplicada en instituciones financieras u ONGs.
- Ahorrar gastos de papelería, en los formularios en donde se detallan las caracteristicas de la vivienda. Debido a que el sistema SAMBP, permite este ingreso de manera digital.
- Rapidez y eficiencia en el proceso de predicción para aprobar microcrédito, debido a que se extrae patrones de pobreza mediante técnicas de minería de datos que ayudan a simplificar la toma de decisiones.
- Simplificar trámites internos para analizar los datos, debido a que el sistema extrae conocimiento en el cual se basa para realizar futuras predicciones.

Viabilidad del sistema

Una de las razonas mas importantes por la cual el sistema "SAMBP" será viable en el Ecuador, se debe a que principalmente es una herramienta de predicción y referencia de aprobación de microcrédito orientado hacia los más pobres.

Considerando que existen actualmente en el mercado ecuatoriano alrededor de 34 cooperativas de ahorro y crédito. Como meta inicial se establecería vender el sistema "SAMBP" por lo menos al 25% de los potenciales clientes establecidos en 1 año y 6 meses.

Por lo tanto las estimaciones de utilidades serian:

Inversión Inicial	$6,200	
Estimación de venta de unidades		11
Estimación precio de venta		$3,000
Estimación de Utilidad Bruta		$11,000
Estimación Utilidad Neta		$6,000

Tabla 15. Viabilidad del sistema

4.2. Análisis Comercial

4.2.1. Análisis de FODA

Fortalezas:

- El sistema usa técnicas de minerías de datos, técnicas que ya son comprobadas y verificadas en su confianza y exactitud.
- El sistema está desarrollado en un ambiente Web con lo que tiene mayor alcance para ser usado en distintas partes.
- El sistema Web está desarrollado con Open Source, lo que le hace tener un bajo costo y ser implantado en distintas plataformas.

Oportunidades:

- Instituciones financieras interesadas en el otorgamiento de microcrédito.
- Personas con escasos recursos económicos que no cumplen con los requisitos normales para la obtención de microcrédito.

- En el Ecuador no existe un sistema que aplique alguna metodología como CASHPOR para identificar a los más pobres.

Debilidades:

- Poco interés por parte de ciertas instituciones financieras, de realizar préstamos sin fines de lucro.

- Falta de censos por parte del INEC, donde se establezcan los niveles de pobreza entre los pobres, aplicando alguna metodología como: Cashpor.

Amenazas:

- Falta de emprendimiento por parte de los pobres.
- Falta de compromiso de los adquirientes del microcrédito para pagarlo.

4.2.2. Cuota de mercado y volumen de ventas

La cuota de mercado que se pretende alcanzar es aquella que actualmente está ocupado por las 34 cooperativas de ahorro y crédito existentes en el Ecuador.

En cuanto al volumen de venta esperado, se pretende vender inicialmente al 25% de las cooperativas de ahorro y crédito, las cuales son: [18]

- 11 de junio
- 15 de abril
- 23 de julio
- 29 de octubre
- 9 de octubre ltda.
- Alianza del valle Ltda.
- Andalucía
- Atuntaqui
- CACPE Biblián Ltda.
- Cacpeco
- Calceta Ltda.
- Cámara de Comercio de Quito Ltda.
- Financoop.
- Chone Ltda.
- Codesarrolo
- Comercio ltda.
- Cotocollao
- De la pequeña empresa de Pastaza
- El sagrario
- Guaranda
- Jesús del gran poder ltda.

- Juventud ecuatoriana progresista ltda.
- La dolorosa
- Nacional
- Oscus
- Pablo Muñoz Vega
- Previsión, ahorro y desarrollo
- Progreso
- Riobamba
- San Francisco de Asís
- San José ltda.
- Santa Ana
- Santa Rosa
- Tulcán

4.2.3. Cliente objetivo

Específicamente el segmento de mercado al cual vamos a dirigir nuestros servicios, es para todas aquellas instituciones financieras que dan crédito a personas de bajos recursos económicos, las cuales le es difícil seleccionar el candidato idóneo para conceder un microcrédito de manera: óptima, confiable y eficiente.

Como estrategia inicial de venta nos podríamos enfocar a aquellas entidades microfinancieras tales como: las cooperativas de ahorro y crédito, debido a que por lo general conceden microcrédito a la pequeña y mediana empresa ecuatoriana.

4.2.3.1 Política de producto

El sistema de aprobación de microcrédito estará alojado en un servidor al que podrá ser accesado por varias usuarios simultáneamente, el sistema está desarrollado y estructurado de una manera que es escalable a futuras modificaciones y necesidades que presente el cliente.

Una vez entregado el sistema, se otorgara un mes El sistema, el cual fue desarrollado basándose en la necesidades y requerimientos del cliente, tendrá un periodo de un mes para proceder a realizar cambios de diseño, no de requerimiento, de manera gratuita; luego de este periodo los cambios que requiera el cliente tendrán un costo. El sistema se encuentra garantizado por un

periodo de un año; debido a que se estima, que en dicho tiempo se deba dar mantenimiento al sistema, a través, de actualizaciones, las cuales serían mejoras de rendimiento.

Un mes antes de la entrega del sistema se tendrá un periodo de capacitación, en el cual se dictará un curso a los usuarios sobre el manejo adecuado del sistema.

4.2.3.2 Política de precios

Para realizar un adecuado establecimiento del precio de venta del sistema, se procederá a analizar los tres puntos de referencia:

Costo de fabricación: $6200.00

Precio de la competencia: Actualmente en el mercado ecuatoriano no existe sistema alguno que posea el sentido de orientación social, y que a la vez, utilice técnicas de extracción de conocimiento con la finalidad de predecir perfiles idóneos para la aprobación de microcrédito.

Sin embargo, para citar precios podemos mencionar a Cooperativa Nacional, la cual recientemente ha adquirido un sistema financiero netamente transaccional con un costo aproximado de $22.000, cabe resaltar que dicho valor no incluye precio de implantación.

Demanda: Con un adecuado sistema de comercialización nos enfocaremos a atraer la demanda de las cooperativas de ahorro y crédito.

4.2.3.3 Política de distribución

La distribución la podemos efectuar de dos formas.

Distribución directa: Este tipo de distribución se realizará con las instituciones interesadas en adquirir el producto, se realizarán visitas programadas previamente, para hablar con el gerente de sistemas y el director financiero de la entidad. Y realizar la debida exposición del sistema analizando las ventajas de la aplicación de "SAMBP" en su negocio.

Distribución indirecta: Este tipo de distribución se realizará mediante intermediarios, llegando a aquellos lugares en donde reciben mantenimiento la infraestructura computacional de las instituciones financieras y a donde se dirigen las instituciones financieras para pedir consultoría, además estos pedidos se realizarán en territorio nacional e internacional.

4.2.3.4 Política de comunicación

La política de comunicación que usaremos para dar a conocer nuestro producto será por medio de dos medios:

Publicidad: A través de periódicos, revistas tecnológicas, revistas financieras y dirigidos mayoritariamente a las mujeres de escasos recursos y a personas en general entre los 35 y 55 años de edad. [19]

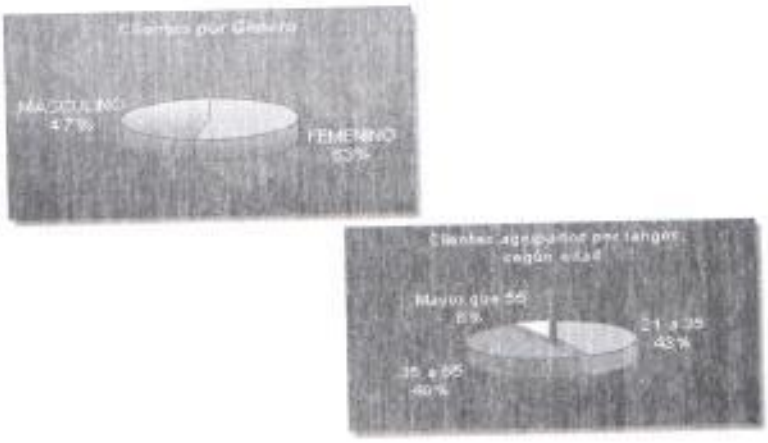


Figura 42. Clientes de microcrédito por género y por edad[15]

Promociones de venta: Las promociones para la venta de nuestro producto y para incentivar a terceros que vendan el producto, será proveer de un porcentaje de descuentos para aquellos que adquieren el producto en cantidades considerables.

