ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLÓGÍA

CENTRO DE POSGRADOS

MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DEMAGÍSTER EN: GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

MODELO PREDICTIVO DEL COMPORTAMIENTO DE LA CARTERA CREDITICIA PARA COOPERATIVAS DE AHORRO Y CRÉDITO

AUTORA: ING. TOSCANO PALOMO, GLADYS NATALI

DIRECTOR: MSC. PARRAGA VILLAMAR, VIVIANA CRISTINA

SANGOLQUÍ

2019

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y

TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación **"MODELO PREDICTIVO DEL COMPORTAMIENTO DE LA CARTERA CREDITICIA PARA COOPERATIVAS DE AHORRO Y CRÉDITO"** fue realizado por la señora **Toscano Palomo, Gladys Natali,** el mismo que ha sido revisado en su totalidad, analizado por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 15 de julio del 2019

Firma:

.....

Ing. Viviana Parraga Villamar MSc.
C.C.: 1721903407

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORÍA DE RESPONSABILIDAD

Yo, **Toscano Palomo, Gladys Natalí**, con cédula de identidad n° 0502967136, declaro que el contenido, ideas y criterios del trabajo de titulación: **"Modelo predictivo del comportamiento de la cartera crediticia para Cooperativas de Ahorro y Crédito"** es de mi autoría y responsabilidad, cumpliendo con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando citas bibliográficas.

Consecuentemente el contenido de la investigación mencionada es veraz.

Sangolquí, 11 de julio del 2019

Firma:

.....

Ing. Gladys Natali Toscano Palomo
C.C. 0502967136

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN

Yo, **Toscano Palomo, Gladys Natalí**, autorizo a la Universidad de la Fuerzas Armadas ESPE publicar el trabajo de titulación **"Modelo predictivo del comportamiento de la cartera crediticia para Cooperativas de Ahorro y Crédito"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 11 de julio del 2019

Firma:

.....

Ing. Gladys Natali Toscano Palomo
C.C. 0502967136

DEDICATORIA

A Dios

Por ser el motor de mi vida y darme la fuerza para continuar en este proceso para alcanzar esta meta tan anhelada.

A mis Padres

Que con su ejemplo, amor, trabajo y sacrificio, han sido mi mayor inspiración para llegar hasta aquí y convertirme en lo que soy. He tenido el privilegio de ser su hija y adquirir enseñanzas junto a los mejores padres.

A mis hermanos

Por estar siempre presentes, por el apoyo moral, que me han brindado a lo largo de esta experiencia de vida.

A todas las personas

Que con su apoyo, han aportado para que este trabajo llegue a culminarse con éxito.

AGRADECIMIENTO

Deseo expresar mi amor y mi gratitud, a mi hija y a mi esposo que con su paciencia e incansable ayuda, gracias a ellos he llegado alcanzar mis objetivos.

A mi Tutora Viviana Parraga, quien desde el primer momento supo guiarme y fue de gran apoyo en todo momento.

Al personal académico y administrativo de la Universidad de las fuerzas armadas ESPE, quienes supieron impartir sus conocimientos con mucho esfuerzo, para hacer posible el desarrollo de esta investigación.

Mi agradecimiento infinito a mi hermana Adriana Paola, quien supo apoyarme en todo momento y de todas las formas posibles.

Mi agradecimiento a todos, mi familia, mis amigos que de una u otra manera me brindaron su colaboración e incondicional apoyo.

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Las cooperativas de ahorro y crédito son organizaciones jurídicas que realizan actividades de intermediación financiera, aceptan depósitos, otorgan créditos y ofrecen una amplia variedad de otros servicios financieros. El producto central de una cooperativa son los microcréditos que se otorgan a quienes no puede justificar fácilmente sus ingresos. Cuando una persona ingresa a la cooperativa a solicitar un crédito, un asesor de crédito analiza su buró, luego se analiza los documentos que respaldan los ingresos y garantías en función a la actividad, el perfil de los socios y el destino al cual se va a dar uso el crédito; al final de acuerdo a su experiencia resuelve otorgar o negar el crédito. En este contexto, se creó un modelo predictivo para una cooperativa del comportamiento de la cartera crediticia, mediante un modelo de minería de datos, que determinó factores que influyen en el otorgamiento de créditos, ejemplo: tipo de crédito, plazo, frecuencia, estado civil, entre otros. La metodología se definió en 4 fases: Evaluación de herramientas y metodologías, Diseño del modelo, Implementación del modelo y Validación del modelo. Como resultado el modelo predictivo de la cartera crediticia se obtuvo una confianza del 99.9 % y se conocieron los patrones que cumplen un buen pagador.

PALABRAS CLAVES:

- **MODELO DE MINERÍA DE DATOS**
- **CARTERA CREDITICIA**
- **PATRONES DE COMPORTAMIENTO**
- **COOPERATIVAS DE AHORRO Y CRÉDITO**

ABSTRACT

Credit unions are legal organizations that perform financial intermediation activities, accept deposits, grant loans and offer a wide variety of other financial services. The central product of a cooperative is the microcredits that are granted to those who cannot easily justify their income. When a person enters the cooperative to apply for a loan, a credit counselor analyzes his bureau, then analyzes the documents that support the income and guarantees based on the activity, the profile of the members and the destination to which it is going to use the credit; in the end according to his experience he decides to grant or deny the credit. In this context, a predictive model was created for a cooperative of the behavior of the credit portfolio, through a data mining model, which determined factors that influence the granting of credits, for example: type of credit, term, frequency, marital status, among others. The methodology was defined in 4 phases: Evaluation of tools and methodologies, Design of the model, Implementation of the model and Validation of the model. As a result, the predictive model of the loan portfolio obtained a confidence of 99.9% and the patterns that meet a good payer were known.

KEYWORDS:

- **DATA MININGMODEL**
- **CREDIT CARD**
- **BEHAVIOR PATTERNS**
- **COOPERATIVES OF SAVING AND CREDIT**

CAPÍTULO I

INTRODUCCIÓN

1.1.Antecedentes

El sistema financiero está conformado por instituciones que tiene como objetivo canalizar el ahorro de las personas. Esta canalización de recursos permite el desarrollo de la actividad económica del país haciendo que los fondos lleguen desde las personas que tienen recursos monetarios excedentes hacia las personas que necesitan estos recursos. Los intermediarios financieros se encargan de captar depósitos del público dando lugar al ahorro para prestarlos a los demandantes de recursos y es aquí donde se generan los créditos.

En el Ecuador la Junta de Política y Regulación Monetaria y Financiera establece las políticas públicas, y la regulación y supervisión monetaria, crediticia, cambiaria, financiera, de seguros y valores. Los organismos de supervisión y control son la Superintendencia de Bancos para los bancos, mutualistas y sociedades financieras, la Superintendencia de Economía Popular y Solidaria (SEPS) para las cooperativas y mutualistas de ahorro y crédito de vivienda y la Superintendencia de Compañías, Valores y Seguros controla a las compañías de seguros.

Actualmente las Cooperativas de Ahorro y Crédito se conforman por la unión de un grupo de personas que tienen como fin ayudarse los unos a los otros con el fin de alcanzar sus necesidades financieras, no está formada por clientes sino por socios, ya que cada persona posee una pequeña participación dentro de esta. Es una organización jurídica que se encuentra legalmente constituida en el país; realiza actividades de intermediación financiera y de responsabilidad social con sus socios; y previa autorización de la SEPS con socios y terceros con sujeción a las regulaciones y a

los principios reconocidos en la Ley Orgánica de la Economía Popular y Solidaria y del Sector Financiero Popular y Solidario, a su Reglamento General, a las Resoluciones de la Superintendencia de Economía Popular y Solidaria y del ente regulador

El producto central de una cooperativa que da lugar al ahorro es el microcrédito, que está dirigido a microempresarios que cuenten con unidades de producción, comercio; para iniciar o ampliar su negocio, para socios comerciantes en diferentes mercados populares del país para capital de trabajo, incremento y ampliación del negocio, a aquellos cuyos ingresos se obtienen de un comercio en algunos casos informal y donde es difícil justificar de manera segura sus ingresos.

El riesgo crediticio es uno de los principales desafíos que enfrentan las cooperativas, ya que afecta negativamente la rentabilidad y estabilidad de la institución. Además, normalmente se presentan fallos al momento de elegir a quien dar préstamos incurriendo en aumento de índices de morosidad y por ende riego crediticio (Alborzi & Khanbabaei, 2016).

Existen aplicaciones que ayudan a la toma de decisiones en grandes instituciones financieras en todo del mundo. Estos modelos nacen como una necesidad de poder evaluar de forma ágil y rápida las capacidades de endeudamiento de sus clientes ante la solicitud de un crédito (Alborzi & Khanbabaei, 2016).

Además, las agencias bancarias tienen almacenada bastante información en sus bases de datos, que mediante técnicas de extracción de datos tienen numerosas aplicaciones en la calificación crediticia de clientes. Una de las técnicas de minería de datos más populares es el método de clasificación. Donde utilizando técnicas de extracción de datos se puede predecir y clasificar el puntaje crediticio del cliente para superar los riesgos futuros de otorgar préstamos a clientes que no pueden pagar (Gahlaut, Tushar, & Singh, 2017).

Los resultados de investigaciones muestran como análisis crediticio en entidades bancarias utilizando minería de datos arrojan factores como: rentabilidad, capacidad de pago, solvencia, duración de un informe de crédito, garantías, tamaño de la empresa, número de préstamo, estructura de propiedad y la duración de la relación con la banca corporativa que resultaron claves para predecir el incumplimiento. Además, se encontró que los resultados de la clasificación dependen de lo apropiado de las características de los datos y del algoritmo de análisis apropiado para los conjuntos de datos. La selección de variables financieras y no financieras, así como la resolución de desequilibrios de clase permiten a las empresas evaluar su riesgo de crédito con éxito (Khemakhem & Boujelbene, 2018).

Este estudio determinó el comportamiento de la cartera de crédito de una cooperativa de ahorro y crédito caso de estudio, prediciendo si un cliente es apto o no para acceder a un crédito, sin importar, que éste no tenga un historial crediticio.

1.2.Justificación e Importancia

El crédito bancario es de gran importancia para el desarrollo de la economía dentro de un país, mucho más en vías de desarrollo, debido a que es una de las principales fuentes de financiamiento para personas, microempresas y macroempresas. Los indicadores crediticios implican en menor o mayor medida un nivel de riesgo, dicha probabilidad está dada por los patrones de comportamiento que el socio puede tener en el futuro y que vuelven peligrosa la inversión bancaria. Por cuanto la concesión de créditos es la principal actividad de una entidad bancaria, pero a la vez uno de los mayores riesgos (Tello, Eslava, & Tobías, 2013).

El proyecto realizado obtuvo un modelo de predicción del comportamiento de la cartera crediticia, para cooperativas de ahorro y crédito, buscando reducir el índice de morosidad y

riesgo crediticio identificando a un buen pagador. Con la investigación realizada se consiguió optimizar el otorgamiento de créditos, minimizar el tiempo de cobranza y dar mayor tiempo al análisis de nuevos créditos.

Para crear el modelo seutilizaronalgoritmos que predijeron el comportamiento de los datos adaptándose a la necesidad del negocio.Posteriormente, se establecióquepatrones identifican a un mal pagador, utilizando un Dashboardpara visualizar los factores y el resultado final de buen o mal pagador.

1.3.Objetivo general

Desarrollar un modelo de predicción del comportamiento de la cartera crediticia, mediante la aplicación de patrones para determinar los factores que influyen en el otorgamiento de créditos.

1.4.Objetivos específicos

OE1: Analizar la situación actual del proceso de otorgamiento de crédito en cooperativas e investigar herramientas y metodologías para el análisis del comportamiento de cartera crediticia y patrones mediante una revisión sistémica de literatura parcial.

OE2: Diseñar el modelo predictivo del comportamiento de la cartera crediticia mediante minería de datos.

OE3: Implementar el modelo predictivo de la cartera crediticia mediante la aplicación de patrones.

OE4: Validar los resultados obtenidos de la aplicación de los patrones del modelo predictivo de la cartera crediticia.

1.5.Formulación del problema

De acuerdo con los objetivos específicos del proyecto de análisis, diseño e implementación de un modelo de predicción del comportamiento de la cartera de créditos basado en scoring de crédito de las Cooperativas de Ahorro y Crédito, se respondieron las siguientes preguntas para cada objetivo planteado:

OE1 – RQ1.1: ¿Cuáles son las herramientas y metodología que ayudarán a determinar los factores que influyen en el otorgamiento de créditos?

OE1 – RQ1.2: ¿Cuál es la situación actual del proceso de otorgamiento del crédito?

OE2 – RQ2.1: ¿Qué herramienta de minería de datos facilitará la implementación del modelo a desarrollar?

OE2 – RQ2.2: ¿Cuál es el gestor de bases de datos que permitirá manipular la información del datawarehouse de una manera eficaz y óptima?

OE3 – RQ3.1: ¿Cuál es el algoritmo de minería de datos que mejor se ajuste a los datos proporcionados?

OE3 – RQ3.2: ¿Cuál es la herramienta ETL más factible para el diseño y creación de una bodega de datos, a partir de las fuentes de información de las Cooperativas de Ahorro y Crédito?

OE5 – RQ5.1: ¿Cuál es el nivel de confianza aceptable para determinar a un modelo como eficiente?

OE5– RQ5.2: ¿Cuál es el margen de error aceptable para un modelo implementado?

CAPÍTULO II

FUNDAMENTACIÓN TEÓRICA

El marco teórico pretende establecer la relación existente entre la parte teórica y la hipótesis, de acuerdo con las variables del problema, con la finalidad para que la investigación este encaminada jerárquicamente por categoría hasta llegar a la categoría que comprende y explica las variables dependientes e independientes del tema de estudio, para esto se propone la siguiente jerarquía de estudio:

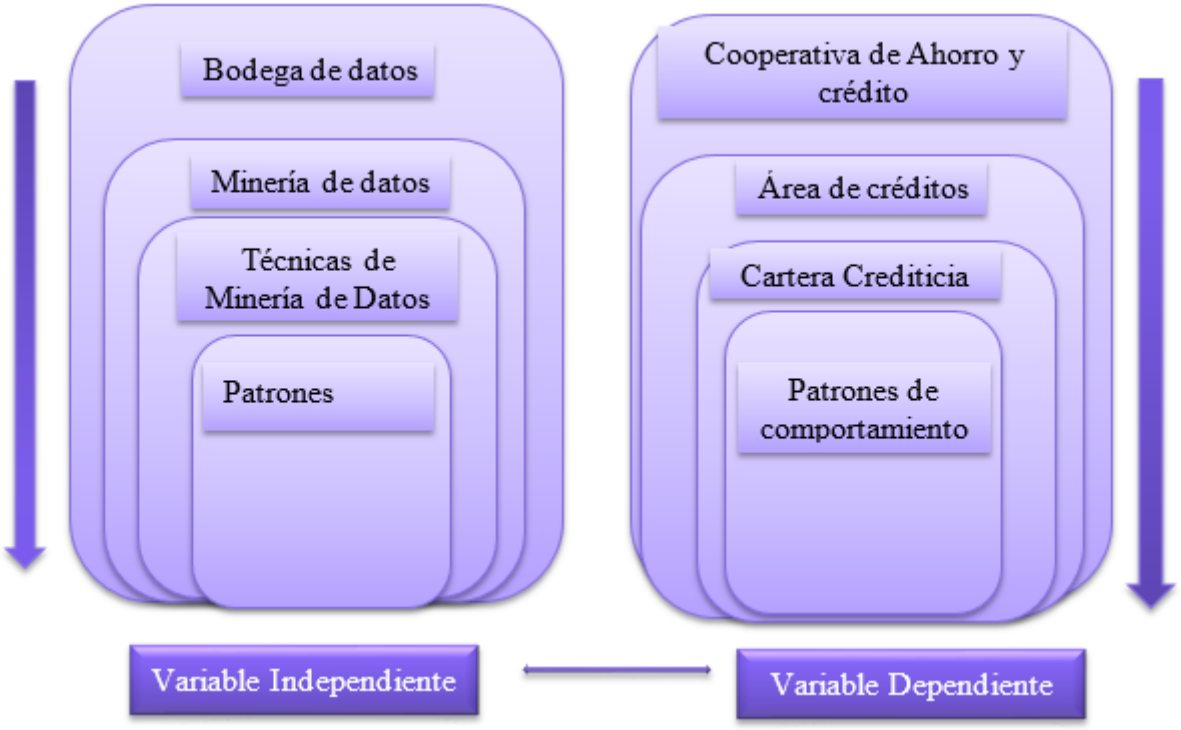


Figura 1.Relación de variables

2.1.Base de datos

2.1.1. Sistema Manejador de Base de Datos (DBMS)

Un sistema manejador de bases de datos (SGBD) o *DataBase Management System* (DBMS) es la interfaz entre la base de datos y el usuario mediante un determinado software que permite además utilizar distintas aplicaciones.

Los sistemas manejadores de base de datos tienen como objetivo manejar un conjunto de datos para convertirlos en información relevante para la organización, ya sea a nivel operativo o estratégico. Esto lo realiza utilizando programas que permiten manejar los datos de una manera segura, sencilla y ordenada permitiendo un mejor control a los administradores de sistemas ymejores resultados a la hora de realizar consultas que ayuden a la gestión.

Un sistema SGBD es conocido por tener características como:

- Independencia
- Redundancia mínima
- Consistencia de la información
- Abstracción de la información
- Acceso seguro
- Asegurar Integridad de los datos

Mediante estas características un SGDB se enfoca en sus procesos esenciales como son la manipulación y construcción de las bases de datos, así como la definición de los mismos. Además, estas características facilitan el cumplimiento de una serie de funciones relacionadas como: definición de los datos, su fácil manipulación, una rápida gestión, poder representar

relaciones complejas entre datos y otros aspectos relacionados con la seguridad y validez de los datos(PowerData, 2015).

2.1.2. Bodega de datos (Data Warehouse)

Se denomina bodega de datos a la colección de estos que son integrados para que sean no volátiles, variante en el tiempo y orientados a temas que den soporte a la toma de decisiones empresariales.

Mediante una bodega de datos se procede a integrar datos de diferentes fuentes y obtener datos consolidados que pueden ser almacenados en un dispositivo de memoria no volátil.

Los datos extraídos e integrados se suelen someter a transformaciones para eliminar las inconsistencias y resumir la información, con el propósito de tomar decisiones en función de mejorar la gestión del negocio a partir de datos depurados.

Para consolidar los datos se utiliza un proceso estandarizado denominado ETL (Extracción, Transformación y Carga de datos), que lleva a cabo un conjunto de procedimientos necesarios para la adecuada alimentación de los datos históricos de una bodega, y cargarlos en una nueva base de datos (Parraga & Zaldumbide, 2018).

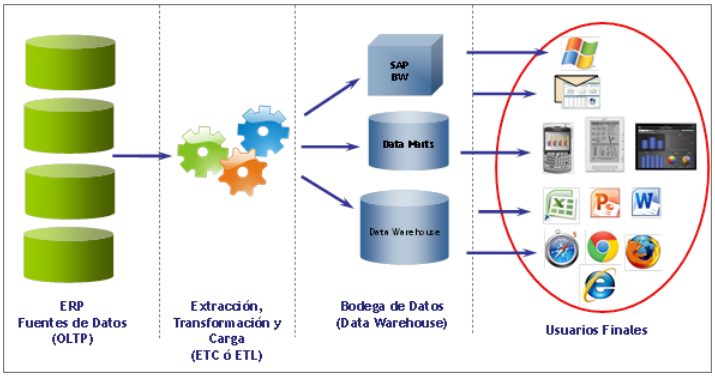


Figura 2.Proceso ETL
Fuente:(Parraga & Zaldumbide, 2018).

Las metodologías más utilizadas para la creación de una bodega de datos son la de Kimball e Immon en las que se indica cómo realizar el diseño y creación de un DW, aunque existen metodologías impuestas por los fabricantes de software de inteligencia de negocios con sus productos. La más a fin al proyecto a desarrollar es la de Kimball que propone como tareas para el diseño e implementación de una bodega de datos los mostrados en la figura 4.