[15] Credife – Desarrollo Microempresarial: http://www.credife.com/index.html

CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

1. El número de variables o predictores que inicialmente eran nueve (Tipo_vivienda, Piso, Pared, Techo, Combustible, Agua_beber, Sanitario, Basura y Tenencia), por medio del algoritmo "ABN" se redujeron a tres, lo cual indica que solamente las variables "Piso", "Combustible", y Tipo_vivienda" tiene mayor incidencia en el atributo clase "Condición" debido al análisis de entropía que ABN utiliza para establecer la significancia entre los predictores y el atributo clase.

2. Los patrones predictivos obtenidos a través del algoritmo "Adaptive Bayes Network", permitieron clasificar a una persona como pobre o muy pobre en base al análisis de caracteristicas de su vivienda.

3. De acuerdo al análisis de "Elbow", el cual realiza un análisis previo de los datos de la muestra poblacional para obtener un número determinado de clúster que optimice la posterior agrupación de los datos; determinó que el número de cluster sea 5.

4. Se pudo obtener información descriptiva por medio de la aplicación del algoritmo "Enhanced K-Means", con la final de proveer un marco referencial acerca de cómo esta compuesta la población a analizar.

5. Por medio de la obtención de la "matriz de confusión" se pudo proveer confianza acerca del modelo ABN obtenido, ya que muestra probabilidades de certeza del 97.06% y una precisión del 97.73%.

6. El análisis del espacio ROC permitió denominar a nuestro modelo predictivo ABN como "Buen Clasificador", debido a que el punto resultado es cercano a 1.

7. La minería de datos permite generar conocimiento que ayuda a mejorar la toma de decisiones en los procesos fundamentales de un negocio.

8. Todas las operaciones de preparación, creación de modelos y aplicación, permanecen en la base de datos de nuestro sistema lo que resulta en una mejora de la productividad, automatización e integración.

9. Como hemos trabajado con la base de datos de Oracle nos limitamos a que cualquier algoritmo que sea adaptado al sistema en algún futuro, tenga que ser basado en la estructura y modelo de la base de datos.

10. En vista que en el Ecuador no existe un software que aplique alguna metodología como CASHPOR para medir la pobreza, el sistema "SAMBP" tendrá gran acogida entre aquellas instituciones financieras

u ONGs interesadas en realizar un verdadero análisis para promover microcréditos en las comunidades pobres del país.

11. Las perspectivas de adopción de la herramienta "SAMBP", son muy altas; partiendo del principio de la llamada lucha contra la pobreza, en la cual todas las instituciones financieras están llamadas a colaborar, ejerciendo una "Banca Ética", que dedique parcial o totalmente inversiones para financiar actividades de alto rendimiento social a través del microcrédito. Haciendo a un lado el mal concepto de que pobreza no significa morosidad y que los pobres son gente en la que se puede confiar.

RECOMENDACIONES

1. Para que el sistema "SAMBP" pueda ser usado, requiere necesariamente de la plataforma Oracle, cuya versión incluya Data Mining.

2. Debemos usar la técnica de minería de datos cuando tenemos grandes volúmenes de información y deseamos predecir algún evento o suceso para tomar decisiones a futuro.

3. En un futuro nuestro sistema puede mejorarse agregándole más técnicas de mineria de datos, en el caso de que se le quiera dar otro enfoque de análisis a nuestros datos.

4. Otra mejora que se podria añadir al sistema seria proveer reportes históricos de las aprobaciones de microcrédito.

5. Un factor importante en nuestro sistema es que la base de datos con la cual trabajemos debe de ser de origen no dudoso, es decir no manipulada por otros, debido a que los resultados que van a ser generados depende de la credibilidad de la misma.

6. El sistema "SAMBP" podría ser utilizado para obtener puntajes de referencias para calcular el valor "P", en base al análisis de las caracteristicas de la vivienda de los estudiantes, siempre y cuando se realicen las debidas modificaciones que se ajusten a los requerimientos de la ESPOL.

BIBLIOGRAFÍA

[1] MARIANA MARTÍNEZ. (21 Febrero 2004), El poder del Microcrédito. BBC Mundo-Economía. Obtenido el 15 de junio del 2006 en: http://news.bbc.co.uk/hi/spanish/business/barometro_economico/newsid_3509000/3509551.stm.

[2] PLANET FINANCE: http://planetfinance.org/ES/ong-microfinanzas/ong-presentacion.php

[3] EL COMERCIO. (13 Febrero del 2003). Banca Ecuatoriana dirige su apoyo al microcrédito. Obtenido el 15 de junio del 2006 en: http://www.elpanamaamerica.com.pa/archive/02132002/finance07.html

[4] CAROLA BONDE BONFIL. (2005). Orientación de los servicios microfinancieros hacia los más pobres. Economía, Sociedad y Territorio, vol. V, núm. 17. Pág. 1-170. Obtenido el 15 de junio de 2006 en: http://www.cmq.edu.mx/documentos/Revista/revista17/est17_6.pdf

[5] THE MICROCREDIT SUMMIT CAMPAIGN. (5 Noviembre de 2003). Estado de la Campaña de la Cumbre del Microcrédito Informe 2003. Obtenido

el 15 de junio de 2006 en: http://www.microcreditsummit.org/spanish/index.html

[6] ORACLE. (17 de Diciembre de 2004). Oracle Data Mining. Pág. 1. Obtenido el 15 de junio del 2006 en: http://www.oracle.com/global/es/database/docs/oracle_data_mining.pdf

[7] MICROSOFT CORPORATION. (2006). Conceptos de mineria de datos. (2006). Obtenido el 15 de junio de 2006 en: http://msdn2.microsoft.com/es-es/library/ms174949.aspx

[8] ORACLE. (Junio 2005). Oracle Data Mining Concepts 10g Release 2 (10.2) Part Number B14339-1. Obtenido el 15 de junio del 2006 en:http://downloadeast.oracle.com/docs/cd/B19306_01/datamine.102/b14339.pdf

[9] ORACLE. (Noviembre 2005). Oracle Data Mining Administrator's Guide 10g Release 2 (10.2) Part Number B14338-02B14338-02. Obtenido el 15 de junio de 2006 en: http://download-east.oracle.com/docs/cd/B19306_01/datamine.102/b14338.pdf

[10] ORACLE. (Junio 2005). Oracle Data Mining Application Developer's Guide 2 (10.2) Part Number B14340. Obtenido el 15 de junio de 2006 en: http://download-east.oracle.com/docs/cd/B19306_01/datamine.102/b14340/toc.htm

[11] ORACLE. (Agosto 2005). Oracle Database Installation's Guide 10g Release 2 (10.2) for Microsoft Windows Part Number B14320. Obtenido el 15 de junio de 2006 en: http://download-east.oracle.com/docs/cd/B19306_01/win.102/b14320/toc.htm

[12] QUINLAN, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers. Obtenido el 15 de julio del 2006 en: http://en.wikipedia.org/wiki/C4.5_algorithm

[13] LIC. ENRIQUE JOSÉ FERNÁNDEZ. (Diciembre 2004). Análisis De Clasificadores Bayesianos. Pág. 6-10. Obtenido el 16 de julio del 2006 en: http://www.itba.edu.ar/capis/webcapis/trabajosfinalesdeespecialidad/fernadez-trabajofinaldeespecialidad.pdf

[14] EDUARDO MORALES. (Enero 1999). Heurísticas de Selección de Atributos. Obtenido el 29 de septiembre del 2006 en: http://ccc.inaoep.mx/~emorales/Cursos/KDD/node41.html

[15] SERGIO M. SAVARESI AND DANIEL L. BOLEY. (1997). On the performance of bisecting K-means and PDDP. Pág. 1-4. Obtenido el 15 de julio del 2006 en: http://www.siam.org/meetings/sdm01/pdf/sdm01_05.pdf

[16] JUAN PABLO GIRALDO RENDÓN - UNIVERSIDAD DE MANIZALES. (2000). Métricas. Obtenido el 10 de junio del 2006 en: http://www.monografias.com/trabajos15/ingenieria-software/ingenieria-software2.shtml

[17] ORACLE – STORE (2006). Precios licencias Oracle Estándar Edition. Obtenido el 10 de junio del 2006 en: http://oraclestore.oracle.com/OA_HTML/ibeCZzpHome.jsp?minisite=10021&respid=22372&grp=STORE&language=US

[18] SUPERINTENDENCIA DE BANCOS Y SEGUROS – ECUADOR. (Junio 2006). Cooperativas - Estadísticas de Depósitos. Obtenido el 10 de junio del 2006 en: https://www.superban.gov.ec/pages/c_coop_estadisticas_depositos.htm

[19] CREDIFE. (2005). Desarrollo Microempresarial – Nuestros Clientes - Perfil del Cliente. Obtenido el 10 de marzo del 2006 en: http://www.credife.com/index.html

[20] HISPAVISTA. (2006). Mi proyecto empresarial-Estrategia Empresarial. Obtenido el 10 de marzo del 2006 en http://www.trabajos.com/informacion/index.phtml?n=10&s=4

[21] HERNÁNDEZ, J.; RAMÍREZ, M.J. y FERRI. Prentice Hall, España, primera edición. 2004. Introducción a la Minería de Datos.

[22] HAIR, J.F.; ANDERSON, R.E.; TATHAM, R.L. y BLACK, W.C. Prentice Hall, Madrid, 1999. Análisis multivariante.

[23] ORACLE. (Septiembre 2005). Oracle Data Mining "Know More, Do More, Spend Less". Obtenido el 15 de agosto del 2006 en: http://www.oracle.com/technology/products/bi/odm/pdf/bwp_db_odm_10gr2_0905.pdf

[24] ORACLE. (2003). Descriptive Data Mining Models. Obtenido el 15 de julio del 2006 en: http://www.stanford.edu/dept/itss/docs/oracle/10g/datamine.101/b10698/4descrip.htm

[25] Ángel M. Ramos Domínguez. Análisis de Cluster. Obtenido el 12 de agosto del 2006 en: webpages.ull.es/users/aramos/CLUSTERS.ppt

ANEXOS

ANEXO A

Diagrama de Gannt

Id.	Nombre de tarea	Comienzo	Fin	Duración
1	Propuesta de tema	6/1/2006	6/5/2006	3d
2	Reunion con director de tópico para aprobación de tema	6/6/2006	6/9/2006	4d
3	Redacción de temario	6/9/2006	6/12/2006	2d
4	Reunion con director de tópico para aprobacion de temario	6/13/2006	6/14/2006	2d
5	Investigación para obtención de datos internet	6/15/2006	6/16/2006	2d
6	Redacción de solicitud inec para obtención de datos	6/16/2006	6/16/2006	1d
7	Establecimiento de plataforma de trabajo	6/16/2006	6/19/2006	2d
8	Descarga de herramientas de desarrollo	6/19/2006	6/20/2006	2d
9	Instalación de herramientas de desarrollo	6/20/2006	6/23/2006	4d
10	Preparacion de datos para extraccion a la BD	6/23/2006	6/23/2006	1d
11	Creacion de la Base de datos	6/23/2006	6/23/2006	1d
12	Preparacion y creacion del ambiente web	6/23/2006	6/26/2006	2d
13	Investigación del mecanismo de extracción de datos.	6/27/2006	6/29/2006	3d
14	Desarrollo de Algoritmo Kmean y ABN sobre java	6/30/2006	7/14/2006	11d
15	Implementación y pruebas del proceso de extracción de datos	7/14/2006	7/17/2006	2d
16	Implementacion del proceso de extraccion de patrones.	7/17/2006	7/20/2006	4d
17	Implementación y pruebas del proceso de clasificación de nuevos datos	7/20/2006	7/24/2006	3d
18	Implementación de consultas	7/25/2006	7/25/2006	1d
19	Ultimación de detalles en la parte web	7/26/2006	7/31/2006	4d

Id	Nombre de tarea	Comienzo	Fin	Duración
1	Propuesta de tema	6/1/2006	6/5/2006	3d
2	Reunion con director de tópico para aprobación de tema	6/6/2006	6/9/2006	4d
3	Redacción de temario	6/9/2006	6/12/2006	2d
4	Reunion con director de tópico para aprobacion de temario	6/13/2006	6/14/2006	2d
5	Investigación para obtención de datos internet	6/15/2006	6/16/2006	2d
6	Redacción de solicitud inec para obtencion de datos	6/15/2006	6/16/2006	1d
7	Establecimiento de plataforma de trabajo	6/16/2006	6/19/2006	2d
8	Descarga de herramientas de desarrollo	6/19/2006	6/20/2006	2d
9	Instalacion de herramientas de desarrollo	6/20/2006	6/23/2006	4d
10	Preparación de datos para extraccion a la BD.	6/23/2006	6/23/2006	1d
11	Creacion de la Base de datos	6/23/2006	6/23/2006	1d
12	Preparacion y creacion del ambiente web	6/23/2006	6/26/2006	2d
13	Investigación del mecanismo de extracción de datos.	6/27/2006	6/29/2006	3d
14	Desarrollo de algoritmo Khosan y ABN sobre java	6/30/2006	7/14/2006	11d
15	Implementación y pruebas del proceso de extracción de datos	7/14/2006	7/17/2006	2d
16	Implementacion del proceso de extracion de patrones.	7/17/2006	7/20/2006	4d
17	Implementacion y pruebas del proceso de clasificación de nuevos datos	7/20/2006	7/24/2006	3d
18	Implementación de consultas	7/25/2006	7/25/2006	1d
19	Ultimación de detalles en la parte web	7/26/2006	7/31/2006	4d

Id	Nombre de tarea	Comienzo	Fin	Duración
1	Propuesta de tema	6/1/2006	6/5/2006	3d
2	Reunion con director de tópico para aprobación de tema	6/6/2006	6/9/2006	4d
3	Redacción de temario	6/9/2006	6/12/2006	2d
4	Reunion con director de tópico para aprobación de temario	6/13/2006	6/14/2006	2d
5	Investigación para obtención de datos internet	6/15/2006	6/16/2006	2d
6	Redacción de solicitud ined para obtención de datos	6/16/2006	6/16/2006	1d
7	Establecimiento de plataforma de trabajo	6/16/2006	6/19/2006	2d
8	Descarga de herramientas de desarrollo	6/19/2006	6/20/2006	2d
9	Instalación de herramientas de desarrollo	6/20/2006	6/23/2006	4d
10	Preparacion de datos para extraccion a la BD	6/23/2006	6/23/2006	1d
11	Creacion de la Base de datos	6/23/2006	6/23/2006	1d
12	Preparacion y creacion del ambiente web	6/23/2006	6/26/2006	2d
13	Investigación del mecanismo de extracción de datos.	6/27/2006	6/29/2006	3d
14	Desarrollo de algoritmo Id3 y ANN sobre java	6/30/2006	7/14/2006	11d
15	Implementación y pruebas del proceso de extracción de datos	7/14/2006	7/17/2006	2d
16	Implementacion del proceso de extracción de patrones.	7/17/2006	7/20/2006	4d
17	Implementación y pruebas del proceso de clasificación de nuevos datos	7/20/2006	7/21/2006	3d
18	Implementación de consultas	7/25/2006	7/25/2006	1d
19	Ultimación de detalles en la parte web	7/26/2006	7/31/2006	4d