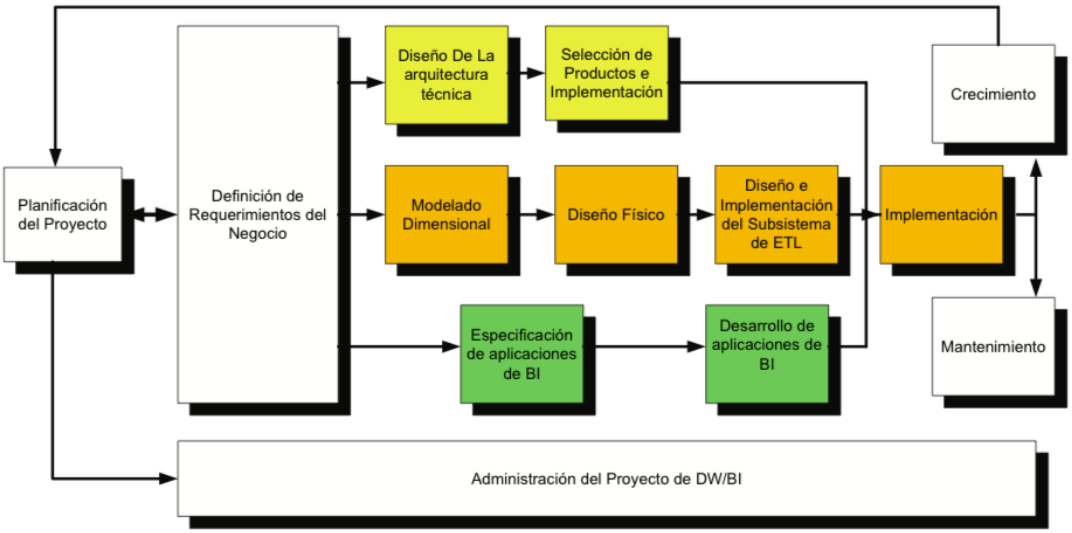


Figura 3.Tarea metodología Kimball
Fuente:(Parraga & Zaldumbide, 2018).

2.1.3. Modelos de bases de datos multidimensionales

Las bases de datos multidimensionales (MDB) se utilizan generalmente para crear aplicaciones OLAP(*On-Line AnalyticalProcessing*)cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Están constituidas de varias tablas de hechos y de dimensionesque contienen datos resumidos de grandes bases de datos o Sistemas Transaccionales.

Estas bases de datos se han optimizado para data warehouse y aplicaciones de procesamiento analítico en línea y generalmente se crean usando entradas de las bases de datos relacionales existentes. Mediante ellas, es posible procesar rápidamente los datos en la base de datos a fin de

que las respuestas se pueden generar rápidamente utilizando la idea de un cubo de datos para representar las dimensiones de los datos disponibles para un usuario.

Este tipo de base de datos facilitan el análisis para el negocio ya que permiten extraer datos de forma selectiva y realizar consultas de distinto tipo, aunque no es posible modificar la estructura de estas bases de datos multidimensionales, por lo que cuando sea preciso introducir cambios, habrá que diseñarlos de nuevo.Se usan en informes de negocios de ventas, marketing, informes de dirección, minería de datos y áreas similares(Rouse, Margaret, 2015).

2.1.4. Minería de datos

Es un proceso por el cual se pretende determinar patrones de comportamiento de una gran cantidad de datos. La minería de datos combina la estadística, las bases de datos y la inteligencia artificial para descubrir automáticamente situaciones interesantes en un mar de datos. El análisis de minería de datos se lleva a cabo con dos actividades para obtener conocimiento no conocido:

a) Describir en detalle a los generadores de datos.

b) Predecir su comportamiento en su entorno.

Para descubrir el conocimiento mediante análisis de minería de datos se utiliza la historia almacenada en la bodega de datos. De acuerdo al comportamiento de los generadores datos se puede ayudar a las personas que toman decisiones a identificar futuras situaciones deseadas o no deseadas, aun con datos faltantes, y poder indicar el valor de éstos con cierta certidumbre (Martínez, 2011).

2.1.5. Técnicas de minería de datos

La Minería de Datos se podría abstraer como la construcción de un modelo que ajustado a unos datos proporciona un conocimiento. Es así que se definen dos procedimientos a seguir: elección del modelo y ajuste final de éste a los datos.

La elección del modelo se determina de acuerdo al tipo de los datos y el objetivo que se quiera obtener. Mientras que la relación del modelo con el objetivo depende del nivel de comprensibilidad que se quiera obtener del modelo final.

El segundo paso consiste en realizar una "fase de aprendizaje" con los datos disponibles para ajustar el modelo anterior a nuestro problema particular, buscando un modelo con valores que intenten aumentar la eficiencia del mismo.

Las técnicas más representativas que se utilizan en la minería de datos son:

Las redes neuronales: Es una técnica basada en el funcionamiento del sistema nerviosos conformado por un paradigma de aprendizaje y un procesamiento automatizado que permite interconectar las neuronas en una red (red neuronal) que presta colaboración para la producción de estímulos de salida.

Regresión lineal: Es aquella en la que se forman relaciones entre los datos por cuanto es una de las técnicas más utilizada. Además, incorpora un sistema eficaz y rápido, aunque cuenta con insuficiencias cuando se requiere relacionar más de 2 variables.

Árboles de decisión: Técnica utilizada en el campo de la inteligencia artificial mediante la cual partiendo de una base de datos se construyen diagramas de construcciones lógicas. Es un sistema basado en reglas para representar condiciones sucesivas que dan solución a un problema similar a la predicción.

Modelos estadísticos: Técnica que utiliza una expresión simbólica para identificar los factores que modifican la variable de respuesta utilizada para diseños experimentales y en la regresión.

Agrupamiento: Se basa en la agrupación de determinados criterios para formar vectores de entrada de los cuales de acuerdo a su disposición se los agrupará en base a los que estén más cercanos porque tienen características comunes.

Según el objetivo que tenga la realización del análisis, los algoritmos se pueden clasificar como algoritmos supervisados, que predicen un dato desconocido inicialmente a partir de otros datos que son de conocimiento previo. Y los algoritmos no supervisados, los cuales descubren patrones y tendencias que se presentan los datos(Martínez, 2011).

2.1.6. Herramientas ETL

Actualmente, las empresas crean una gran cantidad de datos e información que es preciso recolectar y analizar. En un *Business Intelligence* (BI) el proceso de recolectar dicha información conlleva entre el 60 y el 80% del tiempo, por cuanto es importante contar con herramientas ETL para que todo el proceso de BI se lleve a cabo correctamente.

ETL (del inglés *extract, transform and load*) es el proceso por el cual se extrae datos de diferentes fuentes y con distintos formatos, se validan, limpian y transforman para ser analizados de una manera sencilla. Finalmente, los datos son cargados en una nueva base de datos, data warehouse o data mart, donde se encuentran listos para ser explotados, según los objetivos del negocio.

De acuerdo al tamaño de los datos un proceso ETL puede llegar a ser muy complejo, es así, que las herramientas ETL juegan un papel fundamental ya que son la base para cualquier

estrategia de análisis de datos y de inteligencia de negocio.Su uso reporta a la empresa una gran cantidad de beneficios:

- Gobernabilidad de datos.
- Generación de documentación apta para la toma de decisiones.
- Detección, análisis y corrección de errores encontrados en la base de datos.
- Posibilidad de conectores disponibles para mejorar su capacidad.
- Integración con otras herramientas de Business Intelligence.

Sin embargo, no siempre es sencillo saber cómo elegir la herramienta correcta y que mejor se adapte a nuestros objetivos debido a que hay distintas herramientas ETL en el mercado, cada una con sus características concretas.

En la actualidad se puede diferenciar 4 categorías:

Tabla 1

Categorías de Herramientas ETL

Herramientas ETL Enterprise.	• Productos propietarios. • Muchas funcionalidades incluidas. • Soporte para conexión con una gran cantidad de fuentes. • coste de adquisición es elevado.
Herramientas ETL open source.	• Herramientas de código libre. • Uso gratuito. • Mayor accesibilidad para empresas pequeñas. • Productos con un enfoque general. • Requiere consultoría especializada para que se adapten a objetivos.
Herramientas ETL personalizadas.	• Herramientas desarrolladas a medida. • Específica para una empresa o proyecto en concreto. • Requieren un grande esfuerzo inicial de desarrollo. • Resultado se ajusta mejor a los requerimientos.
Herramientas ETL Cloud.	• Alta flexibilidad. • Pago por uso se ofrecen como servicio.

Es así, que para escoger la mejor herramienta es importante tener claro para qué nos va a servir, ya que cada herramienta tiene sus puntos fuertes que pueden hacer que encaje con el proyecto y presupuesto.

Para comparar las herramientas ETL se pueden considerar las siguientes características:

- El coste, que debe incluir costos de adquisición, soporte, formación y consultoría para decidir entre una herramienta propietaria o de código libre.
- El riesgo,para poder asegurar que el proyecto tenga éxito, lo que debe considerar cumplir con el presupuesto, el calendario o con los requerimientos o expectaciones de los clientes.
- La facilidad de uso, en donde debe considerarse si la herramienta dispone de una interfaz gráfica amigable, quereduciría el tiempo de aprendizaje.
- El soporte y la atención al cliente. En este sentido hay que tener en cuenta si se ofrece en varios idiomas y países.
- Los requerimientos de despliegue de la herramienta, lo que incluye la compatibilidad con las distintas plataformas y sistemas operativos, así como los requisitos de sistema en cuanto a hardware.
- La velocidad, que depende en larga medida de la cantidad de datos que hay que transferir a través de la red y de la capacidad de cálculo requerida para las transformaciones.
- La calidad de datos, quizás la característica más importante de las herramientas ETL ya que permite disponer de datos validados y limpios.

- Herramientas de control, que permiten identificar y solucionar los problemas a lo largo de la fase de desarrollo y después.

Una referencia importante para saber cuáles son las mejores herramientas ETL, es el Cuadrante Mágico de Gartner, que cada año indica cuáles son los proveedores leader del mercado, entre los cuales se encuentran:

Figura 4Cuadrante Gartner Herramientas ETL
Fuente:(Gartner, 2018).

Informática: líder según Gartner. Su suite empresarial de integración de datos incluye la solución PowerCenter, una de las más populares.

IBM: proporciona la suite de soluciones InfoSphere, en la cual destaca su herramienta DataStage.

Talend: conocido por su software de integración de código abierto gratuito Open Studio.

SAP: ofrece la herramienta ETL Data Services como parte de SAP BO (Business Objects)

SAS: proporciona una solución de integración de datos llamada Data Management

Oracle: proporciona la herramienta ELT Data Integrator, que permite gestionar procesos de integración de datos en sistemas de inteligencia de negocio(Carisio, Emanuele, 2018).

2.2.Antecedentes del estado del arte

En el presente estado del arte se consideraron las fases de un estudio de mapeo sistemático SMS[1], mismo que mediante criterios de inclusión y exclusión de búsqueda ayudaron a determinar los estudios relevantes entorno al trabajo a desarrollar. Como fuentes de búsqueda de la información para la investigación se usaron los siguientes repositorios académicos: Scopus, Springer, IEEExplore y ACM Digital Library.

2.2.1. *Definición de objetivo*

El objetivo del estudio del estado del arte está enfocado en resolver las preguntas de los objetivos específicos planteados.

2.2.2. Definición de los criterios de inclusión y exclusión

Al momento de buscar información sobre un tema específico en los repositorios de búsqueda se puede encontrar un sinfín de resultados que no permiten determinar cuáles aportarán a la investigación. Es por esto por lo que se realizó una revisión de dichos resultados considerando los siguientes criterios:

- Criterios de inclusión

Buscando obtener estudios actuales, basados en casos similares al planteado se consideraron artículos a partir del 2014.

[1]SystematicMappingStudy (SMS): estudio de alcance que analiza un amplio conjunto de estudios primarios (artículos, publicaciones) para identificar qué y cuantas evidencias hay disponibles sobre un determinado tema.

Se analizaron únicamente artículos científicos y documentos de conferencias publicados en el idioma inglés y español

En su mayoría se tomaron en cuenta artículos científicos y documentos de conferencias.

Se tomaron en cuenta artículos que apliquen la inteligencia de negocios para la predicción de riegos crediticios.

Se revisaron artículos que realices minería de datos en estudios de datos de bancos e instituciones financieras.

- Criterios de exclusión

Artículos que tengan temas de inteligencia de negocios no relacionados con instituciones financieras.

Artículos que no consideren como fase principal la minería de datos.

2.2.3. Definición de la estrategia de búsqueda

Revisión inicial: En base a las preguntas de investigación planteadas se realizó una revisión inicial de artículos en los repositorios de búsqueda digital.

Validación cruzada de estudios: Con el propósito de encontrar el listado inicial de los artículos se procedió a revisar aquellos que cumplan los criterios establecidos anteriormente. Con estos estudios se procedió a pasar a la siguiente fase.

Integración del grupo de control: En esta fase se procedió a realizar un análisis inicial del título de los estudios, introducción, conclusiones y palabras claves de los artículos seleccionados en la revisión inicial. Estos artículos permitieron definir grupos de control definidos de la siguiente manera:

Tabla 2

Estudios por Grupo de Control

Grupo Control	Titulo	Palabras Clave
EC1	Credit scoring using cart algorithm and binary particle swarm optimization	Credit scoring, Data mining,
EC2	Credit risk assessment for unbalanced datasets based on data mining, artificial neural network and support vector machines	Artificial intelligence, Credit risk, Credit scoring, Data mining, Unbalanced data
EC3	Some methods for estimating financial risks in banking	Altman mode, Backpropagation, Beta index, Financial risks, Linear regression, Neural networks, Probability of bankruptcy, Z-score
EC4	Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method	Banking, Credit scoring, Customer behaviour analysis, Data mining, Neural networks, RFM analysis method
EC5	Feature selection in credit scoring model for credit card applicants in XYZ bank: A comparative study	Data Mining, Bank, Credit card, Credit scoring, Feature selection
EC6	A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring	Classification, Credit scoring, Data mining, Ensemble learning, Feature selection

2.2.4. Construcción de la cadena de búsqueda

En esta fase se construyó la cadena de búsqueda que permitió encontrar los estudios primarios, de acuerdo con los grupos de control se verificó las palabras claves que más se repitan y se definió contextos de análisis. Para el presente estudio se definieron los siguientes contextos:

Tabla 3
Construcción de cadenas de búsqueda

Contexto	Palabra Clave	EC1	EC2	EC3	EC4	EC5	EC6	Número de Repeticiones
Algoritmos de Inteligencia de Negocios	Data mining	x	x		x	x	x	5
	Artificial intelligence	x						1
	Linear regression			x				1
	Neural networks			x	x			2
	Classification						x	1
Análisis Crediticio	Creditscoring	x	x		x	x	x	5
	Creditrisk		x					1
	Financialrisks			x				1
Entorno de Análisis	Unbalanced data		x					1
	Customerbehaviouranalysis				x			1
	Banking				x	x		2
	Creditcard					x		1

La cadena de búsqueda está formada por la unión de las palabras claves que más se repiten en cada contexto, los conectores usados son OR para las palabras que están dentro del mismo contexto y el conector AND para las palabras que están en contextos distintos, de esta manera se establece la siguiente cadena de búsqueda.

(Data Mining OR Neural networks) AND Credit scoring AND Banking

Luego de definir la cadena de búsqueda se procedió a buscarla dentro de los repositorios de búsqueda digital, considerando además filtros como: cuya fecha de publicación sea mayor al

2014, que sean artículos científicos o documentos de conferencias y que tengan asociado la palabra clave Data Minig.

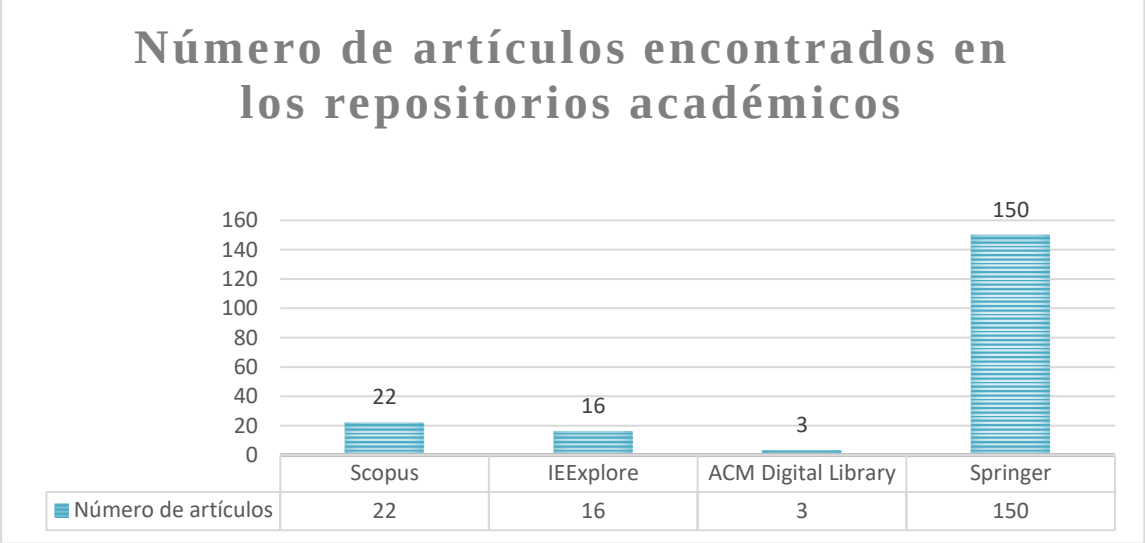


Figura 5. Artículos encontrados en los repositorios académicos.

2.2.5. Artículos primarios

Una vez obtenidos los resultados se realizó la revisión de los documentos encontrados verificando a nivel de resumen y contenido si ayudaran en la investigación propuesta, definiendo como principales los que se listan a continuación:

(Khemakhem & Boujelbene, 2018) Predicting credit risk on the basis of financial and non-financial variables and data mining

En este artículo se analizan datos académicos del desempeño de los estudiantes de pregrado seleccionados al azar. Se utiliza estadística descriptiva y distribuciones de frecuencia de los datos de rendimiento académico en tablas y gráficos para facilitar la interpretación de los datos. Además, se realizan análisis de varianza de una vía (ANOVA) y pruebas de comparación múltiple post hoc para determinar si las variaciones en los rendimientos académicos son significativas. Los datos proporcionados en este artículo ayudarán a la comunidad de

investigación educativa global y a los responsables de la política regional a comprender y optimizar el entorno de aprendizaje hacia la realización de campus inteligentes y la educación sostenible.

(Serrano-Cinca & Gutiérrez-Nieto, 2016)The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending

Este estudio va más allá de los sistemas de calificación crediticia de préstamos peer-to-peer (P2P) al proponer una puntuación de ganancias. Los sistemas de calificación crediticia estiman la probabilidad de incumplimiento del préstamo, en este estudio los autores se enfocan en predecir la rentabilidad esperada de invertir en préstamos P2P, medida por la tasa interna de rendimiento. Se analizan los factores que determinan la rentabilidad del préstamo, y se encuentra que estos factores difieren de los factores que determinan la probabilidad de incumplimiento. Los resultados muestran que los préstamos P2P no son actualmente un mercado totalmente eficiente. Esto significa que las técnicas de extracción de datos son capaces de identificar los préstamos más rentables, o en la jerga financiera, "ganarle al mercado". En la muestra analizada, se encuentra que un prestamista que selecciona préstamos aplicando un sistema de calificación de ganancias usando una regresión multivariable supera a los resultados obtenidos mediante el uso de un sistema de calificación crediticia tradicional, basado en regresión logística.

(Alborzi & Khanbabaei, 2016) Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method

Este estudio considera que la calificación crediticia es una de las principales actividades en los bancos y otras instituciones financieras y que actualmente se utilizan técnicas de minería de datos y el método de análisis RFM para ayudar a los bancos a desarrollar sistemas de análisis del

comportamiento del cliente y calificación crediticia. En este documento, se presenta un nuevo modelo híbrido de calificación de comportamiento y calificación de crédito basado en técnicas de minería de datos y redes neuronales para el campo de la banca. En este modelo híbrido, se desarrolla un nuevo método de análisis WRFMLC mejorado utilizando técnicas de agrupación y clasificación. Los resultados demuestran que el modelo propuesto se puede implementar para segmentar y clasificar efectivamente a los clientes bancarios valiosos.

(Koutanaei, Sajedi, & Khanbabaei, 2015) A hybrid data miningmodel of featureselectionalgorithms and ensemblelearningclassifiersforcreditscoring

El presente estudio desarrolló un modelo de minería de datos híbrido de algoritmos de selección de características y clasificación de aprendizaje en conjunto sobre la base de tres etapas. La primera etapa, como se esperaba, trata sobre la recolección de datos y el pre procesamiento. En la segunda etapa, se emplean cuatro algoritmos FS, para la implementación del algoritmo de clasificación de la máquina de vectores de soporte (SVM). Después de elegir el modelo apropiado para cada característica seleccionada, se aplican a los algoritmos de clasificación de base y conjunto. En esta etapa, se indica el mejor algoritmo FS con su configuración de parámetros para la etapa de modelado del modelo propuesto. En la tercera etapa, los algoritmos de clasificación se emplean para el conjunto de datos preparado a partir de cada algoritmo FS. Los resultados mostraron que, en la segunda etapa, el algoritmo PCA es el mejor algoritmo FS. En la tercera etapa, los resultados de la clasificación mostraron que el método de refuerzo adaptativo de la red neuronal artificial (ANN) (AdaBoost) tiene una mayor precisión de clasificación. En última instancia, el documento verificó y propuso el modelo híbrido como un modelo operativo y sólido para realizar la calificación crediticia.

(Gahlaut et al., 2017) Prediction analysis of risky credit using Data mining classification models

En este documento, se verifica si las técnicas de extracción de datos son útiles para predecir y clasificar el puntaje crediticio del cliente (bueno / malo) para superar los riesgos futuros de otorgar préstamos a clientes que no pueden pagar. Se usó el conjunto de datos históricos de un banco para el modelo predictivo (modelos generales), los bancos pueden usarlos para el mejor resultado de su sistema de crédito general. Por ejemplo, si a un cliente se le asigna un puntaje de crédito malo después de aplicar estos modelos de clasificación predictiva, entonces el banco no permitirá otorgarle un crédito futuro a ese cliente y analizará rápidamente todos los demás créditos de riesgo.

(Lohokare, Dani, & Sontakke, 2017) Automated data collection for credit score calculation based on financial transactions and social media

Actualmente la credibilidad de un usuario bancario se basa en el "puntaje crediticio" de la persona que se calcula a partir del desempeño anterior de esta en obligaciones de deuda. Este documento proporciona una solución alternativa única para recopilar estos datos. Aprovechando el hecho de que casi todo el mundo tiene teléfonos inteligentes en la actualidad, puede haber una aplicación para teléfonos inteligentes que recopile todos estos datos y los envíe al organismo oficial. Este documento propone acceder a los datos de las redes sociales para obtener información sobre el estado social general de una persona. La solución propuesta tiene una aplicación de teléfono inteligente que captura datos de transacciones bancarias y datos relacionados con compras en línea a través de SMS. El uso de redes neuronales artificiales permitió calcular el puntaje de credibilidad final en función de los diversos parámetros de datos recopilados.

(Okesola, Okokpujie, Adewale, John, & Omoruyi, 2017) An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach

La calificación crediticia es una herramienta de decisión utilizada por las organizaciones para otorgar o rechazar solicitudes de crédito de sus clientes. Se han utilizado una serie de enfoques artificiales inteligentes y tradicionales para construir un modelo de calificación crediticia y una evaluación del riesgo crediticio. A pesar de estar clasificado entre los 10 mejores algoritmos en minería de datos, el algoritmo BayesianNaive no se ha utilizado ampliamente en la construcción de tarjetas de puntuación de crédito. Utilizando indicadores demográficos y materiales como variables de entrada, este documento investiga la capacidad del clasificador bayesiano para construir un modelo de calificación crediticia en el sector bancario.

(Shi, 2012) China's National Personal Credit Scoring System: A Real-Life Intelligent Knowledge Application

El Centro de Referencia de Créditos (CRC) del Banco Popular de China (PBC) ha creado la mayor base de datos de crédito personal en el mundo con 800 millones de cuentas de personas en China desde 2003. Se desarrolló el Sistema Nacional de Calificación de Créditos Personales de China, conocido como "Puntuación de China", que es un Aplicación KDD única y avanzada bajo la gestión inteligente del conocimiento en estos grandes datos. El sistema finalmente servirá a todos los 1.300 millones de habitantes de China para sus actividades financieras diarias. En este artículo se presentan los componentes clave del proyecto China Score que incluye objetivos, proceso de modelado, técnicas KDD utilizadas en los proyectos, gestión inteligente del conocimiento y experiencia del desarrollo del proyecto. Además, se describe una serie de recomendaciones de políticas basadas en el proyecto China Score, que ha tenido un impacto

potencial en el gobierno chino en su toma de decisiones estratégicas para el desarrollo económico de China.

2.2.6. Conclusión

Luego de realizar este análisis de literatura se pudo visualizar como los autores se enfocan en la misma problemática de la investigación propuesta como es buscar métodos para evaluar el riesgo de crédito, teniendo en cuenta no solo las variables financieras y no financieras, sino también el desequilibrio de clase. De esta manera ellos estiman la probabilidad de incumplimiento de préstamos. Además, en otros casos usan sistemas de análisis del comportamiento del cliente y calificación crediticia, para de esta manera predecir y clasificar el puntaje crediticio del cliente (bueno / malo) y superar los riesgos futuros de otorgar préstamos a clientes que no pueden pagar. Los estudios antes mostrados utilizan técnicas como redes neuronales, arboles de decisión, regresión logística, utilizan técnicas de minería de datos y el método de análisis RFM, NaiveBayesian, entre otros para realizar el análisis de tos históricos de entidades bancarias y en otro caso de comportamiento de los usuarios usando sus teléfonos inteligentes. Los resultados obtenidos fueron factores claves para predecir el incumplimiento, una calificación crediticia, segmentación y clasificación efectiva de los clientes bancarios valiosos, puntajes de credibilidad, entre otros, que están enmarcados dentro de los que se quiere conseguir en el análisis propuesto. En el que se buscó encontrar un modelo que detecte los factores para calificación crediticia óptima y en base a datos históricos de créditos otorgados en la cooperativa permita otorgar créditos a buenos pagadores que no posean un buró crediticio.

2.3.Metodología de investigación

Las organizaciones requieren contar con modelos que permitan evidenciar patrones de comportamiento útiles para optimizar sus procesos, sin embargo, es necesario un lineamiento oportuno en cuanto a metodologías, modelos y herramientas aplicables (Benalcazar & Vinueza, 2017).

El proyecto está enfocado a la realización de un modelo de minería de datos por cuanto utilizará una metodología de investigación AD-HOC (propia), para asegurar el éxito en el desarrollo de este.

Esta metodología consideró las siguientes fases:

2.3.1. Evaluar herramientas y métodos

Fase en la que se determinaron las herramientas y métodos que ayudarán a determinar el comportamiento de cartera crediticia.

Se consideró que la fase más importante de la investigación desarrollada es la creación del modelo de minería de datos por cuanto se realizó un análisis de las herramientas de minería de datos existentes en el mercado para determinar la más óptima para el trabajo desarrollado.

Las herramientas de minería de datos permiten realizar el análisis de patrones y relaciones en los datos de acuerdo a los requerimientos del cliente. Un software de minería de datos puede permitir: crear clases de información, identificar asociaciones y patrones secuenciales, entre otros, paraen base a estos, sacar conclusiones sobre las tendencias en el comportamiento de los datos.

Generalmente, el proceso de minería de datos consta de pasos como: recopilar datos que son cargados en sus almacenes de datos, a continuación, se almacenan los datos en servidores

internos o en la nube, de esta manera los analistas de negocios, los equipos de administración y los profesionales de TI tienen acceso a los mismos. Finalmente, el usuario final presenta los datos en una vista accesible, como un gráfico o una tabla(ITpedia, 2018).

A continuación, se realiza la comparación entre algunas herramientas para minería de datos:

Herramientas de Minería de Datos	Tipo de Software	Plataforma	Arquitectura	Algoritmos	Tipo de modelo
MD(*Data Mining*)					
Weka	Libre	Todas las plataformas		Clustering,Regresion	Predictivo
Clemetine	Libre	Windows, Linux	Cliente/servidor	Red neuronal, GRI A priori, logística, QUEST, CHAID, KARMA	Predictivo
Knime	Libre	Windows, Linux, Mac Os		Algoritmos segmentación, árboles de decisión, redes neuronales, SVM	Predictivo
IBM SPSS	Comercial	Windows, Linux		Ecuaciones estructurales	Predictivo
RapidMiner	Libre	Windows, Linux	Cliente/servidor	Clustering, arboles de decisión, redes neuronales	Predictivo

Figura 6. Comparación de Herramientas de Minería de Datos
Fuente:(Lara Hernández et al., 2014)

Luego de realizar un análisis de la mismas se decidió utilizar KNIME (KonstanzInformationMiner), que es una herramienta de software libre que permite el desarrollo de modelos sobre la plataforma Eclipse y es extremadamente flexible y potente. Su versión inicial fue lanzada en Julio de 2006, la versión actual es la 3.7.2, pero se encuentra en continuo desarrollo por los creadores de la misma en Konstanz (Alemania) y toda la comunidad que quiere participar.

Figura 7. Logo herramienta KNIME
Fuente: (Biosilveit, 2016)

Knime es una plataforma analítica y modularde exploración de datos, que permite alusuario crear flujos de datos, de forma visual e intuitiva para manipularlos mediante workflows en base al modelo de nodos quese conectan entre si y hacen cosas.Además, esta herramienta,permite ejecutar de forma selectiva algunos de los pasos creados, así como,todo el flujo desarrollado. Tras la ejecución, los resultados se puedeninvestigar mediante vistas interactivas tanto de los datos como de los modelos.

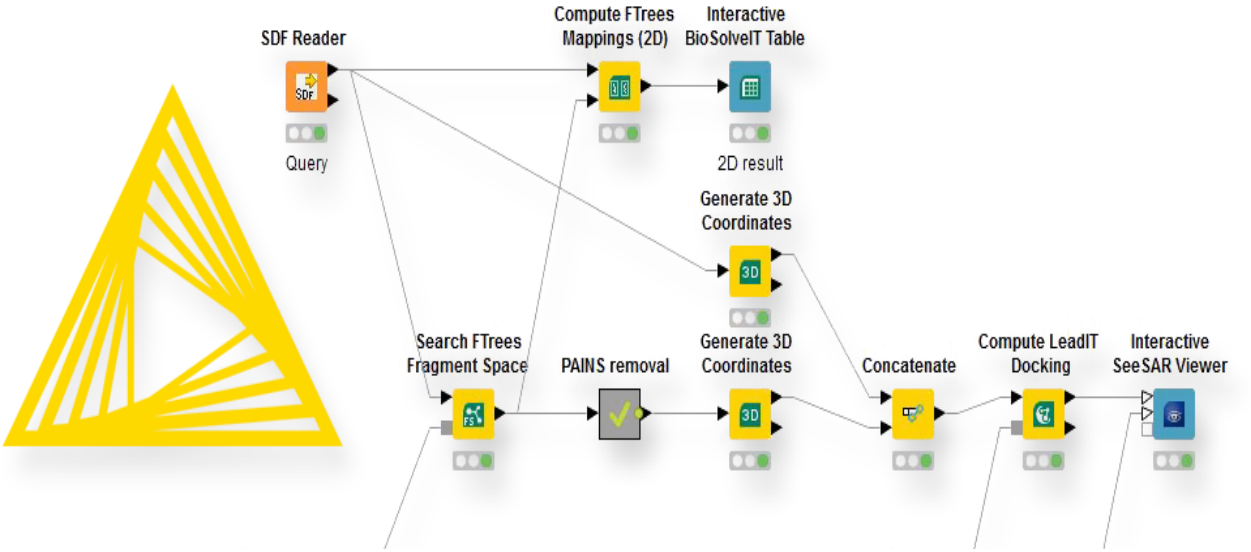


Figura 8. Workflow en KNIME
Fuente: (Biosilveit, 2016)

Además, una de sus principales bondades es su posibilidad de ampliación y de conexión con otras herramientas para poder utilizar las funcionalidades de las mismas como son: Weka, Python, R, Tableau, entre otros.

Los operadores que presenta la herramienta están agrupados en 12 grupos:

- I/O: Nodos para procesos de extracción y carga de datos desde diferentes orígenes de datos.
- *Manipulation:*Nodos para el análisis y transformación delos datos.
- *Views:* Nodos que presentan herramientas para visualización de tablas, gráficos entre otros.
- *Analytics:* Nodos para realizar minería de datos, cálculos estadísticos entre otros.
- *Database:* Nodos para realizar operaciones y conexiones con diferentes gestores de bases de datos.
- *Other Data Types:* Nodos para operaciones con otras bases de datos.
- *Structured Data:* Nodos para realizar flujos utilizando JSON y XML.
- *Scripting:* Nodos para realizar flujos relacionados con desarrollos en Java.
- *Tools &Services:* Nodos para procesos desarrollados con Web Services.
- *KNIME Labs*: Nodos para operaciones con bases de datos MongoDB.
- *Workflow Control:* Nodos para automatizar los flujos creados.
- *Reporting:* Nodos para crear reportes.

En la Fig. 9 se puede observar los operadores que ofrece la herramienta y su entorno que es bastante amigable e intuitiva.

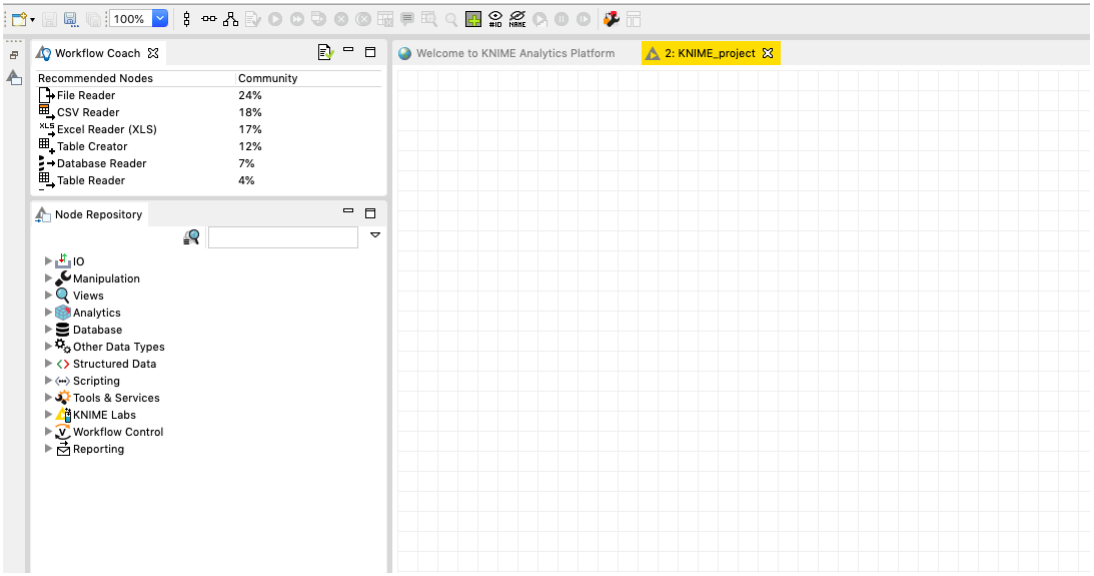


Figura 9. Entorno de trabajo KNIME

Mediante estos operadores, se puede observar que la herramienta además de permitir realizar la minería de datos, también brinda la posibilidad de realizar los procesos ETL, por lo que todo el proceso de este trabajo se lo desarrolló en KNIME.

2.3.2. Diseño del modelo

Aquí se realizó la recolección y selección inicial de los datos. Además, en esta fase se describieron, exploraron y verificaron los datos con los que se trabajó en el proyecto. Se identificó el conjunto de datos con el que se trabajó creando una bodega de datos. Se realizó la revisión de los parámetros de configuración para crear y evaluar el modelo que determine los patrones de comportamiento de los buenos pagadores de crédito.

Durante esta fase se recolectó información de la cooperativa, la misma quecontenía 7 archivos de Excel:

- C3. anexo detallado de cartera al 31/12/2018

- Base de datos clientes
- Solicitudes crédito rechazados
- Socio estudio mercado 18/10/2018
- Créditos desembolso diario 25/03/2019
- Cartera crédito
- C5.1 anexo detalle créditos castigados 2018

Luego de analizar cada uno de los archivos se procedió a realizar el diseño de un modelo multidimensional, considerando los atributos que estaban directamente relacionados con el objetivo del proyecto. En base al modelo creado, se realizó el respectivo proceso ETL para creación de la bodega de datos que se cargó en una nueva base de datos.

Con la data ya procesada, transformada y optimizada se procedió a analizar el comportamiento actual de los buenos pagadores de créditos en la cooperativa para de esta manera compararlos con los resultados obtenidos al final.

Desde la nueva base creada desde los procesos ETL se procedió a crear el modelo de minería de datos considerando las técnicas de minería de datos que mejor se acoplen al caso de investigación, donde se obtuvo como resultado un modelo de minería de datos que determina el comportamiento de un buen pagador en la cooperativa caso de estudio considerando un nivel de confianza en un rango aceptable.

2.3.3. Implementación del modelo

Una vez que el modelo se construyó y validó, se pasó a la fase de transformacióndel conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados o que el modelo sea

usado directamente por el personal de créditos en el proceso de otorgamiento de créditos en la cooperativa.

2.3.4. Validación del modelo

Mediante esta fase se evaluaron los resultados obtenidos basados en los requerimientos de La Cooperativa de Ahorro y Crédito, en esta evaluación se consideró el estado anterior del proceso de concesión de créditos y el estado propuesto para así proponer acciones e implementaciones en futuras concesiones de créditos.

CAPÍTULO III

ANÁLISIS Y DISEÑO

3.1.Comprensión del negocio

"Es una organización jurídica que realiza actividades de intermediación financiera y de responsabilidad social con sus socios y previa autorización de la Superintendencia de Economía Popular y Solidaria con socios y/o terceros con sujeción a las regulaciones u a los principios reconocidos en la Ley Orgánica de la Economía Popular y Solidaria y del Sector Financiero Popular y Solidario, a su reglamento General, a las Resoluciones de la Superintendencia de Economía Popular y Solidaria y del ente regulador"(Cooperativa, s. f.).

Esta cooperativa posee 15 agencias distribuidas en el país de la siguiente manera:

- Riobamba: Agencia matriz y 1 agencia
- Quito: 3 agencias
- Guayaquil: 2 agencias
- Guamote: 1 agencia
- Sto. Domingo de los Tsáchilas: 1 agencia
- Santa Elena: 1 agencia
- Alausí: 1 agencia
- Ambato: 1 agencia
- Machala: 1 agencia
- Cañar: 1 agencia
- Latacunga: 1 agencia

3.1.1. Área de créditos

La Cooperativa de Ahorro y Crédito ofrece a sus clientes el servicio de créditos mediante su área correspondiente donde las personas pueden acceder a diferentes tipos de préstamos como:

Microcrédito: Dirigido a microempresarios que cuenten con unidades de producción, comercio y servicios: para iniciar y/o ampliar su negocio, con montos de hasta 80000 dólares y con plazos de hasta 48 meses.

Credimóvil: Dirigido para nuestros socios en los diferentes mercados populares del país para capital de trabajo, incremento y aplicación del negocio, con montos de hasta 25000 dólares y con plazos de hasta 48 meses.

Vivienda: Destinado a la compra, construcción, remodelación, ampliación y mejora de la vivienda o adquisición de terreno para la vivienda, con montos de hasta 30000 dólares y con plazos de hasta 15 años.

Consumo: Para adquisición de bienes de consumo, servicios o gastos no relacionados con una actividad productiva, cuya fuente de pago es bajo relación de dependencia, con montos de hasta 35000 dólares y con plazos de hasta 48 meses.

Agropecuario: Dirigido a micro, pequeños y medianos empresarios, que se dediquen a la actividad agropecuaria, con montos de hasta 25000 dólares y con plazos de hasta 48 meses.

Iglesias: Crédito destinado a los socios agrupados en organizaciones de hecho y/0 jurídicas vinculadas con la religión cuyo destino sea la construcción, remodelación, ampliación, compra de bienes y organización eventos religiosos, con montos de hasta 100000 dólares y plazos de hasta 48 meses.