Id	Nombre de tarea	Comienzo	Fin	Duración
1	Propuesta de tema	6/1/2006	6/5/2006	3d
2	Reunion con director de tópico para aprobación de tema	6/6/2006	6/9/2006	4d
3	Redacción de temario	6/9/2006	6/12/2006	2d
4	Reunion con director de tópico para aprobacion de temario	6/13/2006	6/14/2006	2d
5	Investigación para obtención de datos internet	6/15/2006	6/16/2006	2d
6	Redacción de solicitud mec para obtención de datos	6/16/2006	6/16/2006	1d
7	Establecimiento de plataforma de trabajo	6/16/2006	6/19/2006	2d
8	Descarga de herramientas de desarrollo	6/19/2006	6/20/2006	2d
9	Instalación de herramientas de desarrollo	6/20/2006	6/23/2006	4d
10	Preparación de datos para extracción a la BD.	6/23/2006	6/23/2006	1d
11	Creacion de la Base de datos	6/23/2006	6/23/2006	1d
12	Preparacion y creacion del ambiente web	6/23/2006	6/26/2006	2d
13	Investigación del mecanismo de extracción de datos	6/27/2006	6/29/2006	3d
14	Desarrollo de algoritmo Kmean y ABN sobre java	6/30/2006	7/14/2006	11d
15	Implementación y pruebas del proceso de extracción de datos	7/14/2006	7/17/2006	2d
16	Implementacion del proceso de extracción de patrones	7/17/2006	7/20/2006	4d
17	Implementación y pruebas del proceso de clasificación de nuevos datos	7/20/2006	7/24/2006	3d
18	Implamentación de consultas	7/25/2006	7/25/2006	1d
19	Ultimación de detalles en la parte web	7/26/2006	7/31/2006	4d

ANEXO B

Diccionario de datos

- **Descripción de variables utilizadas :**

Id_variable: Identificador de los posibles valores cualitativos de la variable.

Descripción: Nombre de los valores de la variable.

Valor_Cashpor: Valor numérico, que representa el peso de cada uno de los posibles valores de la variable. Estos valores servirán para obtener un total_cashpor por vivienda.

Las variables que se utilizaron en el sistema "SAMBP" son:

1. Nombre _ variable: TipoVivienda

ID_TIPO_VIVIENDA	DESCRIPCIÓN	VALOR_CASHPOR
1	CASA INDIVIDUAL	4
2	CHOZA O RANCHO	1
3	APARTAMENTO	2
4	CUARTO EN CASA DE VECINDAD	2
5	IMPROVISADA	0.5
6	OTRO	0

7		
8		
9	NR	5

2. Nombre _ variable: Paredes

ID_PAREDES	DESCRIPCIÓN	VALOR_CASHPOR
1	BLOQUE/LADRILLO	4
2	MADERA	3
3	QUINCHA/ADOBE	2
4	METAL	2
5	CAÑA/PAJA/PALOS	1
6	SIN PAREDES	0
7	OTROS MATERIALES	0
8		
9	NR	5

3. Nombre _ variable: techo

ID_TECHO	DESCRIPCIÓN	VALOR_CASHPOR
1	CONCRETO/CEMENTO	4
2	TEJA	3

3	FIBRA	2
4	METAL	2
5	MADERA	1
6	PAJA/PECA	0
7	OTROS	0
8		
9	NR	5

4. Nombre _ variable: Piso

ID_PISO	DESCRIPCIÓN	VALOR_CASHPOR
1	CONCRETO/CEMENTO	4
2	LADRILLO	3
3	MADERA	2
4	TIERRA	0
5	OTROS	0
6		
7		
8		
9	NR	5

5. Nombre _ variable: Tenencia

ID_TENENCIA	DESCRIPCIÓN	VALOR_CASHPOR
1	PROPIA	4
2	HIPOTECADA	3
3	ALQUILADA	2
4	PRESTADA	1
5	OCUPANTES DE HECHO	1
6		
7		
8		
9	NR	5

6. Nombre _ variable: Agua_beber

ID_AGUA_BEBER	DESCRIPCIÓN	VALOR_CASHPOR
1	ACUEDUCTO PUBLICO	4
2	ACUEDUCTO COMUNIDAD	3
3	ACUEDUCTO PARTICULAR	2
4	POZO SANITARIO	1
5	POZO BROCAL	1
6	RIO	0
7	OTRO	0
8		

9	NR	5

7. Nombre _ variable: Sanitario

ID_SANITARIO	DESCRIPCIÓN	VALOR_CASHPOR
1	ALCANTARILLADO	4
2	TANQUE SEPTICO	2
3	LETRINA	1
4	NO TIENE	1
5		
6		
7		
8		
9	NR	5

8. Nombre _ variable: Basura

ID_BASURA	DESCRIPCIÓN	VALOR_CASHPOR
1	SERVICIO MUNICIPAL	4
2	SERVICIO PARTICULAR	4
3	LA BOTAN A OTROS LOTES	3
4	LA BOTAN DENTRO DEL PATIO	2
5	LA BOTAN AL RIO	1

6	LA QUEMAN	1
7	LA ENTIERRAN	0
8	OTRO	0
9	NR	5

9. Nombre _ variable: Combustible

ID_COMBUSTIBLE	DESCRIPCIÓN	VALOR_CASHPOR
1	GAS	4
2	LEÑA	1
3	ELECTRICIDAD	4
4	NO COCINA	2
5	OTRO	0
6		
7		
8		
9	NR	5

Fuente: Encuesta Niveles de Vida 2003. Panamá. Base de Datos de Vivienda y Capital Social/Sección 1 Vivienda y Hogar. Archivo E03HG01.pdf. Obtenido en: http://www.mef.gob.pa/Politicas%20Sociales/Documentos%20a%20Publicar%20en%20el%20WEB/Encuesta%20de%20Niveles%20de%20Vida%202003/Composición%20de%20la%20Base%20de%20Datos.htm

- **Diccionario de la Base de datos**

Tabla: Propietario

Atributos:

Id _ propietario: Clave primaria de la tabla Propietario, de tipo Number (7).

Nombre: Nombre del propietario, de tipo Varchar2 (20).

Apellido: Apellido del propietario, de tipo Varchar2 (20).

Cédula: Cédula del propietario, de tipo Varchar2 (10).

Sexo: Género femenino o masculino del propietario, de tipo Varchar2 (20).

Estado _ civil: Estado civil del propietario, de tipo Varchar2 (20).

Teléfono: Número de teléfono del propietario, de tipo Varchar2 (10).

Tabla: Tipo_vivienda

Atributos:

Id_tipo_vivienda: Clave primaria de la tabla Tipo_vivienda, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Tipo_vivienda, de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla tipo_vivienda. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Tipo_vivienda, de tipo Varchar2 (1).

Tabla: Paredes

Atributos:

Id_paredes: Clave primaria de la tabla Paredes, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Paredes, de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla Paredes. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Paredes, de tipo Varchar2 (1).

Tabla: Techo

Atributos:

Id_techo: Clave primaria de la tabla Techo, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Techo, de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla Techo. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Techo, de tipo Varchar2 (1).

Tabla: Piso

Atributos:

Id_piso: Clave primaria de la tabla Piso, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Piso de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla Piso. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Piso, de tipo Varchar2 (1).

Tabla: Tenencia

Atributos:

Id_tenencia: Clave primaria de la tabla Tenencia, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Tenencia, de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla Tenencia. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Tenencia, de tipo Varchar2 (1).

Tabla: Agua_beber

Atributos:

Id_agua_beber: Clave primaria de la tabla Agua_beber, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Agua_beber, de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla Agua Beber. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Agua_beber, de tipo Varchar2 (1).

Tabla: Sanitario

Atributos:

Id_sanitario: Clave primaria de la tabla Sanitario, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Sanitario, de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla sanitario. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Sanitario, de tipo Varchar2 (1).

Tabla: Basura

Atributos:

Id_basura: Clave primaria de la tabla Basura, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Basura, de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla Basura. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Basura, de tipo Varchar2 (1).

Tabla: Combustible

Atributos:

Id _ combustible: Clave primaria de la tabla Combustible, de tipo Number (7).

Descripción: Nombre de los valores que puede tener la tabla Combustible, de tipo Varchar2 (50).

Valor: Identificador numérico de los posibles valores categóricos de la tabla combustible. De tipo varchar2 (1).

Valor_cashpor: Valor numérico o peso que tiene cada valor de la tabla Combustible, de tipo Varchar2 (1).

Tabla: Vivienda

Atributos:

Id_vivienda: Clave principal de la tabla Vivienda, de tipo Number (7).

Dirección: ubicación de la vivienda, de tipo Varchar2 (30).

Provincia: Nombre de la provincia del Ecuador donde se encuentra la vivienda, de tipo Varchar2 (15).

Ciudad: Nombre de la ciudad donde se encuentra la vivienda, de tipo Varchar2 (15).

Sector: Nombre del sector (norte, sur, este, oeste) donde se encuentra la vivienda, de tipo Varchar2 (10).

Barrio: Nombre del barrio (Guayacanes, Kennedy, etc) donde se encuentra la vivienda, de tipo Varchar2 (10).

Id _ propietario: Clave foránea perteneciente a la tabla Propietario, de tipo Number (7).

Id_tipo_vivienda: Clave foránea perteneciente a la tabla Tipo_vivienda, de tipo Number (7).

Id_paredes: Clave foránea perteneciente a la tabla Paredes, de tipo Number (7).

Id_techo: Clave foránea perteneciente a la tablaTecho, de tipo Number (7).

Id_piso: Clave foránea perteneciente a la tabla Piso, de tipo Number (7).

Id_tenencia: Clave foránea perteneciente a la tabla Tenencia, de tipo Number (7).

Id_agua_beber: Clave foránea perteneciente a la tabla Agua_beber, de tipo Number (7).

Id_sanitario: Clave foránea perteneciente a la tabla Sanitario, de tipo Number (7).

Id_basura: Clave foránea perteneciente a la tabla Basura, de tipo Number (7).

Id _ combustible: Clave foránea perteneciente a la tabla Combustible, de tipo Number (7).

Total_Cashpor: Suma total de los valores Cashpor de cada variable, de tipo Number (3).

Condición: Se refiere a denominar a una vivienda como "pobre" o "muy pobre" dependiendo del total_cashpor previamente obtenido. De tipo Varchar2 (12).

Tabla: Algoritmo

Atributos:

Descripción: Se refiere al nombre del algoritmo que se utiliza para generar el modelo, el cual puede ser ABN o K-Means. El atributo es de tipo Varcha2 (40).

Id_algoritmo: Clave principal de la tabla Algoritmo, de tipo Number (4).

Id_tipo_algoritmo: Clave foránea perteneciente a la tabla Tipo_Algoritmo, de tipo Number (4).

Tabla: Tipo_Algoritmo

Atributos:

Descripción: Se refiere al tipo de algoritmo creado, el cual puede ser Predictivo o Descriptivo. De tipo Varchar2 (40).

Id_tipo_algoritmo: Clave principal de la tabla Tipo_Algoritmo, de tipo Number (4).

Tabla: Modelo

Atributos:

Nombre: Se refiere al nombre que el usuario ingrese cuando desea guardar el modelo del algoritmo generado. De tipo Varchar2 (20).

Descripción: Se refiere a detalles que el usuario ingrese al momento de guardar el modelo. De tipo Varchar2 (20).

Fecha_creación: Se refiere a la fecha de creación del modelo, de tipo Date.

Id_algoritmo: Clave foránea perteneciente a la tabla Algoritmo, de tipo Number (4)

Id_modelo: Clave principal de la tabla Modelo, de tipo Number (4).

Tabla: Aprobados

Atributos:

Id_aprobados: Clave principal de la tabla Aprobados, de tipo Number (4).

Id_modelo: Clave foránea perteneciente a la tabla Modelo, de tipo Number (4).

Id_propietario: Clave foránea perteneciente a la tabla Propietario, de tipo Number (4).

Fecha_aprobación: Se refiere a la fecha en la cual se aprueba el microcrédito al propietario de acuerdo a los patrones generados. Atributo de tipo Date.

PROCEDIMIENTOS:

Procedure Cashpor: Procedimiento que permitirá obtener los valores denominados "total_cashpor" de cada vivienda y fijar el atributo "condición" como "pobre" si "total_cashpor" es mayor que 20 o "muy pobre", si "total_cashpor" es menor que 20.

VISTAS PARA ALGORITMOS DE MINERÍA DE DATOS

Vista Mining_data_build_v2 : selecciona los datos de la Tabla vivienda para que los algoritmos K-means y ABN puedan construir sus respectivos modelos.

Vista Aplicar _ modelo: Guarda los datos del nuevo propietario ingresado por medio del formulario. Para poder aplicar el modelo del algoritmo escogido y predecir su condición como "pobre" o "muy pobre".

- **Reglas/Patrones predictivos generados por el algoritmo "Adaptive Bayes Network" (ABN)**

Patrónes generados	
Nombre:	Abn
Fecha de creación:	2006-07-19 00:00:00.0
Tipo algoritmo:	Predictivo
Algoritmo:	Adaptive Bayes Network
Descripción:	predeterminado
Resultado	0
Soporte:	0.47854785478547857
Confidencia:	0.9978042205975507
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	1
Soporte:	0.14207134999214203
Confidencia:	0.9995829680534777
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	2
Soporte:	0.05799151343705799
Confidencia:	0.9022879929866813
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Lena

	TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	**3**
Soporte:	0.04699041332704699
Confidencia:	0.9997581919241741
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**4**
Soporte:	0.04384724186704365
Confidencia:	0.9429631653752693
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal muy pobre
Resultado	**5**
Soporte:	0.034574886060034574
Confidencia:	0.9996979416451017
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado	**6**
Soporte:	0.029545811724029546
Confidencia:	0.9225851543431105
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	**7**
Soporte:	0.02765990864802766
Confidencia:	0.9986217064041743
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado	**8**

Soporte:	0.026402640264026403
Confidencia:	0.6936503169262099
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resulatdo	**9**
Soporte:	0.022002200220022004
Confidencia:	0.978722895298702
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado	**10**
Soporte:	0.0165016501650165
Confidencia:	0.6577031141060069
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal muy pobre
Resultado	**11**
Soporte:	0.010215307245010215
Confidencia:	0.965552872901333
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	**12**
Soporte:	0.0068008600880088
Confidencia:	0.9986602641346572
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**13**
Soporte:	0.008329404389000833
Confidencia:	0.9997674606854982
Antecedente:	PISO isIn Otros COMBUSTIBLE isIn Lena

	TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**14**
Soporte:	0.005972025774006972
Confidencia:	0.9990963713295841
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado	**15**
Soporte:	0.003928964325003929
Confidencia:	0.9177721094527942
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**16**
Soporte:	0.0033003300330033004
Confidencia:	0.9992352161542554
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado	**17**
Soporte:	0.0026716957410026715
Confidencia:	0.9960846702883455
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado	**18**
Soporte:	0.0026716957410026715
Confidencia:	0.8179577165027515
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**19**