Todos estos créditos requieren como requisito la presentación de documentos como: ser socio, copia de cédula, copia papeleta de votación, planilla de luz, certificado de ingresos y certificado bienes del deudor

3.2. Objetivos de las Cooperativas de Ahorro y Crédito

Las cooperativas de ahorro y crédito o como sus siglas en ingles *Saving and CreditsCooperative* son entidades que tienen como finalidad suplir con las necesidades financieras de untercero, la organización está conformada por personas naturales o jurídicas que unen sus capitalespara formar la cooperativa, en la actualidad se rigen bajo La Superintendencia de la EconomíaPopular y Solidaria y por ende deben sujetarse bajo la presente ley(Guido H. Poveda-Burgos, Edison A. Erazo-Flores y Gabriel J. Neira-Vera, 2017).

Cada país tiene establecidas leyes que rigen la estructura de las cooperativas de ahorro y crédito, que fundamentalmentepretende brindar a sus socios una serie de servicios financierosaccesibles para crear oportunidades de negocio.

El objetivo de las cooperativas de ahorro y crédito es apoyara las personas de lossectores más necesitados a progresar,mediante créditos con tasas moderadas para implementar negocios y así puedan surgir en su vida personal.

Las cooperativas de ahorro y crédito tienen como propósito ofrecer a sus socios laconfianza de poder invertir su dinero con ideas claras y objetivos definidos en un determinado tiempo contemplando los beneficios correspondientes.

Es importante mencionar la misión de estas instituciones es ofrecer un servicio de calidad yrentabilidad financiera y social, que a la vez están comprometidos en el desarrollo

socioeconómicode las zonas de influencia(Guido H. Poveda-Burgos, Edison A. Erazo-Flores y Gabriel J. Neira-Vera, 2017).

3.3. Evaluación de la situación actual

Las cooperativas para otorgar un microcrédito siguen este proceso: la persona que solicita el crédito debe ser socio de la cooperativa, por tanto, debe abrir una cuenta de ahorros y adquirir la cuenta certificados de aportación, de esta manera puede acceder a un crédito.

El asesor de crédito para otorgar un crédito solicita la documentación necesaria como es: una foto tamaño carnet actualizada, copia de la cédula de identidad y papeleta de votación actualizada, planilla del último pago de servicio, certificado de ingresos, bienes del deudor; de ser el caso los mismos documentos para el garante.

Luego se procede analizar el buró de crédito que es consultando desde el aplicativo que emite el score crediticio del socio, seguidamente se analiza la información que respaldan los bienes, ingresos y garantías en función a la actividad económica, el perfil de los socios y el destino al cual se va a dar uso el crédito, posterior a esto se realiza la inspección de campo en el lugar donde labora el socio. Para los créditos de consumo, en la Cooperativa de ahorro y crédito se establece la Tecnología Crediticia Convencional, que se basa en un análisis y énfasis en trabajo de oficina, análisis de información (documentos de respaldo de ingresos) y enfoque en garantía en función a la actividad, el perfil de los socios y el destino principalmente.

Finalmente, se analiza y toma una decisión, de acuerdo a su experiencia personal para dar el visto bueno, y luego defender el otorgamiento del crédito en la reunión del comité de crédito. De esta manera se otorga un crédito a un socio.

El asesor de crédito analizaba la documentación recibida y de acuerdo al buró de crédito otorgaba o no el crédito, sin embargo, este proceso generaba inconvenientes en la recuperación de la cartera e incrementado el tiempo de cobranza, por créditos mal colocados y minimizando el tiempo del asesor de créditos para atender a nuevos créditos.

La Cooperativa de Ahorro y Crédito presentaba un alto índice de morosidad debido a lo difícil que se tornaba la cobranza de créditos que habían sido mal colocados convirtiéndose en algunos casos en incobrables.

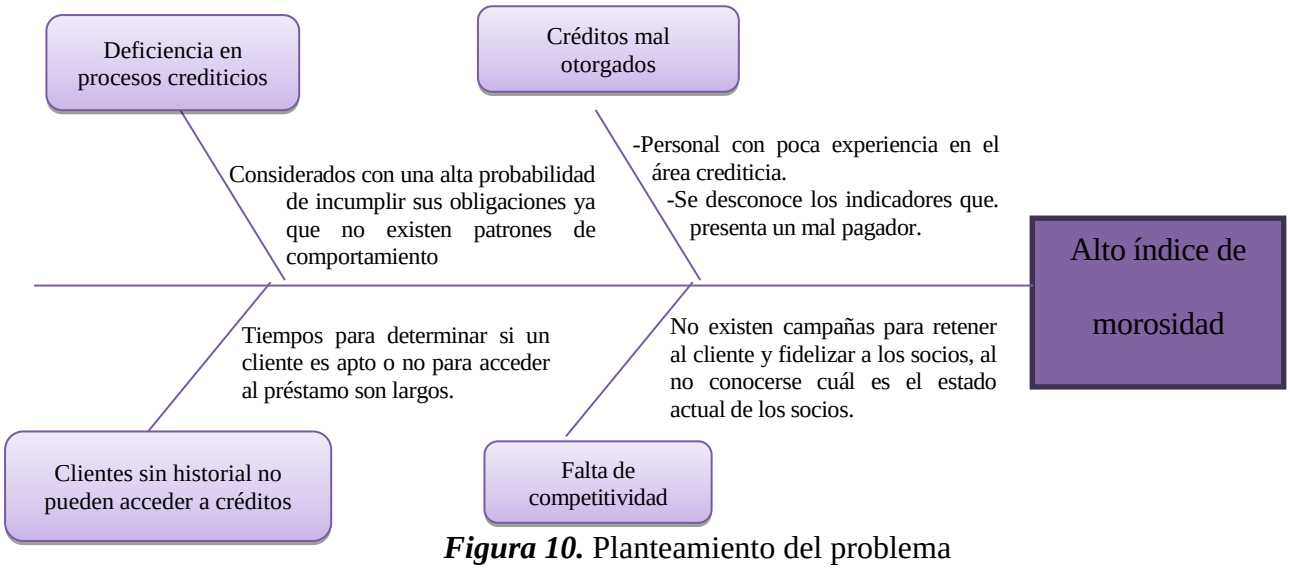


Figura 10. Planteamiento del problema

3.4.Objetivo de minería de datos

- Análisis la situación actual del proceso de otorgamiento de crédito en una cooperativa de Ahorro y Crédito caso de estudios.
- Recopilación información de la cooperativa caso de estudio para su posterior análisis y diseño de una bodega de datos.

- Realizar la extracción de la información obtenida en la cooperativa caso de estudio, transformación en base al diseño de la bodega de datos y carga a una nueva base de datos para la creación del modelo de minería de datos.
- Diseño de un modelo predictivo del comportamiento de la cartera crediticia mediante minería de datos.
- Implementacióndel modelo predictivo de la cartera crediticia mediante la aplicación de patrones.
- Validación de los resultados obtenidos de la aplicación de los patrones del modelo predictivo de la cartera crediticia.

3.5.Evaluación inicial de funciones y algoritmos

Dentro de los procesos que se desarrollan en la cooperativa se consideró la información que la cooperativa receptaba con respecto al ingreso de nuevos socios para de esta manera identificar qué factores se podría encontrar ahí tales como:

- Identificación
- TipoPersona
- NombreCliente
- Sexo
- FechaNacimiento
- Estado Civil
- Instrucción
- Ocupación
- Profesión
- DirecciónDomiciliario
- TipoResidencia
- Teléfonos
- Mail
- ActividadEconómica
- Activo
- Pasivo
- Patrimonio
- Ingresos
- Egresos

Además, se consideró la información receptada con respecto a solicitudes créditos donde se determinaron otros factores como:

- Número Socio
- Identificación
- Nombre Socio
- Tipo Producto
- Tipo Crédito
- Destino Crédito
- Monto
- Plazo
- Forma Pago
- Total, Ingresos
- Total, Egresos
- Patrimonio
- Actividad

También se consideró los datos que la cooperativa generó en base a cada crédito solicitado:

- Fecha Concesión
- Fecha Vencimiento
- Fecha Último Pago
- Días Morosidad
- Tasa Interés
- Tasa Interés Efectiva
- Saldo Crédito
- Cuota Crédito
- Calificación Propia
- Calificación Homologada

De igual manera se consideró información de Créditos Castigados y Rechazados, para determinar más factores relacionados con el modelo.

Con respecto a los algoritmos de minería de datos a utilizar se consideró:

- Árbol de decisión: Esta técnica es considerada una de las más eficaces para la clasificación supervisada, combina técnicas matemáticas y computacionales para ayudar a la descripción, la categorización y la generalización de un conjunto dado de datos y es fácil de entender e interpretar debido a la manera gráfica de presentar los resultados utilizando lógica booleana. Esta técnica es capaz de manejar tanto datos numéricos y categorizados. Grandes cantidades de datos pueden ser analizados utilizando recursos informáticos estándar en un plazo razonable.(CHristianCH., 2018)

- NaiveBayes:Algoritmo que se basa en probabilidades condicionadas con datos conocidos, rápido y simple de usar, por lo que es uno de los clasificadores más usados. Su funcionamiento se basa en calcular probabilidades de datos conocido y de acuerdo a los resultados y una fórmula, se puede calcular la probabilidad de que la entrada sea de una u otra clase.(CHristianCH., 2018)
- Redes Neuronales:Debido a que son una adaptación artificial de lo que hace el cerebro, las funciones son similares a las conexiones neuronales reales: capaces de aprender de la experiencia, generalizar casos anteriores a nuevos casos, abstraer características esenciales a partir de entradas que representan información irrelevante etc.(CHristianCH., 2018)

Estos algoritmos fueron escogidos debido a su popularidad, rapidez, fácil uso y robustez. Además, estos algoritmos permiten manejar datos nominales en sus implementaciones y para el caso de estudio planteado el atributo a predecir es nominal puesto que es: “SI o NO” dependiendo si el crédito fue o no otorgados.

3.6.Selección de la fuente de datos

Luego de determinar los factores que podrían considerarse para la creación del modelo se procedió a solicitar la información disponible en la cooperativa en donde se entregaron 7 archivos Excel:

- C3. anexo detallado de cartera al 31/12/2018
- Base de datos clientes
- Solicitudes crédito rechazados
- Socio estudio mercado 18/10/2018

- Créditos desembolso diario 25/03/2019
- Cartera crédito
- C5.1 anexo detalle créditos castigados 2018

Estos archivos contienen información de datos de socios, cartera de créditos, créditos castigados, rechazados, entre otras. Se procedió a realizar un análisis completo de cada archivo para determinar los atributos relacionados con los factores descritos anteriormente, obteniendo el siguiente análisis:

Tabla 4

Atributos reales y seleccionados del origen de datos: archivo C3. anexo detallado de cartera al 31/12/2018

Atributos reales	**Atributos a considerar**
Numerosocio	
Identificacion	Identificacion
Nombresocio	Sucursal
Sucursal	Numerooperacion
Oficina	Tipoproducto
Numerooperacion	Tipocredito
Tipoproducto	Claseoperacion
Tipocredito	Lineacredito
Claseoperacion	Destinocredito
Lineacredito	Montoconcedido
Destinocredito	Fechaconcesión
Montoconcedido	Fechavencimiento
Fechaconcesión	Estadooperacion
Fechavencimiento	Diasmorosidad
Fechaultimopago	Plazo
Estadooperacion	Saldocredito
Diasmorosidad	Cuotacredito
Plazo	Formapago
Tasainteres	Calificacionpropia
Tasainteresefectiva	Calificacion homologada
Valorvencer1_30	Totalingresos
Valorvencer31_90	Totalegresos
Valorvencer91_180	Patrimonio
Valorvencer181_360	Actividad
Valorvencermas360	Origendeoperacion
Valornodevinteres1_30	
Valornodevinteres31_90	
Valornodevinteres91_180	
Valornodevinteres181_360	
Valornodevinteresmas360	
Valorvencido1_30	

Continúa

Valorvencido31_90 Valorvencido91_180 Valorvencido181_360 Valorvencidomas360 Valorvencido181_270 Valorvencidomas270 Valorvencido91_270 Valorvencido271_360 Valorvencido361_720 Valorvencidomas720 Saldocredito Cuotacredito Formapago Provisionespecifica Provisionconstituida Provisionhomologada Calificacionpropia Calificacion homologada Porcentajeprovision Interesprovisionado Interesordinario Interesmora Valorcarteracastigada Fechacastigo Valordemandajudicial Tipogarantia Valorgarantias Fechacastigo Causalvinculacion Totalingresos Totalegresos Patrimonio Actividad Origendeoperacion	

Tabla 5

Atributos reales y seleccionados del origen de datos: Archivo base de datos clientes

Atributos reales	Atributos a considerar
Sucursal	
Identificacion	Identificacion
Fecha_ingreso	Fecha_ingreso
Numerosocio	Tipo_persona
Tipo_persona	Sexo
Ctipoidentificacion	Fecha_nacimiento
Nombre_cliente	Estadocivil
Sexo	Ocupacion
Fecha_nacimiento	Descripcion_ocupacion
Codigo_pais_nacimiento	Tipo_residencia
Pais	Actividad_economica
Cod_provincia	Total_activo
Codigo_ciudad	Total_pasivo
Codigo_parroquia	Total_patrimonio
Estadocivil	Total_ingresos
Instrucion	Egresos otros_ingresos
Ocupacion	Fecha_actualizacion
Profesion	
Descripcion_ocupacion	
Direccion_domiciliario	
Sector_referencia	
Tipo_residencia	
Telefonodomicilio	
Telefonotrabajo	
Celularpersonal	
Celulartrabajo	
Mail	
Actividad_economica	
Total_activo	
Total_pasivo	
Total_patrimonio	
Total_ingresos	
Egresos	
Otros_ingresos	
Nombre_conyuge	
Actividad_economica_conyuge	
Ingreso_conyuge	
Total_activos_conyuge	
Total_pasivos_conyuge	
Patrimoño_conyuge	
Nombre_apoderado	
Fecha_actualizacion	
Estado	

Tabla 6

Atributos reales y seleccionados del origen de datos: Archivo solicitudes crédito rechazados

Atributos reales	Atributos a considerar
Fechacorta	
Fecha	Fechacorta
Identificacion	Identificacion
Numerosolicitudfit	Numerosolicitudfit
Numerosolicitudmsbi	Monto
Monto	Sucursal
Oficial	Ctipopersona
Sucursal	Fsolicitud
Comentario	Cestatussolicitud
Cpersona_compania	Plazo
Csolicitud	
Ctipopersona	
Nombrelegal	
Nombrecuenta	
Csucursal	
Coficina	
Cusuario_oficialcuenta	
Csubsistema	
Descripcion_subsistema	
Cidioma	
Cmoneda	
Fsolicitud	
Cestatussolicitud	
Cgrupoproducto	
Descripcion_grupoproducto	
Siglas_grupoproducto	
Cproducto	
Descripcion_producto	
Siglas_producto	
Ctipobanca	
Descripcion_tipobanca	
Ctiposegmento	
Descripcion_tiposegmento	
Monto	
Plazo	
Tasa	
Verificadatos	
Verificadocumentos	
Comentariosverificacion	
Cusuario_verificadatos	
Cusuario_verificadocumentos	
Fverificadatos	
Fverificadocumentos	
Cusuario_autorizador	
Cautorizacion	
Identificacion	
Cpersona	
Numerosocio	

Tabla 7

Atributos reales y seleccionados del origen de datos: Archivo Socio estudio mercado 18/10/2018

Atributos reales	Atributos a considerar
Cpersona	Identificacion
Numerosocio	Provincia
Ctipoidentificacion	Canton
Identificacion	Nivel_educacion
Nombrelegal	Profesion
Fnacimiento	Estat_cuentas_ahorro
Direccion	
Sector	
Provincia	
Canton	
Ingresos	
Activos	
Pasivo	
Patrimonio	
Cargas_familiares	
Actividad_econ	
Genero	
Estado_civil	
Nivel_educacion	
Profesion	
Agencia	
Nacionalidad	
Ctipopersona	
Edad	
Estat_cuentas_ahorro	
Numero_fijo	

Tabla 8

Atributos reales y seleccionados del origen de datos: Archivo Créditos desembolso diario 25/03/2019

Atributos reales	Atributos a considerar
Agencia	Identificacion
Nombre legal	Fecha desembolso
Numero socio	Fecha vencimiento
Identificacion	Actividad economica
Fecha nacimiento	Valor cuota cancelar
No. Cuenta debito	Frecuencia
No. Cuenta credito	Plazo
Fecha desembolso	Tipo prestamo
Fecha vencimiento	# num. Creditos activos
Fecha de pago de la cuota	Activo
Actividad economica	Pasivo
Monto + api	Ingreso
S. Desembolsado en cuenta	Egreso
Valor cuota cancelar	Patrimonio

Continúa

Certicados aporte credito	Ingresos netos
Api	Codigo actividad
Saldo certificados a. Fecha de corte	Actividad descripcion
Estado	Femision
Frecuencia	Fcaducidad
Plazo	
Tipo prestamo	
Tasa(%)	
Tasa sistema(%)	
Tasa num. Credito(%)	
# num. Creditos activos	
Asesor	
Autorizador credito	
Estado operación	
# credito	
Dias pago	
Validacion cedula	
Edad	
Telefono propio	
Direccion principal	
Ubicación g.	
Activo	
Pasivo	
Ingreso	
Egreso	
Patrimonio	
Ingresos netos	
Codigo actividad	
Actividad descripcion	
Femision	
Fcaducidad	

Tabla 9

Atributos reales y seleccionados del origen de datos: Archivo Cartera crédito

Atributos reales	**Atributos a considerar**
Fcorte	Csucursal
Csucursal	Identificacion
Nombrelegal	Fnacimiento
Cpersona	Genero
Numerosocio	Provincia
Ctipoidentificacion	Ciudad
Identificacion	Montooriginal
Fnacimiento	Fdesembolso
Genero	Fvencimiento
Direccion	Cestatuscuenta
Ntelefono	Ccondicionoperativa
Provincia	Tipoprestamo
Ciudad	Calificacionbanco
Montooriginal	Calificacioninterna
Tasainteres	Totaldeuda

Continúa

Fdesembolso	Frecuencias
Fvencimiento	Parroquias
Ccuenta	Actividad
Cestatuscuenta	Destino
Ccondicionoperativa	Fsolicitud
Cusuario_oficialcuenta	Tipo de prestamo
Orden	
Relacionproducto	
Tipoprestamo	
Calificacionbanco	
Calificacioninterna	
Capitalvigente	
Capitalnogenerainteres	
Capitalvencido	
Capitalenlegal	
Capitalcastigado	
Totaldeuda	
Diasvencido	
Oficial	
Frecuencias	
Parroquias	
Fsolicitud	
Tipo de prestamo	

Tabla 10

Atributos reales y seleccionados del origen de datos: Archivo C5.1 anexo detalle créditos castigados 2018

Atributos reales	Atributos a considerar
No. Cliente	No. Identificación
Nombre	Fecha Concesión
No. Identificación	Monto Original
Fecha Concesión	Tipo de préstamo
Monto Original	No. Operación
Tipo de préstamo	Fecha de castigo
No. Operación	Fecha de último pago
Fecha de castigo	Días Mora
Fecha de último pago	Saldo Capital
Días Mora	
Saldo Capital	
Valor de provisión	
Agencias	
Informe de Gestión	

3.7.Análisis de datos

Con la data entregada y analizando los atributos que cada una de las tablas Excel tenían, tomando en cuenta que muchos atributos se repetían en algunas tablas se procedió a realizar un modelo Entidad Relación para de esa manera luego crear el modelo multidimensional.

Es así como se obtuvo el modelo mostrado en la Fig. 11.

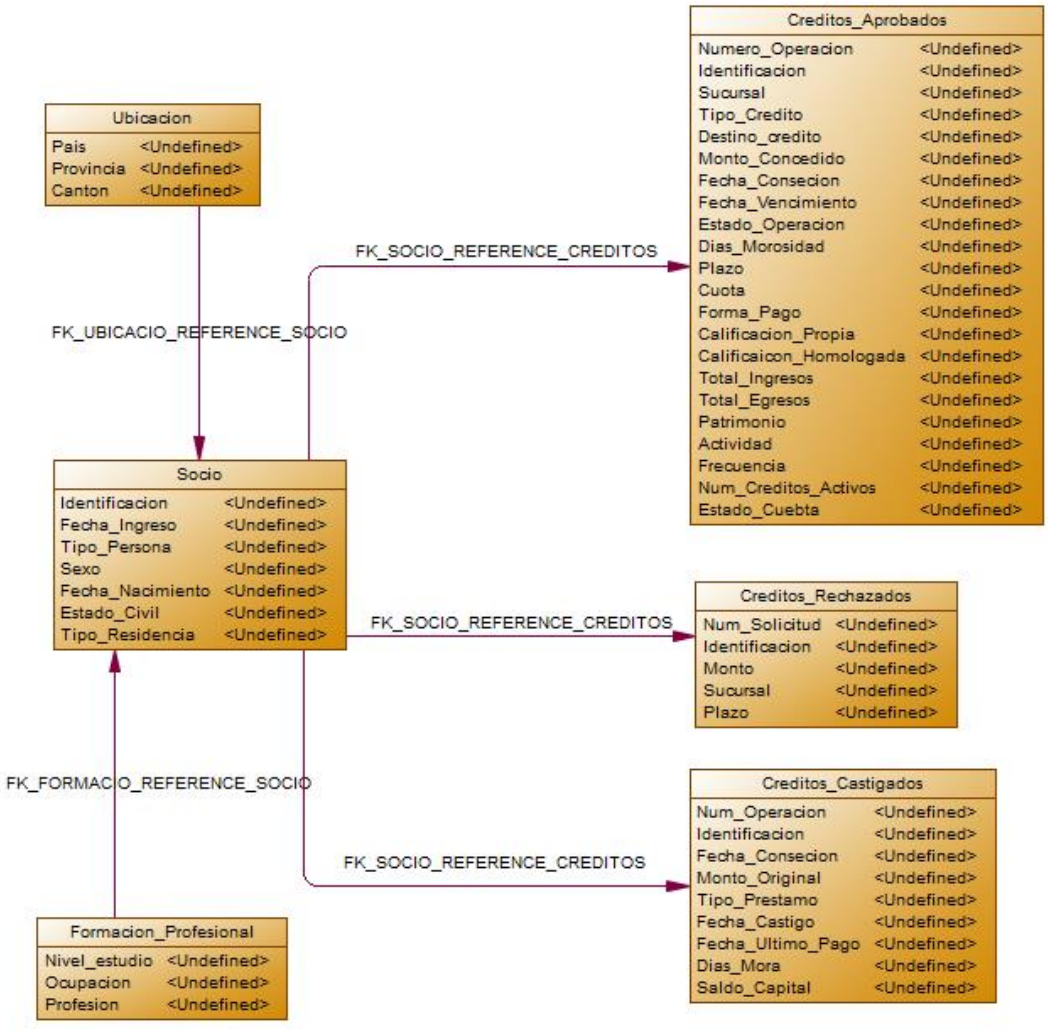


Figura 11. Modelo entidad relación

El modelo Entidad Relación fue creado considerando la relación que el socio tendrá con todas las tablas y los diferentes atributos que brindan los archivos Excel.

3.8.Creación de bodega de datos

A partir del modelo Entidad Relación creado y el análisis desarrollado anteriormente, donde, se consideraron los factores a tomar en cuenta para el modelo de minería de datos, se procedió a realizar el modelo multidimensional, que permitió conocer el esquema que tendrá la bodega de datos creada mediante el proceso ETL.

El modelo multidimensional creado es el mostrado en la Fig. 12 donde se consideraron únicamente los atributos relacionados con el caso de estudio.

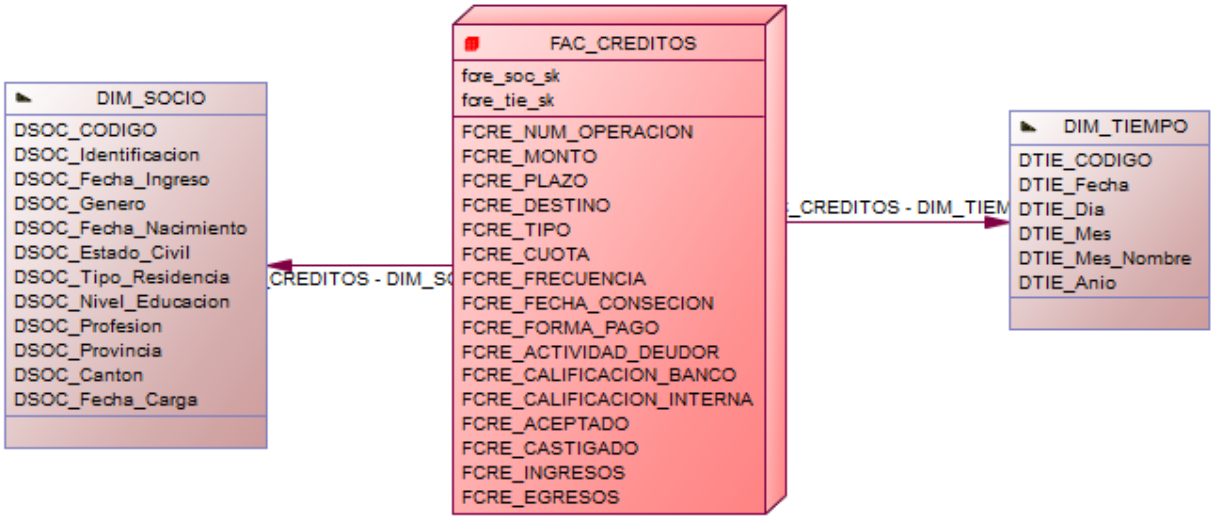


Figura 12. Modelo Multidimensional

El modelo multidimensional creado está compuesto por dos dimensiones y una Tabla de hechos en la que se encuentran relacionados todos los atributos que permiten determinar los patrones de comportamiento de un socio buen o mal pagador de un crédito.

Dimensión Socio: Mediante esta dimensión se agruparon todas las características propias de un socio en la cooperativa, datos que normalmente son obtenidas al momento de ingresar a la cooperativa y que no varían con el tiempo.

Dimensión Tiempo: Tabla que permitirá relacionar todas las fechas del modelo, con el objetivo de poder realizar análisis a lo largo del tiempo.

Tabla de Hechos Créditos: Mediante esta tabla se reunieron todos los atributos relacionados a los créditos solicitados en la cooperativa, donde se consideran todos los atributos que cambian en el tiempo de acuerdo con cuando se solicitó el crédito y relacionado con un socio determinado, además se consideró un atributo que permita determinar si el crédito fue aceptado, rechazado o castigado.

3.9.Preparación de los datos

Luego de crear el modelo multidimensional, se procedió a la implementación de la bodega de datos mediante procesos de Extracción, Transformación y Carga (ETL), donde se fue creando un flujo de datos (workflow) por cada dimensión y tabla de hecho requerida utilizando KIMNE.

EXTRACCIÓN: En la fase de extracción de datos de cada una de las dimensiones y tabla de hechos se procedió a obtener todos los datos provenientes de los archivos Excel, para luego procesarlos en el ETL respectivo.

TRANSFORMACIÓN: Esta fase en primera instancia resume los datos en base al modelo multidimensional diseñado y luego realiza la “limpieza de datos” donde se aseguró la calidad de los datos a procesar, se evitóinformación no veraz o errónea, se ahorró costes de espacio en disco al eliminarse la información duplicada y se agilizó las consultas por la ausencia de datos repetidos o inservibles. Esto se lo realizó aplicandoreglas de unificación de datos, validando completitud de y estandarizando los datos, mediante nodos como: *“StringManipulation”, “CellSplitter”, “ColumnFilter”, “MissingValue”*, entre otros.

CARGA: Fase en la que se guarda en una nueva base de datos las tablas creadas.

Como primer paso se determinó el origen de datos para cada ETL resumido en la Tabla 5de acuerdo con los archivos Excel entregados por la cooperativa.

Tabla 11

Dimensiones con orígenes de datos

Tabla	Atributo	Origen de Dato (Archivo Excel)
DIM_SOCIO	DSOC_Identificacion	Base de datos clientes Socio estudio mercado 18/10/2018
	DSOC_Fecha_Ingreso	Base de datos clientes
	DSOC_Tipo_Persona	Base de datos clientes
	DSOC_Genero	Base de datos clientes
	DSOC_Edad	Base de datos clientes
	DSOC_Estado_Civil	Base de datos clientes
	DSOC_Ocupacion	Base de datos clientes
	DSOC_Tipo_Residencia	Base de datos clientes
	DSOC_Nivel_Educacion	Socio estudio mercado 18/10/2018
	DSOC_Profesion	Socio estudio mercado 18/10/2018
	DSOC_Provincia	Socio estudio mercado 18/10/2018
	DSOC_Canton	Socio estudio mercado 18/10/2018
FAC_CREDITOS	FCRE_NUM_OPERACION	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
	FCRE_MONTO	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
	FCRE_PLAZO	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
	FCRE_DESTINO	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
	FCRE_TIPO	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
	FCRE_CUOTA	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito

Continúa

FCRE_FRECUENCIA	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
FCRE_FECHA CONSECION	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
FCRE_FORMA_PAGO	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
FCRE_ACTIVIDAD_DEUDOR	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019 Cartera crédito
FCRE_CALIFICACION_BANCO	C3. anexo detallado de cartera al 31/12/2018 Cartera crédito
FCRE_CALIFICACION INTERNA	C3. anexo detallado de cartera al 31/12/2018 Cartera crédito
FCRE_INGRESOS	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019
FCRE_EGRESOS	C3. anexo detallado de cartera al 31/12/2018 Créditos desembolso diario 25/03/2019
FCRE_ACEPTADO	Solicitudes crédito rechazados
FCRE_CASTIGADO	C5.1 anexo detalle créditos castigados 2018

El origen de datos para la dimensión Tiempo no se lo exporta desde una fuente de datos, sino que se lo genera con un rango de fechas, de acuerdo con las fechas existentes en los datos.

3.9.1. ETL Dimensión Tiempo

El workflow creado es el mostrado en la Fig. 12, donde el semáforo mostrado bajo cada nodo indica su correcta ejecución cuando se encuentra en verde. El proceso ETL desarrollado para esta dimensión comenzó generando fechas desde enero del 2015 hasta el 2020 en base a los datos de fechas receptados. Luego de la generación de las fechas se procedió a colocarle un id, realizar operaciones con la fecha en *STRING*con el propósito de obtener el día, mes y año por separado

que son atributos de la dimensión y por último renombrar a cada uno de los atributos en base a la notación establecida.

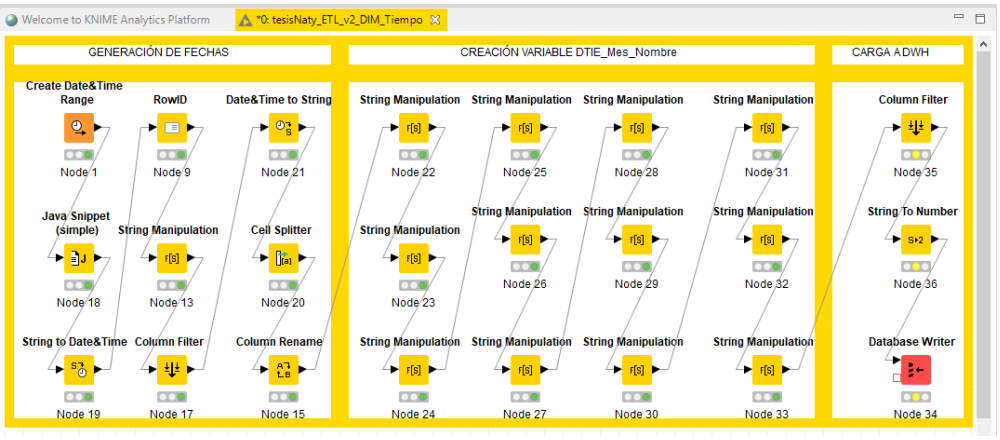


Figura 13. ETL Dimensión Tiempo antes de la ejecución

En la segunda fase se procedió a generar el atributo DTIE_Mes_Nombre basado en el mes generado anteriormente. Finalmente, se realizó la respectiva carga a la nueva base de datos que contiene la bodega creada. Luego de realizar la ejecución se consiguió semáforos en verde en todos los nodos como se ve en la Fig. 14 indicando una correcta carga.

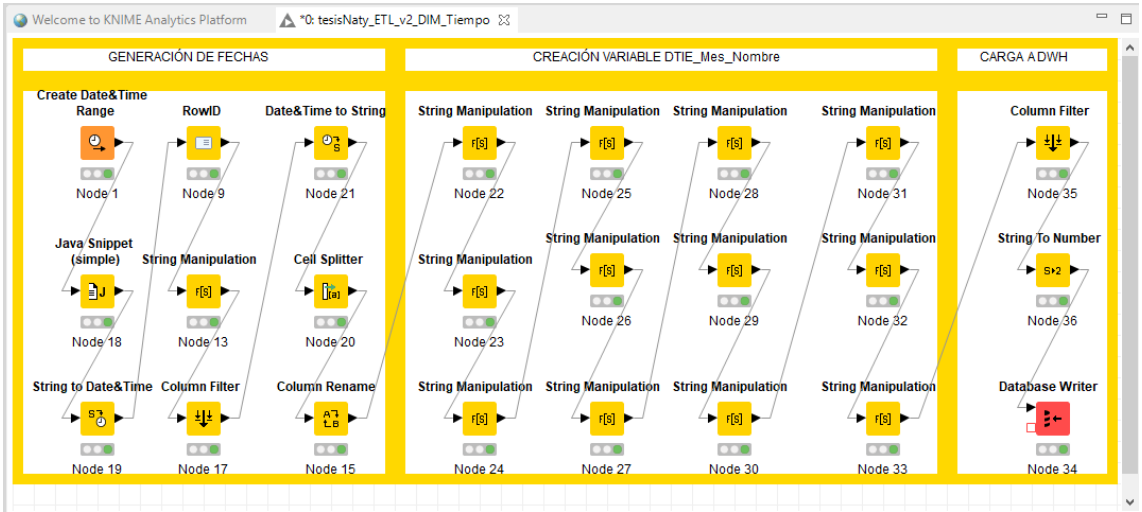


Figura 14. ETL Dimensión Tiempo luego de la ejecución

3.9.2. ETL Dimensión Socio

Para la creación de este ETL se desarrolló el workflow mostrado en la Fig. 15, en este caso la captura fue tomada en el momento de ejecución por tal razón algunos semáforos están en azul, porque están en proceso de ejecución.

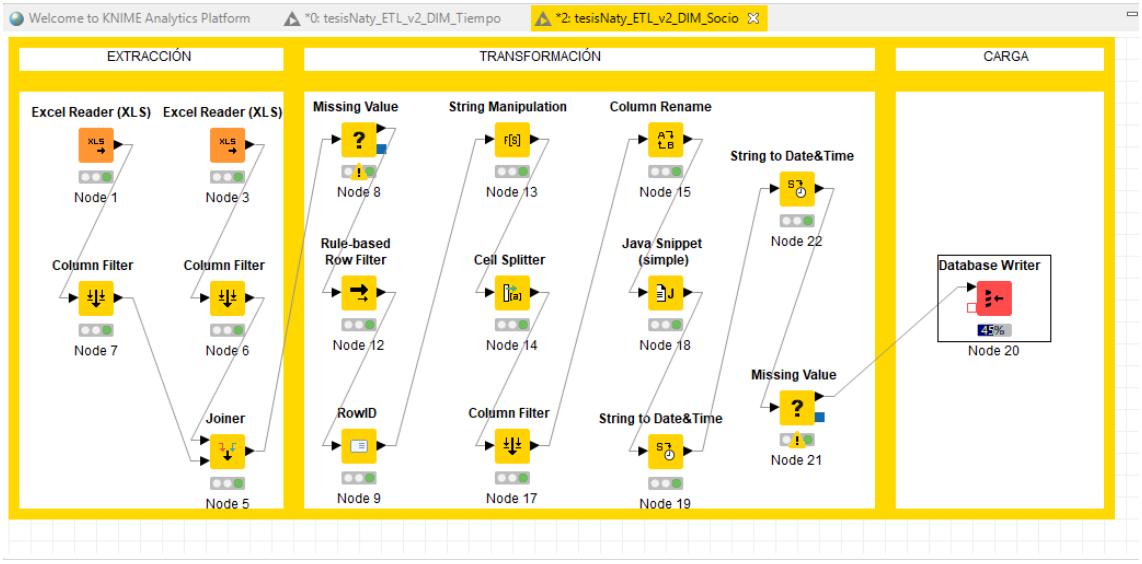


Figura 15. ETL Dimensión Socio antes de la ejecución

Para este proceso ETL se comenzó realizando la extracción de los orígenes de datos, de acuerdo a lo indicado anteriormente, esta dimensión se provee de datos de los archivos Excel: "Base de datos clientes" y "Socio estudio mercado 18/10/2018", de los cuales se realiza un filtro tomando solo los atributos deseados para luego mediante una operación *JOINER*se unen todos los datos en una sola tabla. A continuación, se pasa a la fase de transformación donde se llenan valores faltantes, se coloca id, se renombran a los atributos, se coloca fecha de carga, entre otros. Por último, se realiza la fase da carga a la bodega de datos final. En la Fig. 16 se puede observar la correcta ejecución del workflow.

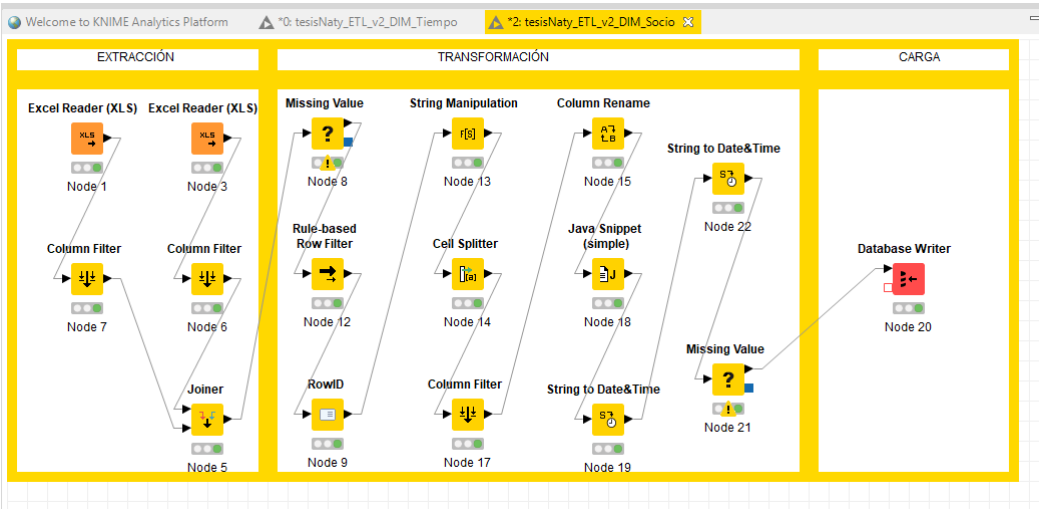


Figura 16. ETL Dimensión Socio luego de la ejecución

3.9.3. ETL Tabla de Hecho Créditos

Finalmente, el workflow para la creación del ETL de la Tabla de hechos, es el mostrado en las figuras 17, 18, 19 donde se puede observar la correcta ejecución del flujo.

Inicialmente se comienza el ETL con la extracción de la información de los orígenes de datos en los archivos Excel, de acuerdo a los analizado anteriormente, primero se extrae datos de los archivos: “C3.anexo detallado de cartera al 31/12/2018”, “Créditos desembolso diario 25/03/2019” y “ Cartera crédito”; para cada uno de los archivos primero se filtran los atributos necesarios y luego mediante el operador JOINER se los une en una sola tabla, a continuación, se relaciona con el Excel de créditos Rechazados consiguiendo un tabla total.