Soporte:	0.0023573785950023575
Confidencia:	0.9997184596108322
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Improvisada
Consecuente:	CONDICION equal muy pobre
Resultado	**20**
Soporte:	0.0023573785950023575
Confidencia:	0.6779455013904367
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal muy pobre
Resultado	**21**
Soporte:	0.002043061449002043
Confidencia:	0.9978491823929265
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado	**22**
Soporte:	0.0018859026760031886
Confidencia:	0.9106822759897968
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**21**
Soporte:	0.0017287443030017287
Confidencia:	0.7278472395904738
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal pobre
Resultado	**22**
Soporte:	0.0015715557300015717
Confidencia:	0.9945870155163875
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn No cocina

	TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado	**23**
Soporte:	0.0014144271570014145
Confidencia:	0.9896886251030018
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Electricidad TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	**24**
Soporte:	0.0011001100110011
Confidencia:	0.9996231829827162
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Otro
Consecuente:	CONDICION equal pobre
Resultado	**25**
Soporte:	9.42951438000943E-4
Confidencia:	0.9890379596944746
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**26**
Soporte:	9.42951438000943E-4
Confidencia:	0.9810111229685315
Antecedente:	PISO isIn Otros COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal muy pobre
Resultado	**27**
Soporte:	9.42951438000943E-4
Confidencia:	0.8355947940741011
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Otro
Consecuente:	CONDICION equal pobre
Resultado	**28**

Soporte:	9.42951438000943E-4
Confidencia:	0.7384479746361146
Antecedente:	PISO isIn Otros COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	**29**
Soporte:	7.857928650007858E-4
Confidencia:	0.9981045026242196
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	**30**
Soporte:	7.857928650007858E-4
Confidencia:	0.60751479941413
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	**31**
Soporte:	6.286342920006286E-4
Confidencia:	0.9989953315192538
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Improvisada
Consecuente:	CONDICION equal muy pobre
Resultado	**32**
Soporte:	6.286342920006286E-4
Confidencia:	0.5046262868303552
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Improvisada
Consecuente:	CONDICION equal pobre
Resultado	**33**
Soporte:	4.7147571900047 15E-4
Confidencia:	0.999948654166345
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Electricidad

	TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado	34
Soporte:	4.7147571900047 15E-4
Confidencia:	0.9995037229270445
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Electricidad TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	35
Soporte:	4.7147571900047 15E-4
Confidencia:	0.9143176847558 36
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	36
Soporte:	4.7147571900047 15E-4
Confidencia:	0.7002917433476 92
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Otro TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	37
Soporte:	4.7147571900047 15E-4
Confidencia:	0.6919101535561 25
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado:	38
Soporte:	4.7147571900047 15E-4
Confidencia:	0.6702343977919115
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Improvisada
Consecuente:	CONDICION equal pobre
Resultado	39

Soporte:	3.143171460003143E-4
Confidencia:	0.9995311354219201
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Otro TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**40**
Soporte:	3.143171460003143E-4
Confidencia:	0.9998771331219242
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Electricidad TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado	**41**
Soporte:	3.143171460003143E-4
Confidencia:	0.9987552629103001
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Electricidad TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal pobre
Resultado	**42**
Soporte:	3.143171460003143E-4
Confidencia:	0.9565569747310805
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado	**43**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9999943360695169
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Otro TIPO_VIVIENDA isIn Otro
Consecuente:	CONDICION equal muy pobre
Rsultado	**44**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9999219453828098
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Electricidad

	TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado	**45**
Soporte:	1.5715857300015716E-4
Confidencia:	0.999855453397922
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Electricidad TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado	**46**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9997266499365852
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Electricidad TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado	**47**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9995450945852061
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado	**48**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9991586887942193
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Otro TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal muy pobre
Resultado	**49**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9990486057095918
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Otro TIPO_VIVIENDA isIn Casa individual
Consecuente:	CONDICION equal muy pobre
Resultado:	**50**

Soporte:	1.5715857300015716E-4
Confidencia:	0.9974776269045451
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Otro
Consecuente:	CONDICION equal muy pobre
Resultado:	**51**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9973169497454752
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Otro
Consecuente:	CONDICION equal muy pobre
Resultado:	**52**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9972937916890955
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Otro
Consecuente:	CONDICION equal muy pobre
Resultado:	**53**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9958324402227082
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn improvisada
Consecuente:	CONDICION equal pobre
Resultado:	**54**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9782026289867042
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Otro TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado:	**55**
Soporte:	1.5715857300015716E-4
Confidencia:	0.9680913150891983
Antecedente:	PISO isIn Concreto/Cemento COMBUSTIBLE isIn Lena

	TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado:	56
Soporte:	1.5715857300015716E-4
Confidencia:	0.9673915976925314
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Cuarto en casa de vecindad
Consecuente:	CONDICION equal pobre
Resultado:	57
Soporte:	1.5715857300015716E-4
Confidencia:	0.9473753163213208
Antecedente:	PISO isIn Madera COMBUSTIBLE isIn No cocina TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado:	58
Soporte:	1.5715857300015716E-4
Confidencia:	0.9412572714653779
Antecedente:	PISO isIn Otros COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre
Resultado:	59
Soporte:	1.5715857300015716E-4
Confidencia:	0.9401068896897929
Antecedente:	PISO isIn Tierra COMBUSTIBLE isIn Gas TIPO_VIVIENDA isIn Apartamento
Consecuente:	CONDICION equal pobre
Resultado:	60
Soporte:	1.5715857300015716E-4
Confidencia:	0.8610597605051491
Antecedente:	PISO isIn Ladrillo COMBUSTIBLE isIn Lena TIPO_VIVIENDA isIn Choza o rancho
Consecuente:	CONDICION equal muy pobre

- **Preparación de los datos categóricos para la aplicación del algoritmo "Enhanced K-Means"**

	NOMBRE DEL ATRIBUTO	BIN_ID	VALOR
1	AGUA_BEBER	1	Otro
2	AGUA_BEBER	2	Pozo brocal
3	AGUA_BEBER	3	Pozo sanitario
4	AGUA_BEBER	4	Acueducto público
5	AGUA_BEBER	5	Acueducto particular
6	AGUA_BEBER	6	Acueducto comunidad
7	AGUA_BEBER	7	Río
8	BASURA	1	La entierran
9	BASURA	2	La botan dentro del patio
10	BASURA	3	La botan a otros lotes
11	BASURA	4	La notan al río
12	BASURA	5	Servicio particular
13	BASURA	6	Servicio municipal
14	BASURA	7	Otro
15	BASURA	8	La queman
16	COMBUSTIBLE	1	Electricidad
17	COMBUSTIBLE	2	Gas
18	COMBUSTIBLE	3	Otro
19	COMBUSTIBLE	4	No cocina
20	COMBUSTIBLE	5	Leña
21	CONDICION	1	muy pobre
22	CONDICION	2	pobre
23	PAREDES	1	Quincha/Adobe
24	PAREDES	2	Otros materiales
25	PAREDES	3	Metal
26	PAREDES	4	Madera
27	PAREDES	5	Cana/Paja/Palos
28	PAREDES	6	Bloque/Ladrillo
29	PAREDES	7	Sin Paredes
30	PISO	1	Otros
31	PISO	2	Madera
32	PISO	3	Ladrillo
33	PISO	4	Concreto/Cemento
34	PISO	5	Tierra
35	SANITARIO	1	Alcantarillado
36	SANITARIO	2	Letrina
37	SANITARIO	3	No tiene
38	SANITARIO	4	Tanque Septico
39	TECHO	1	Metal
40	TECHO	2	Madera
41	TECHO	3	Fibra