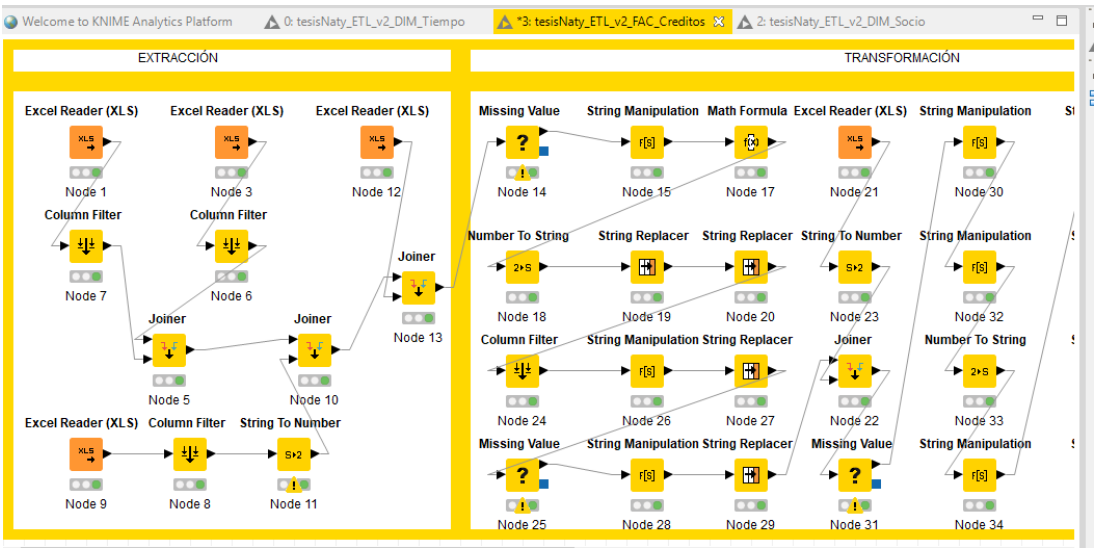


Figura 17. ETL FAC Créditos: Extracción

En la segunda fase de transformación, se procedió a rellenar datos faltantes, verificar tipos de datos, renombrar los atributos, crear variable FCRE_ACEPTADO, que indica mediante un SI o NO si la solicitud de crédito fue aprobada o no. Además, se extrajo la información de créditos castigados y se creó la variable FCRE_CASTIGADO, que mediante un SI o N0 indica si hubo mora en los créditos.

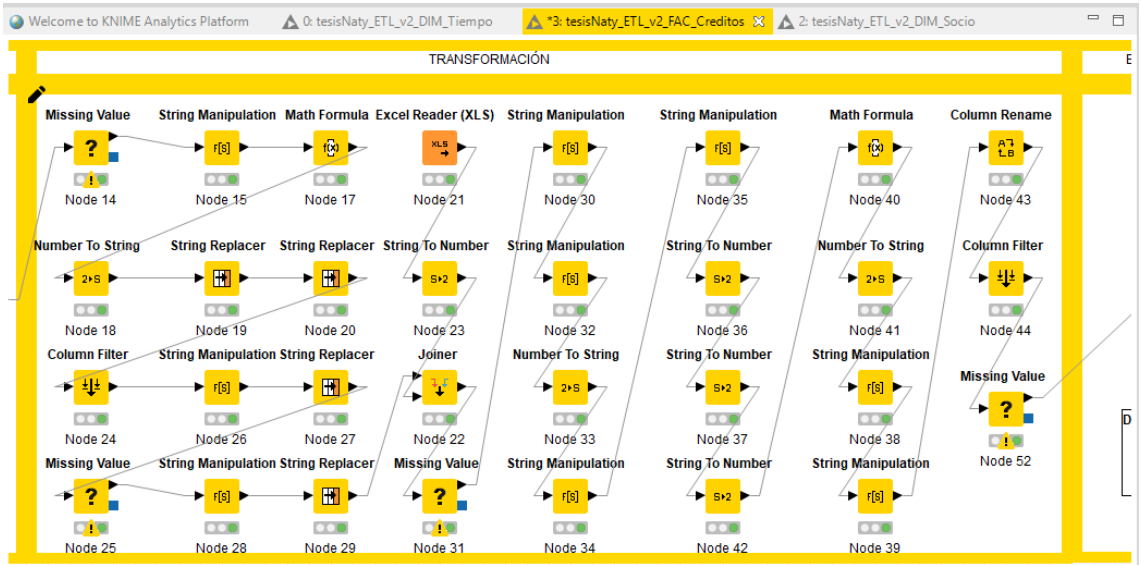


Figura 18. ETL FAC Créditos: Transformación

La siguiente fase de flujo son las relaciones de la tabla de hechos con las dimensiones, donde mediante el operador JOINER se relaciona la tabla ya transformada con los ETL cargados anteriormente en la base de datos de la bodega de datos. Finalmente se carga la tabla total en la nueva base de datos.

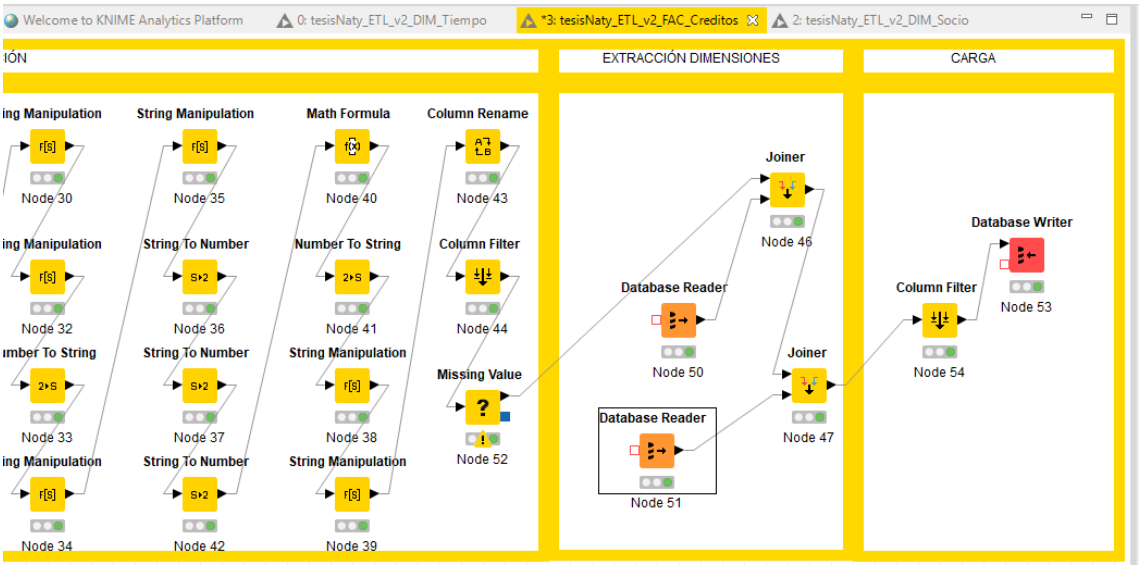


Figura 19. ETL FAC Créditos: Carga

3.10. Creación de la base de datos

El gestor de base de datos utilizado para la carga de la bodega de datos de este proyecto fue MySQL, por ser *Open Source*, por la velocidadal realizar las operaciones, lo que le hace uno de los gestores con mejorrendimiento. También presenta bajocosto en requerimientos para la elaboración de bases de datos, puesto que debido a su bajo consumo puede ser ejecutado en una máquina con escasosrecursos sin ningún problema.MySQL posee facilidadde configuración e instalación y es soportadoen una gran variedad de Sistemas Operativos. El gestor de base de datos MySQL fue implementado mediante la herramienta XAMPP (Fig. 20).

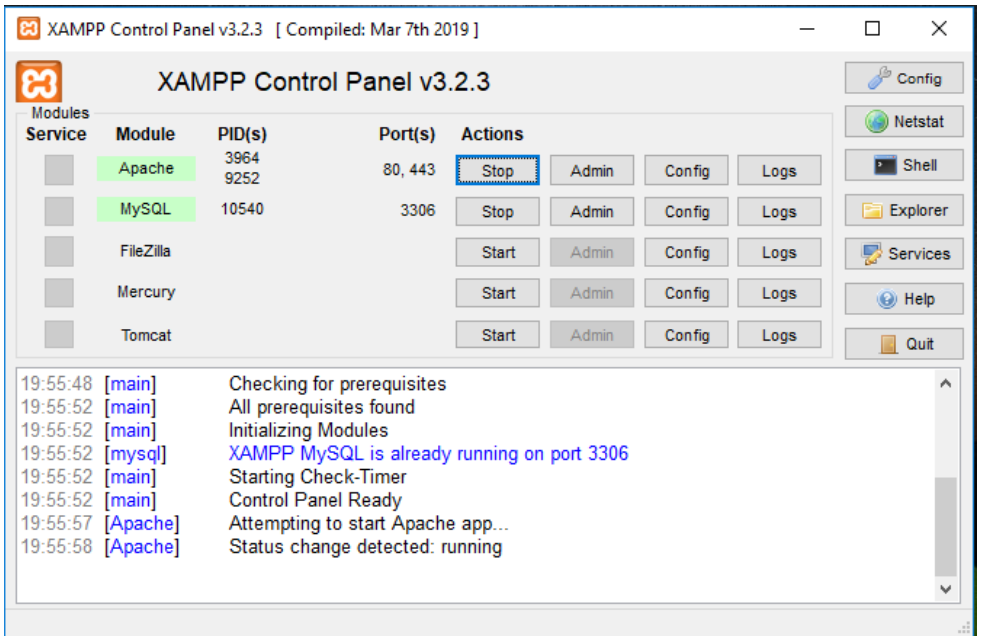


Figura 20. Herramienta XAMPP

Utilizando XAMPP, se levantó un servidor Apache, que permitió utilizar una interfaz web para la creación y manipulación de la base de datos MySQL. La base de datos creada para la bodega de datos se llamó “DWH_Tesis” y se lo puede observar en la Fig. 21.

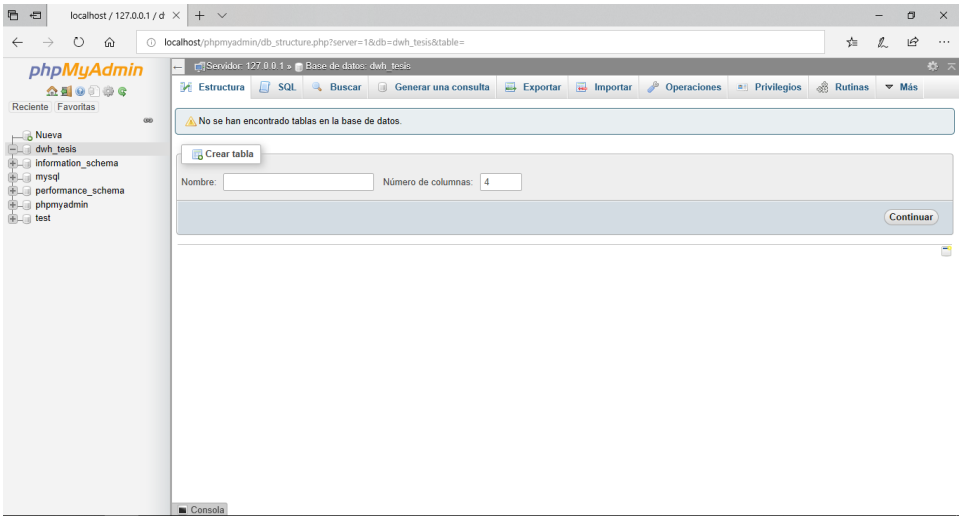


Figura 21. Creación base de datos DWH_Tesis

Desde cada uno de los ETLs creados en KNIME, se fueron cargando las tablas a la base de datos, empezando por la Dimensión Tiempo, como lo indican las Fig. 22, 23.

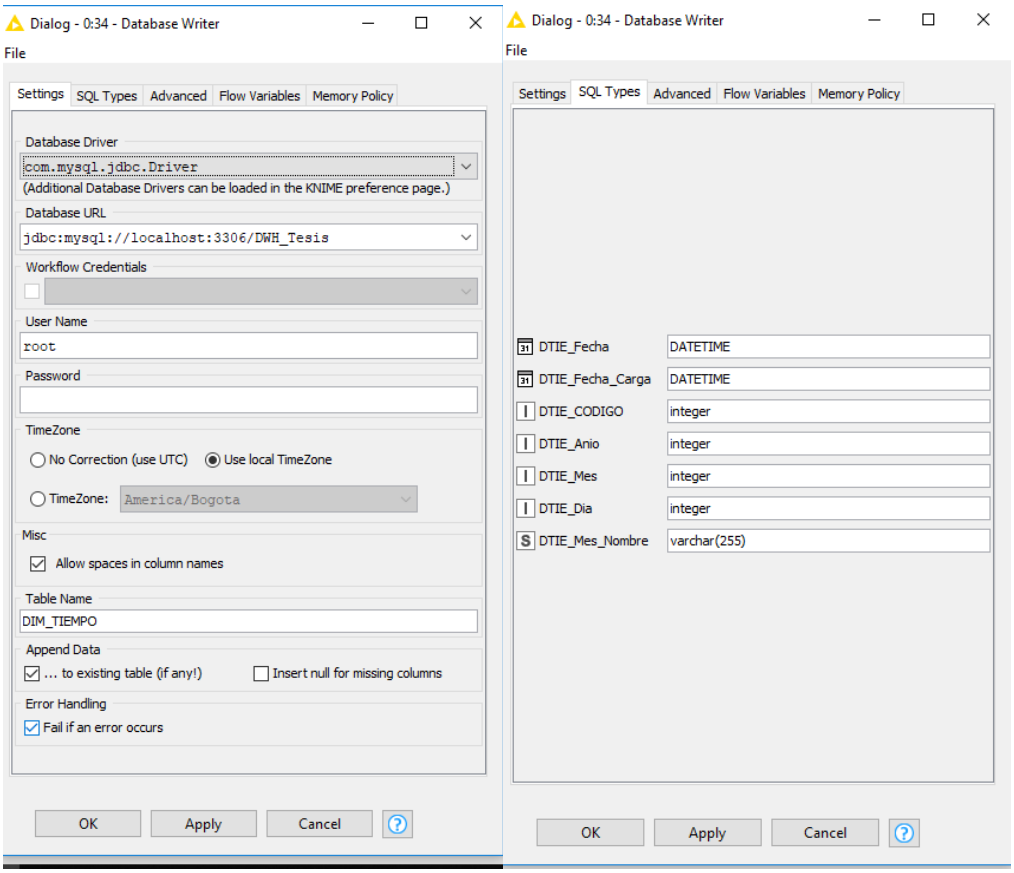


Figura 22. Configuración Carga Dimensión Tiempo en KNIME

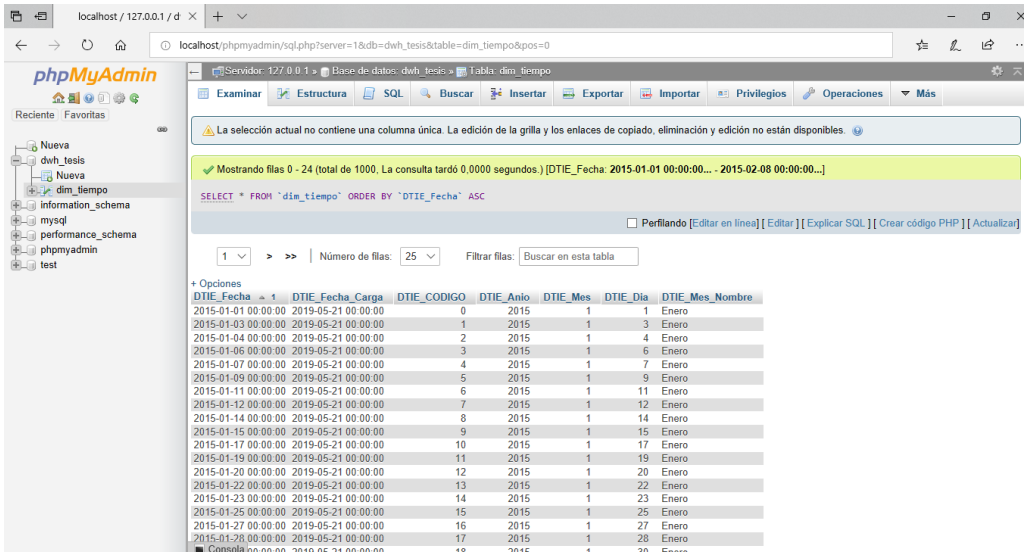


Figura 23. Carga de datos Dimensión Tiempo en MySQL

De la misma manera se lo realizó con el ETL de la Dimensión Socio como se lo puede observar en las Fig. 24,25.

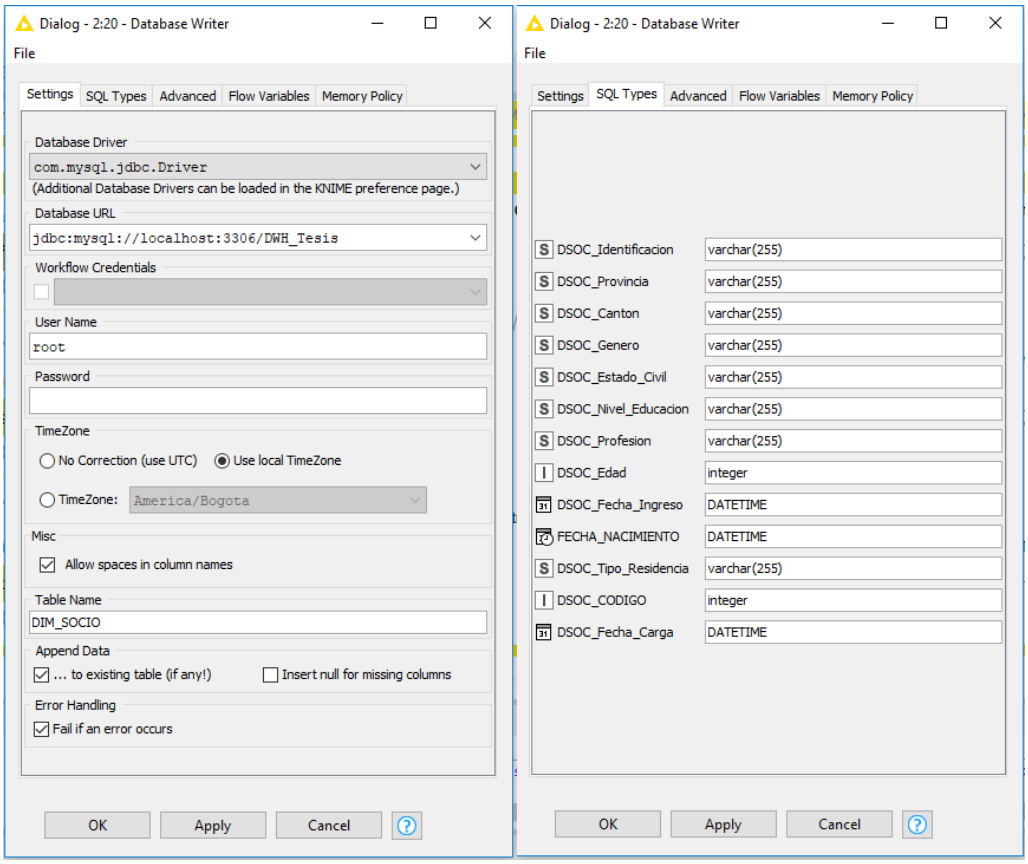


Figura 24. Configuración Carga DimensiónSocio en KNIME

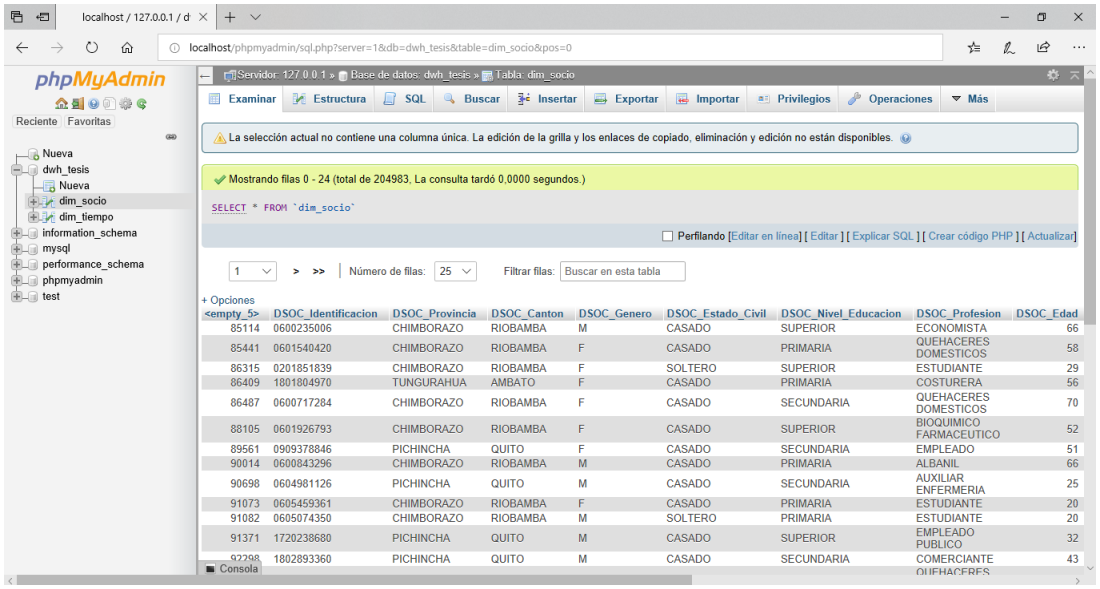


Figura 25. Carga de datos Dimensión Socio en MySQL

Y finalmente se realizó el mismo procedimiento para la tabla de hechos, creando la tabla "FAC_CREDITOS" y como se los observa en las Fig. 26,27.