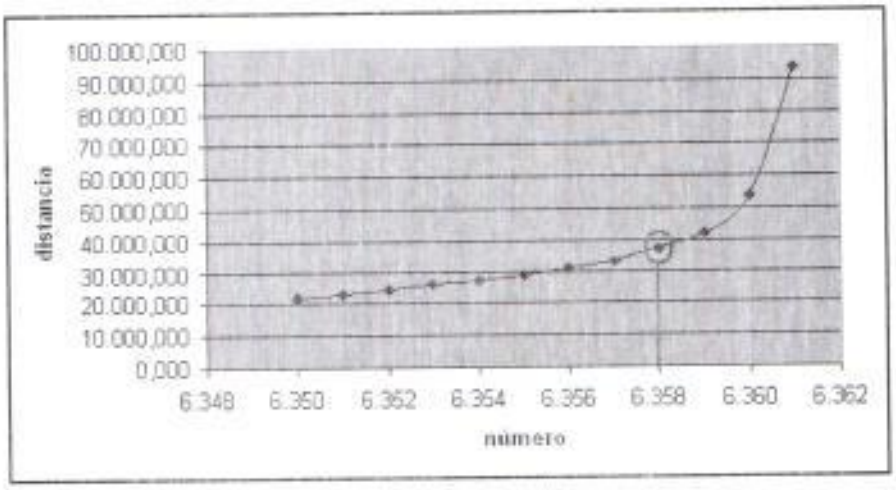
42	TECHO	4	Concreto/Cemento
43	TECHO	5	Otros
44	TECHO	6	Teja
45	TECHO	7	Paja/Penca
46	TENENCIA	1	Alquilada
47	TENENCIA	2	Hipotecada
48	TENENCIA	3	NR
49	TENENCIA	4	Ocupantes de hecho
50	TENENCIA	5	Prestada
51	TENENCIA	6	Propia
52	TIPO_VIVIENDA	1	Apartamento
53	TIPO_VIVIENDA	2	Casa individual
54	TIPO_VIVIENDA	3	Choza o rancho
55	TIPO_VIVIENDA	4	Improvisada
56	TIPO_VIVIENDA	5	Cuarto en casa de vecindad
57	TIPO_VIVIENDA	6	Otro

- **Análisis de "Elbow" para obtener el numero (k) óptimo de clusters para aplicar el algoritmo de K-Means**

De acuerdo al análisis de "Elbow", realizado por medio del programa SPSS, se procede a graficar los datos obtenidos y escoger el número cuya diferencia de distancias es mayor con respecto a las anteriores. [24]

Número	Distancia
6.348	20.129,332
6.349	20.929,590
6.350	21.900,299
6.351	23.056,309
6.352	24.416,762
6.353	25.805,180
6.354	27.212,010
6.355	28.908,961

6.356	30.716,784
6.357	33.168,148
6.358	37.288,398
6.359	42.065,976
6.360	53.531,328
6.361	93.618,776



Luego se realiza una resta entre el número escogido y la cantidad de casos de la población, tal como se muestra abajo:

Número de casos: 6363

Paso "Elbow" 6358

Número de clusters: 5

ANEXO C

Manual de Usuario

Contenido

- Requerimientos mínimos de hardware
- Requerimientos de software
- Base de datos y Minería de datos
- Instalación de la aplicación "SAMBP"
- Manejo del sistema "SAMBP"

- Requerimientos mínimos de hardware

Requerimientos de procesador	Intel 486 DX PC (o similar) o superior
Requerimientos de Memoria	1 GB
Requerimientos de Software	Windows XP PRO
Requerimientos aproximados de espacio en Disco duro	Arriba de 4 GB de espacio libre, dependiendo de los componentes que se instalen.

- Requerimientos de software

- Oracle 10g release 10.2.1
- Eclipse 3.2.1
- Exadel 4.0.2
- J2sdk 1.5.0

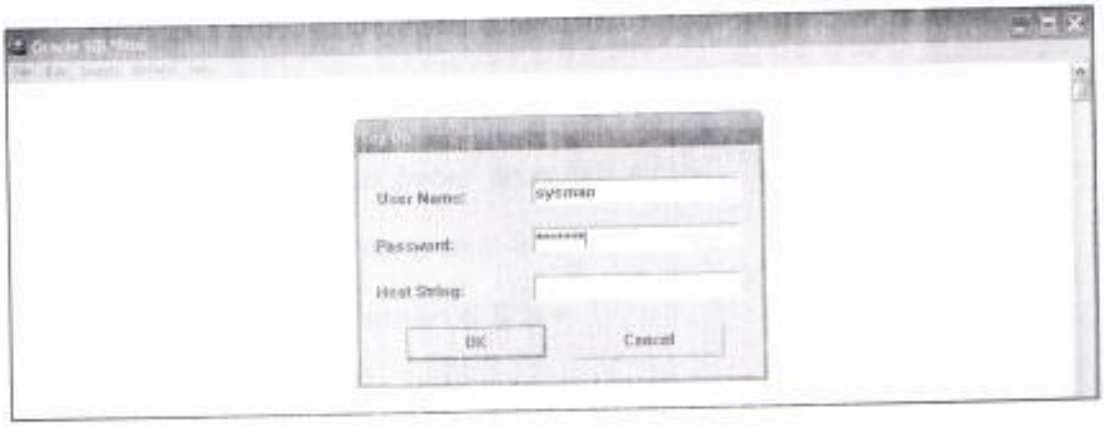
- Base de datos y Minería de datos

Se procede a crear usuarios, tablas, vistas y procedeimientos necesarios para el funcionamiento del sistema.

1.- Abrir Sql plus

Inicio- Programas – Oracle – Application Developer – Sql+

2.- Ingresar como Usuario: sysman y Password: (de acuerdo al password de la base que se puso al momento de la instalación).



3.- Creación de usuario y privilegios en la base de datos

- Ejecutar el script *usuario* que viene en el CD de instalación en el SQL+
- Conectado como sysman, ejecutar la siguiente linea en el SQL+

```
CREATE OR REPLACE DIRECTORY DM_DUMP AS 'c:\oracle\product\10.2.0';
```

- Ejecutar el script *object privileges* en el SQL+
- Conectarse con Usuario: mineria y Password: mineria en el SQL+
- Ejecutar el script *role privileges* en el SQL+
- Ejecutar el script *system privileges* en el SQL+

4.- Creación de tablas en la base de datos

Conectado a la base de datos con el usuario: mineria:

- Ejecutar el script *propietario* en el SQL+
- Ejecutar el script *tipo_vivienda* en el SQL+
- Ejecutar el script *paredes* en el SQL+
- Ejecutar el script *techo* en el SQL+
- Ejecutar el script *piso* en el SQL+
- Ejecutar el script *tenencia* en el SQL+
- Ejecutar el script *agua_beber* en el SQL+
- Ejecutar el script *sanitario* en el SQL+
- Ejecutar el script *basura* en el SQL+

- Ejecutar el script *combustible* en el SQL+
- Ejecutar el script *vivienda* en el SQL+
- Ejecutar el script *usuario* en el SQL+
- Ejecutar el script *system privileges* en el SQL+
- Ejecutar el script *tipo_algoritmo* en el SQL+
- Ejecutar el script *algoritmo* en el SQL+
- Ejecutar el script *modelo* en el SQL+
- Ejecutar el script *aprobados* en el SQL+

5.- Cargar datos a las tablas

- Copiar los archivos *.ctl y *.csv que se encuentran en el CD de instalación a la unidad c:\ de su disco duro.
- Abrir la consola de Windows

Inicio-Ejecutar - cmd

- Ejecutar en la consola de Windows la siguiente línea:

sqlldr mineria/mineria control=c:\<archivo ctl>.ctl

Ej:

sqlldr mineria/mineria control=c:\propietario.ctl

Realizar el mismo proceso con cada uno de los archivos *.ctl que se copio anteriormente en el mismo orden con que se crearon las tablas anteriormente.