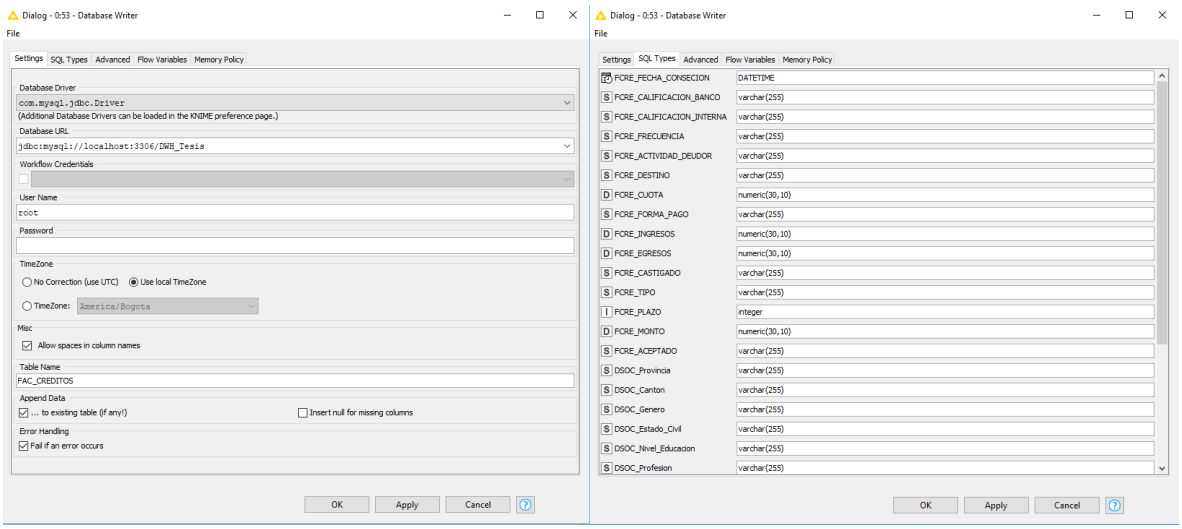


Figura 26. Configuración Carga Tabla de Hechos Créditos en KNIME

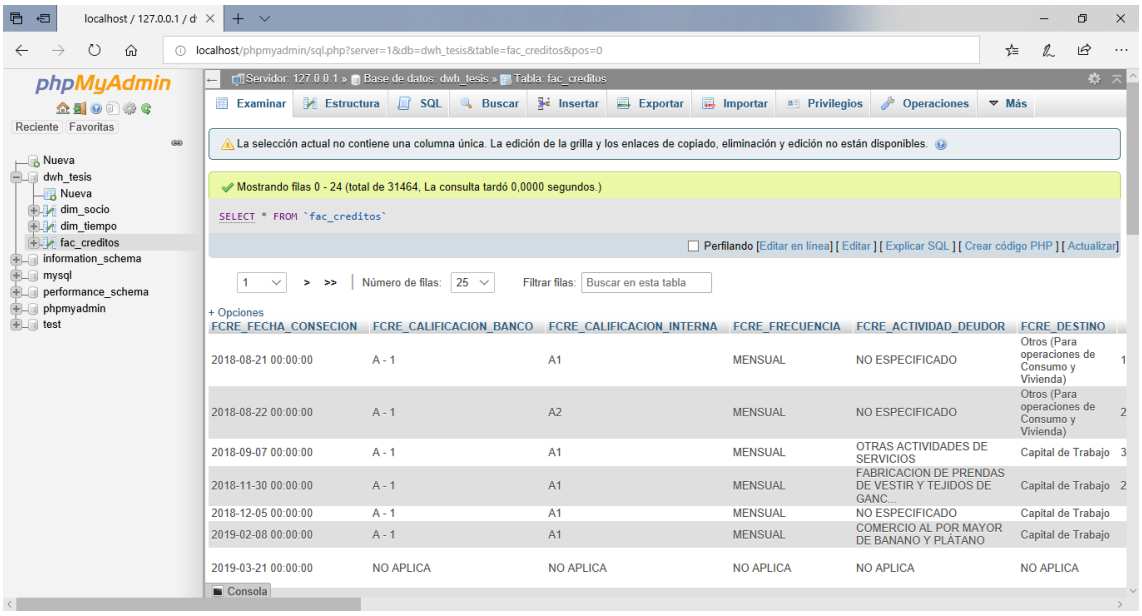


Figura 27. Carga de datos Tabla de Hecho Créditos en MySQL

3.11. Análisis de bodega de datos

Luego de la creación de la bodega de datos, se procedió a realizar un análisis de la información contenida ahí, para compararla con la obtenida luego de la creación del modelo. Es

así como utilizando KNIME se procedió a crear reportes sobre el comportamiento de los datos primero creando un workflow como se puede ver en la Fig. 28 y luego creando un reporte mediante una extensión de KNIME mostrado en la Fig. 29.

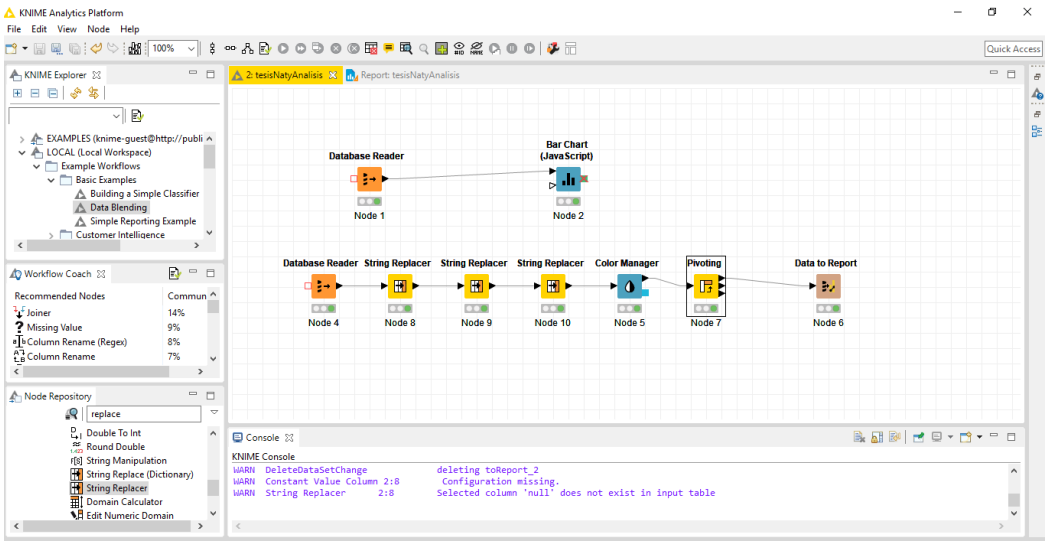


Figura 28. Workflow para crear reportes

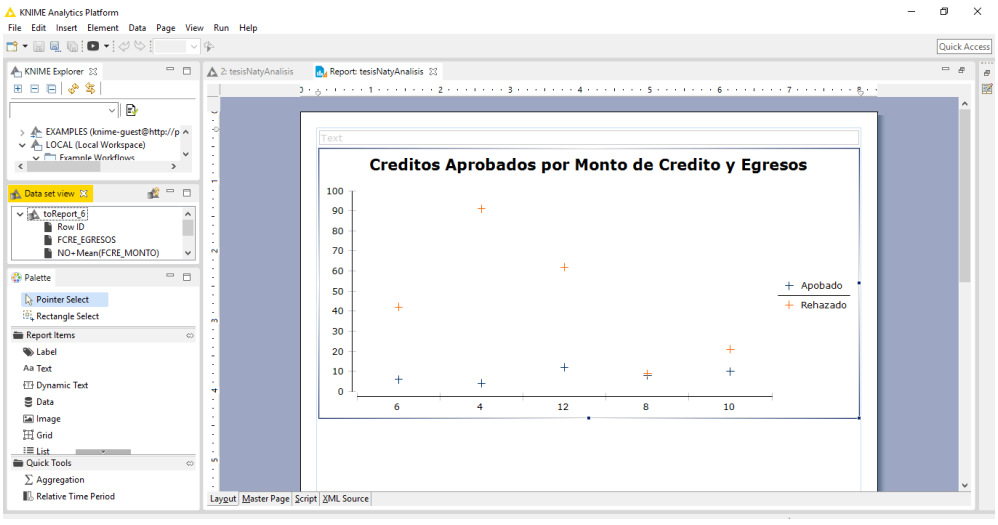


Figura 29. Extensión de reportes en KNIME

Los reportes generados pueden ser cargados a diferentes formatos, en este caso se los envió a una interfaz web obteniendo los siguientes resultados.

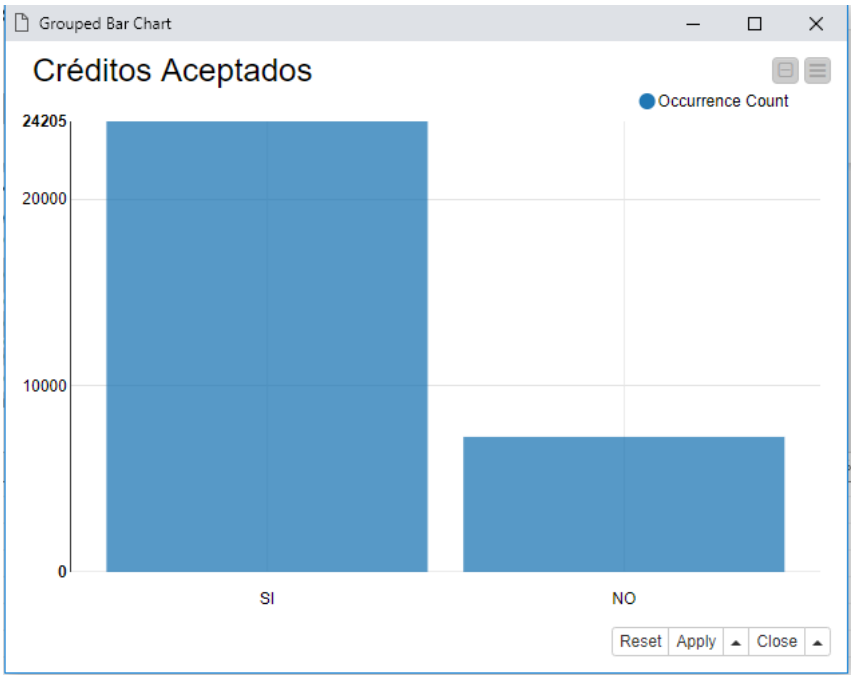


Figura 30. Reporte de número de casos

En base a la Fig. 30, se puede indicar que dentro de los datos la mayoría de casos responden a créditos aprobados.

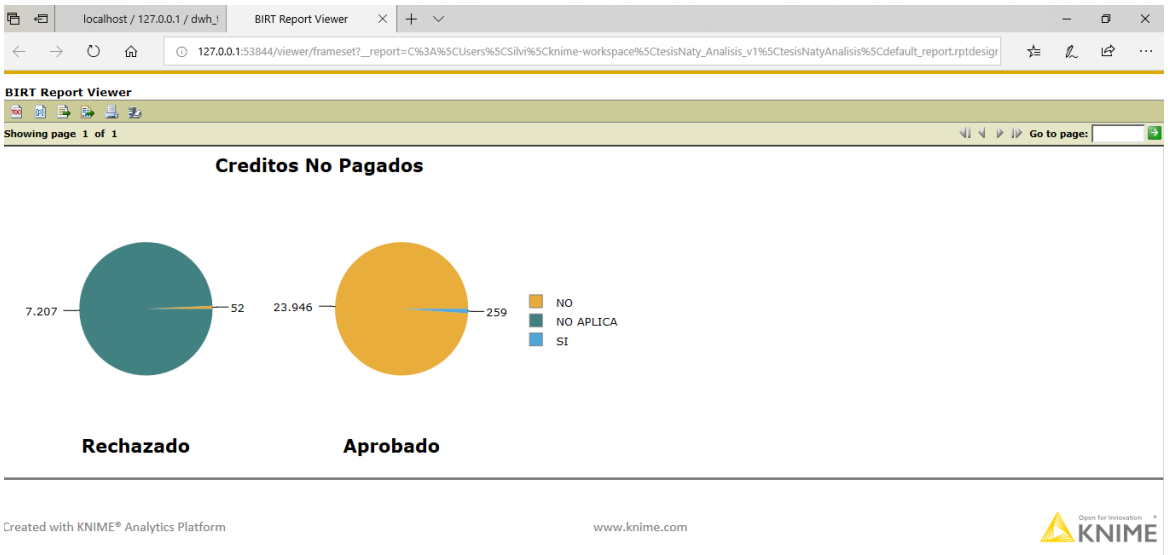


Figura 31. Reporte de créditos no pagados

De acuerdo a la Fig. 31 un porcentaje pequeño de los créditos aprobados fueron castigados, es decir que no se pagaron.

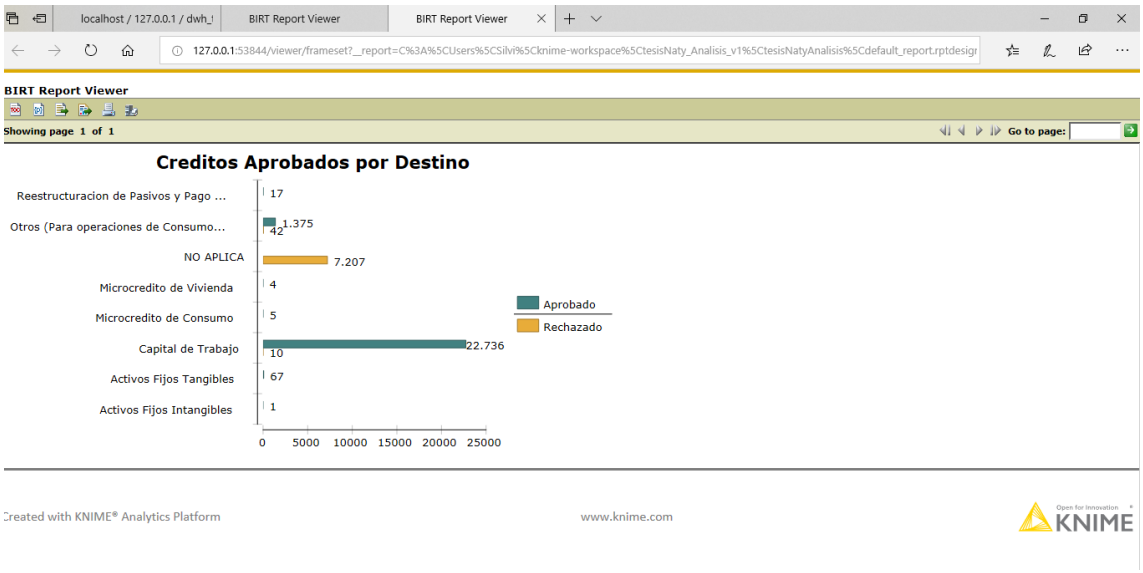


Figura 32. Reporte Créditos aprobados por destino de crédito

En la Fig. 32 se indica como actualmente el destino de crédito que genera más créditos aprobados es “Capital de trabajo”, mientras que el que más genera rechazo es “Otros (Para operaciones de consumo)”.

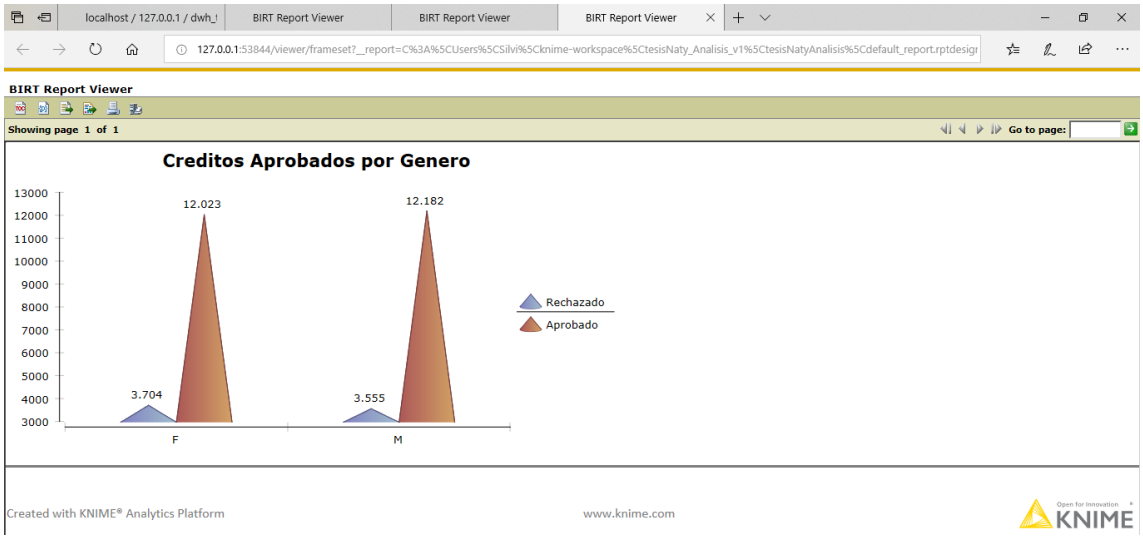


Figura 33. Reporte Créditos aprobados por género

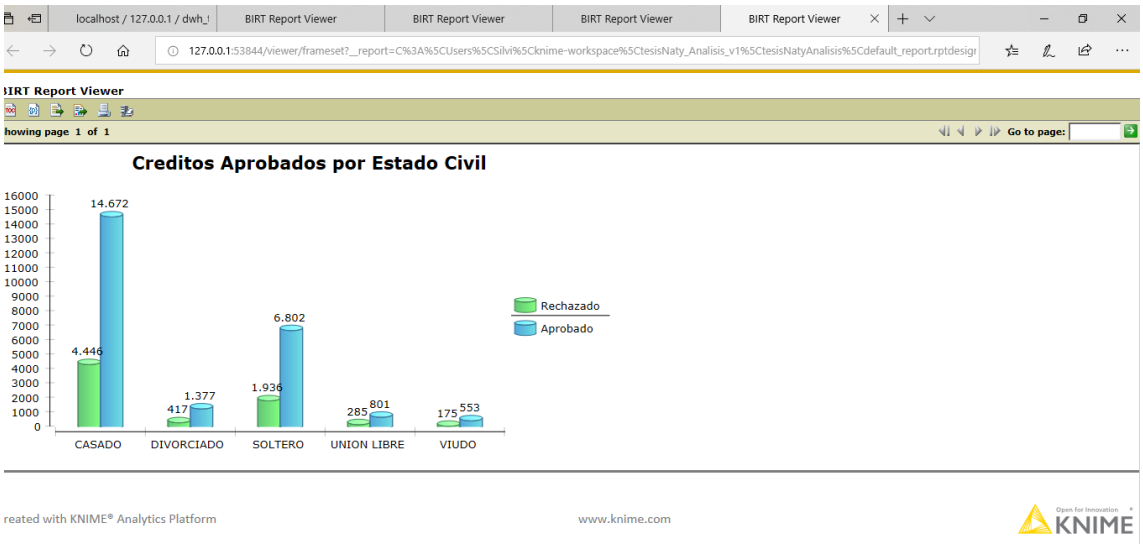


Figura 34. Reporte Créditos aprobados por estado civil

Las Fig. 33,34, indican como los hombres poseen más créditos aprobados mientras que las mujeres tienen más créditos rechazados. Por otro lado, las personas en estado civil casado tienen mayor porcentaje de aprobación y rechazo de créditos.