6.- Creación de procedimientos

- Abrir el SQL+
- Conectarse con el usuario: mineria
- Ejecutar el script cashpor que se encuentra en el CD de instalación
- Ejecutar el siguiente comando: *execute cashpor;*

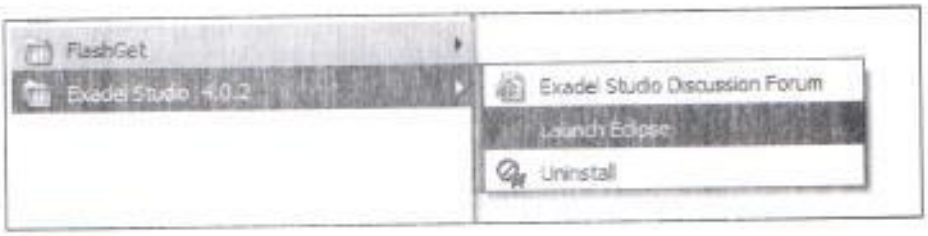
7.- Creación de vistas

- Ejecutar el script *mining_data_build_v2* que se encuentra en el CD de instalación.
- Ejecutar el script *mining_data_test_v2* que se encuentra en el CD de instalación.
- Ejecutar el script *mining_data_apply_v2* que se encuentra en el CD de instalación.

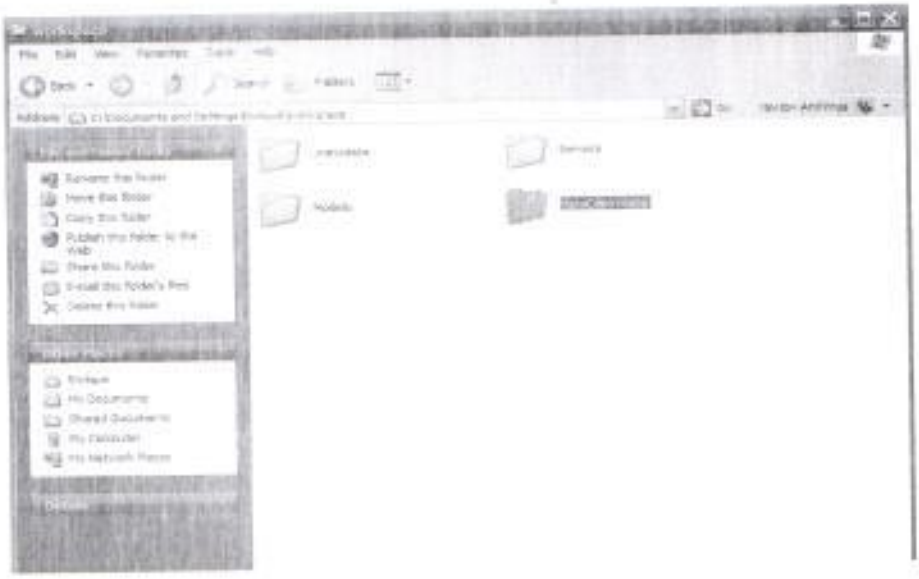
- Instalación de la aplicación "SAMBP"

1.- Abrir el programa Exadel Studio

Inicio-Programas – Exadel Studio – launch Eclipse



2.- Copiar la carpeta Modelo y VistaControlador que se encuentran en el CD en la carpeta workspace del usuario logoneado en Windows.



3.- En ExadelStudio importar las carpetas Modelo y VistaControlador. Para ello ir a: File- Import – StrutProject – Finish. Repetir los pasos para importar la carpeta VistaControlador.

- Manejo del sistema "SAMBP"

1.- Ingreso al sistema

2.- Mensaje de bienvenida al sistema. Muestra el menú en la parte superior:

Cargar información, Extraer Patrones y Aprobación microcrédito.

3.- Modulo de Cargar información, mediante el cual se procede a llenar la base de datos.

4.- Módulo para extraer patrones, el cual permite aplicar los algoritmos de K-Means y Adaptive Bayes Network para obtener modelos de predicción.

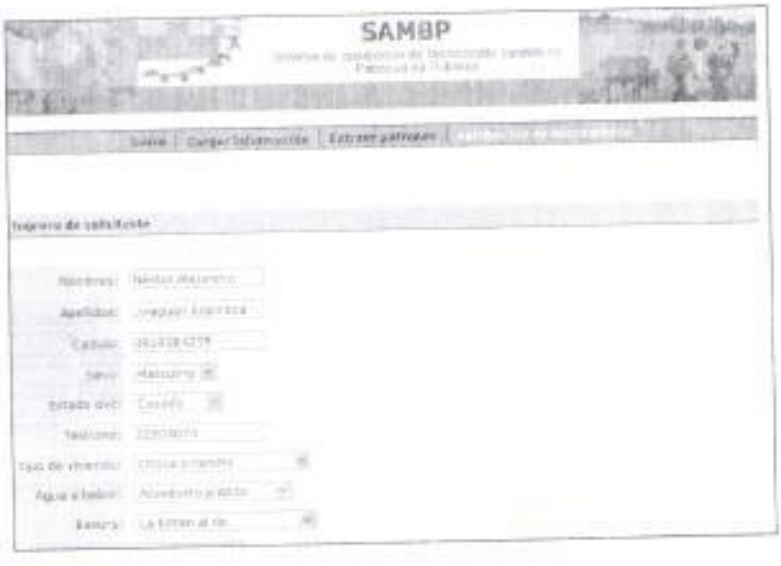
5.- Una vez que se escoge el algoritmo con el cual se extraerá los patrones, el sistema permite guardarlos.

6.- Pantalla donde se muestran los resultados de los patrones extraídos.


SAMBP


SAMBP

7.- Módulo de ingreso de información del solicitante de microcrédito



8.- Pantalla donde se muestra los resultado de aprobación del microcrédito en base a los análisis de los algoritmos de mineria de datos.

- **Solicitud presentada al Instituto Nacional de Estadísticas y Censos del Litoral**

Guayaquil, junio 12 del 2005

Sociólogo
Roberto Sáenz Ozaeta
Director Regional del Litoral (E) del INEC
Ciudad

De mis consideraciones:

Yo, **Pedro Fabricio Echeverría**, profesor de la **Escuela Superior Politécnica del Litoral**, solicito a usted se facilite información del VI Censo de la Población y V de vivienda para motivos académico de tesis a la Srta. **Maria Eugenia Andrade Ramirez**, estudiante politécnica.
Para lo cual detallo a continuación las zonas de las cuales se requiere dicha información:

LAMINAS	ZONAS	BARRIO

Z	349-350-353-354-359-360-361-362-363-364-365-366-367-371-372-373-374-375-376-377	Guasmo Este-Rio Guayas-Unión Bananeros
A-1	356-357-358-368-369-370	Puerto Nuevo-Guasmo Oeste(parte)
B-1	282-285-289-290-291-292-293-294-295-296-297-298-306-308-309-310-311-312-313-314-315-316-331-332-333-334-335-336-337-338	Isla Trinitaria-Luz del Guayas-Batallón
F	3-4-5-6-7-8	Pascuales

La información se la requiere digitalizada por vivienda:

1. Tipo de vivienda
   - Casa o Villa
   - Departamento
   - Cuarto(s) en casad e Inquilinato
   - Mediagua
   - Rancho
   - Covacha
   - Choza
   - Otro (especifique)

- Materiales predominantes en la vivienda.
- Techo o cubierta
- Paredes exteriores
- Piso
- Armazón o estructura

3. Abastecimiento de agua en la vivienda.
- ¿Como se obtiene el agua para la vivienda?
- ¿De donde provine el agua que recibe?

4. ¿Cómo se eliminan las aguas negras o servidas de esta vivienda?
5. ¿Dispone la vivienda de luz eléctrica?
6. ¿Como se elimina la basura de la vivienda?
7. ¿Cual es el principal combustible o energía que se utiliza en este hogar para cocinar?
8. El servicio Higiénico que tiene ese hogar es:
9. La vivienda que ocupa este hogar es:

Agradeciéndole por la ayuda prestada para realizar un proyecto con fines académicos.

Atentamente

Ing. Pedro Fabricio Echeverría
Profesor de la ESPOL