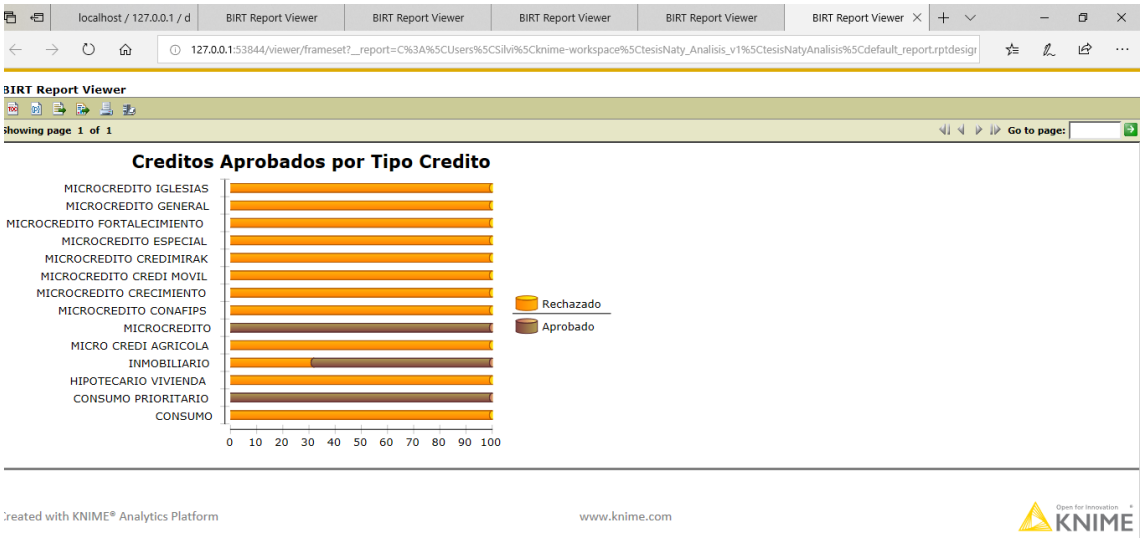


Figura 35. Reporte Créditos aprobados por tipo de crédito

En la Fig. 35 se puede ver que los tipos de crédito “Microcrédito” y “Consumo Prioritario” tienen mayor porcentaje de Aprobación, mientras quela mayoría de tipos tiene alto porcentaje de rechazo.

Figura 36. Reporte Créditos rechazados por nivel de estudios y profesión

De acuerdo a la Fig. 36 un acuicultor con nivel de estudios primario tiene mayor número de créditos rechazados y entre mayor es el nivel de estudios menor número de créditos rechazados se tiene.

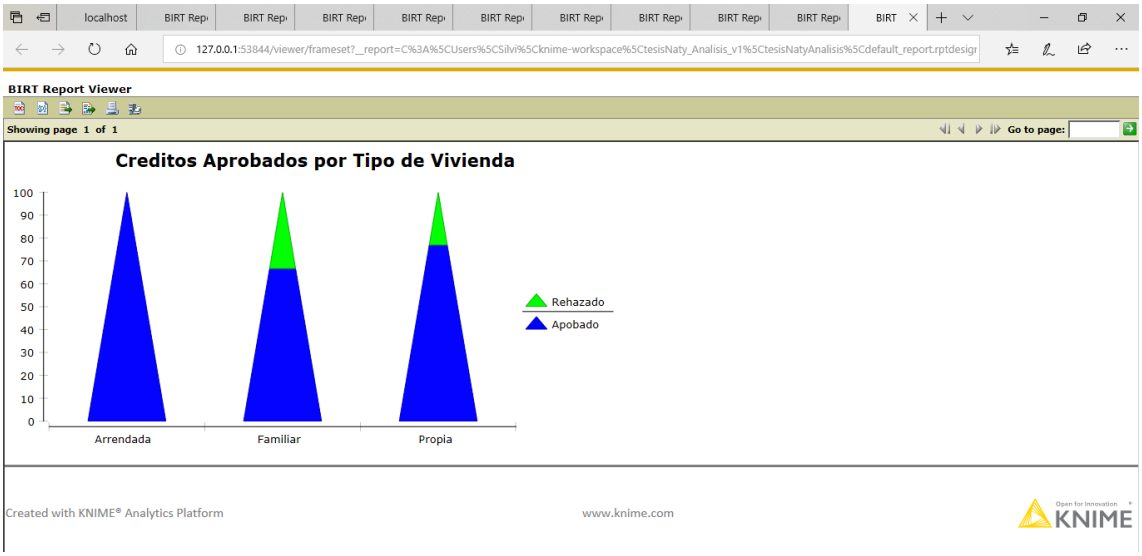


Figura 37. Reporte Créditos aprobados por tipo de vivienda

Se realizó un análisis por tipo de vivienda en la Fig. 37 donde se observa que se tiene mayor número de créditos aprobados en tipos de vivienda “Arrendadas” y mayor rechazo en “Familiar”.

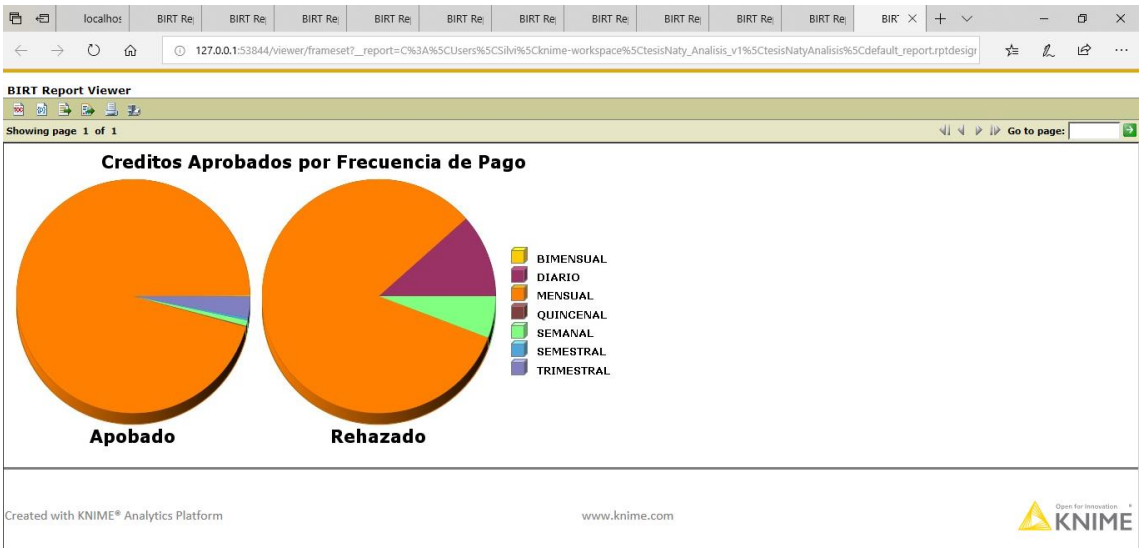


Figura 38. Reporte Créditos aprobados por frecuencia de pago

Con respecto a la frecuencia de pago, la Fig. 38 indica que en ambos casos la mayoría aplica por frecuencia "Mensual", pero luego se tiene que la frecuencia "Diario" genera más créditos rechazados.

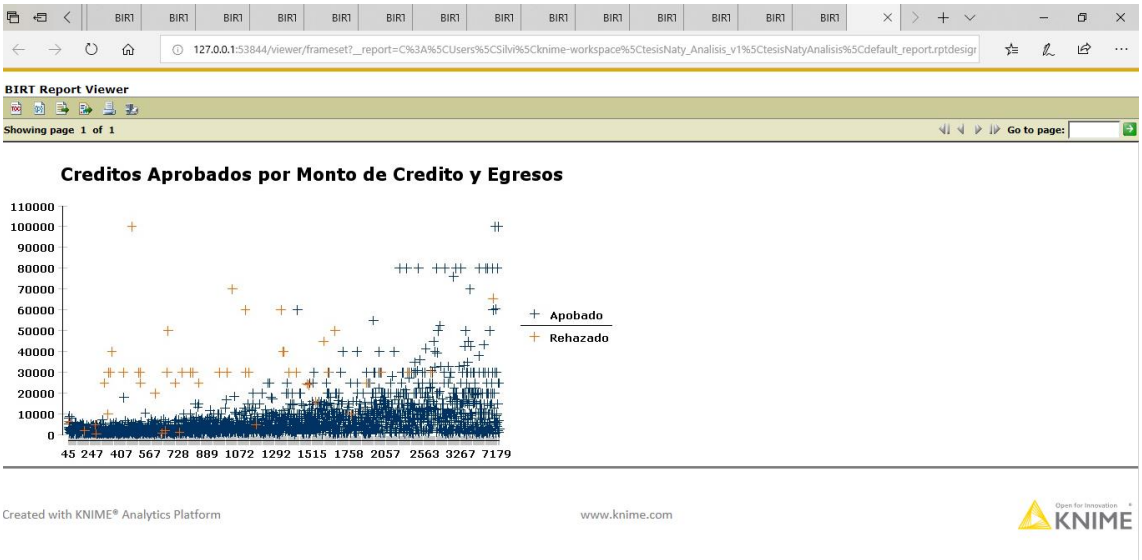


Figura 39. Reporte Créditos aprobados por monto de créditos y egresos.

Por el monto y los egresos, la Fig. 39, muestra que entre más monto y más egresos los créditos han sido aprobados, por otro lado, que entre menos monto y más egresos estos han sido rechazados.

3.12. Creación del modelo de minería de datos

Utilizando KNIME se procedió a realizar el análisis de minería de datos a la información transformados de la bodega de datos mediante tres técnicos:

- Árbol de decisión
- Redes neuronales
- NaiveBayes

Se debe considerar que el atributo a predecir seleccionado es FCRE_ACEPTADO, que indica si un crédito será o no aprobado.

Árbol de decisión

El algoritmo Decisión Tree permitió obtener un modelo de minería de datos más amigable para el usuario, ya que lo presenta de manera gráfica. Mediante el workflow de la Fig. 40 se realizó la creación de dicho modelo, donde se comenzó haciendo la partición de la data para la posterior evaluación. Luego se aplicó el algoritmo y se utilizó algunos operando para visualizar el modelo creado y sus reglas como se lo observa en la Fig. 41, 42.

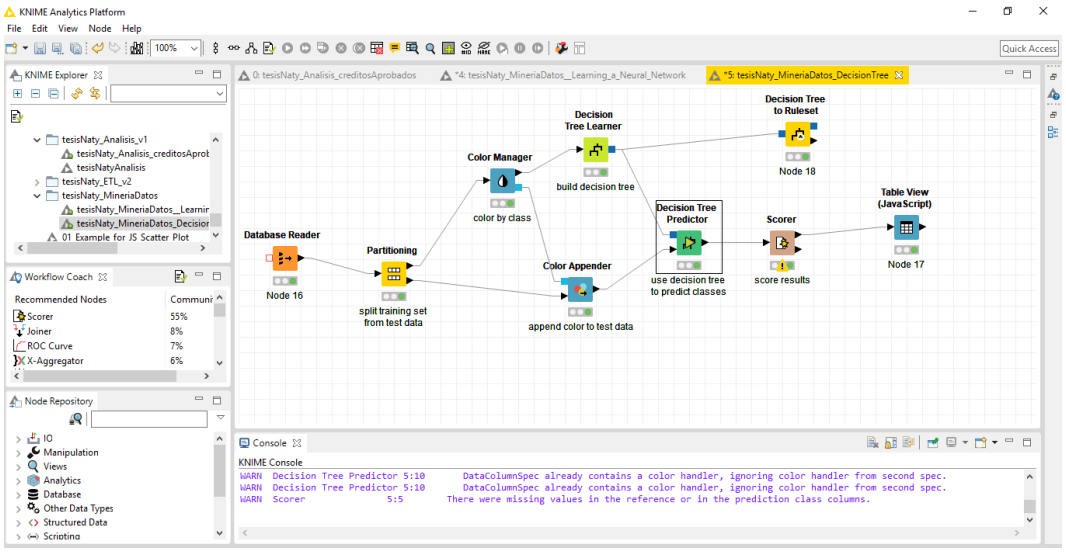
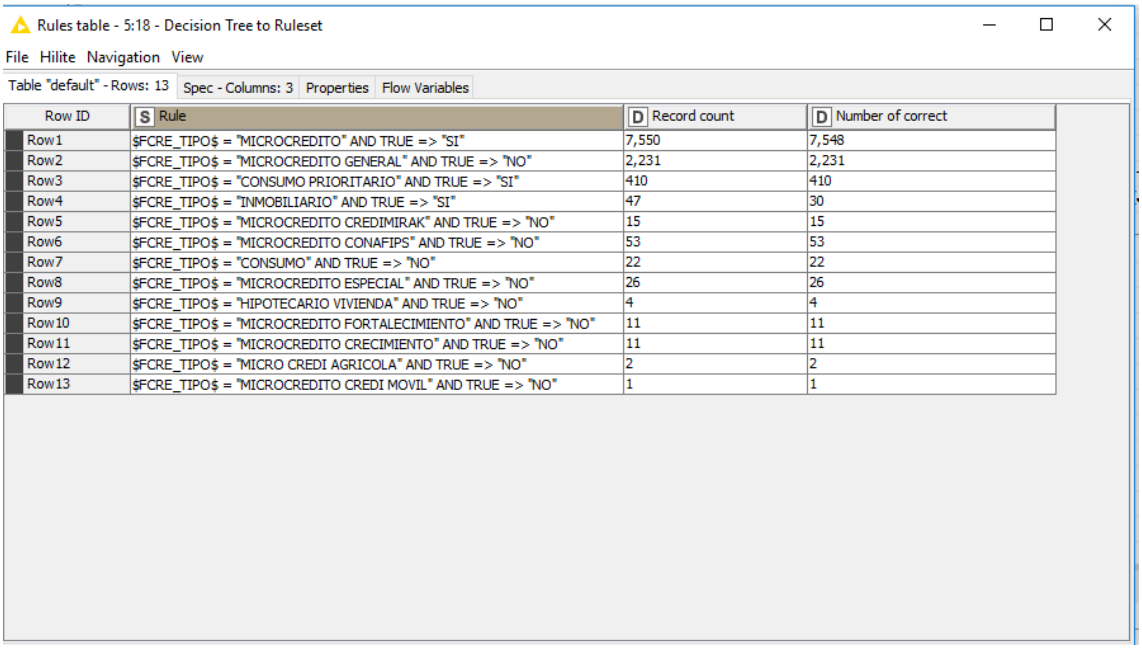


Figura 40. Workflow de modelo DecisionTree

Figura 41. Modelo DecisionTree



Row ID	Rule	Record count	Number of correct
Row1	$FCRE_TIPO$ = "MICROCREDITO" AND TRUE => "SI"	7,550	7,548
Row2	$FCRE_TIPO$ = "MICROCREDITO GENERAL" AND TRUE => "NO"	2,231	2,231
Row3	$FCRE_TIPO$ = "CONSUMO PRIORITARIO" AND TRUE => "SI"	410	410
Row4	$FCRE_TIPO$ = "INMOBILIARIO" AND TRUE => "SI"	47	30
Row5	$FCRE_TIPO$ = "MICROCREDITO CREDIMIRAK" AND TRUE => "NO"	15	15
Row6	$FCRE_TIPO$ = "MICROCREDITO CONAFIPS" AND TRUE => "NO"	53	53
Row7	$FCRE_TIPO$ = "CONSUMO" AND TRUE => "NO"	22	22
Row8	$FCRE_TIPO$ = "MICROCREDITO ESPECIAL" AND TRUE => "NO"	26	26
Row9	$FCRE_TIPO$ = "HIPOTECARIO VIVIENDA" AND TRUE => "NO"	4	4
Row10	$FCRE_TIPO$ = "MICROCREDITO FORTALECIMIENTO" AND TRUE => "NO"	11	11
Row11	$FCRE_TIPO$ = "MICROCREDITO CRECIMIENTO" AND TRUE => "NO"	11	11
Row12	$FCRE_TIPO$ = "MICRO CREDI AGRICOLA" AND TRUE => "NO"	2	2
Row13	$FCRE_TIPO$ = "MICROCREDITO CREDI MOVIL" AND TRUE => "NO"	1	1

Figura 42. Reglas modelo DecisionTree

De acuerdo con este algoritmo, el factor que mayor determina si un crédito debe o no ser aprobado es el tipo de crédito.

Redes Neuronales

El workflow para la creación del modelo mediante esta técnica está en la Fig. 43, donde se realiza un proceso similar al de DecisionTree y al final se gráfica los resultados obtenido como se lo observa en la Fig. 44. Además, en la Fig. 45 se puede observar le modelo creado.

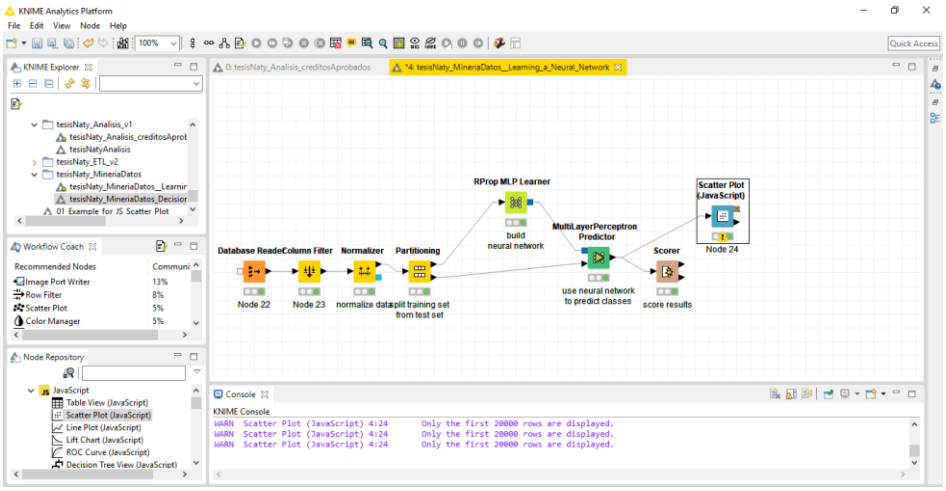


Figura 43. Workflowmodelo de Neural Network

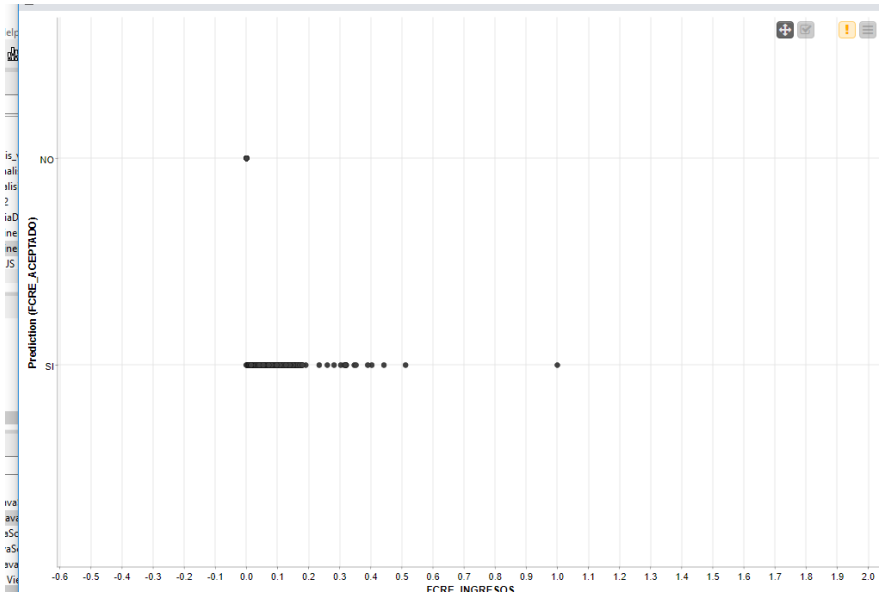
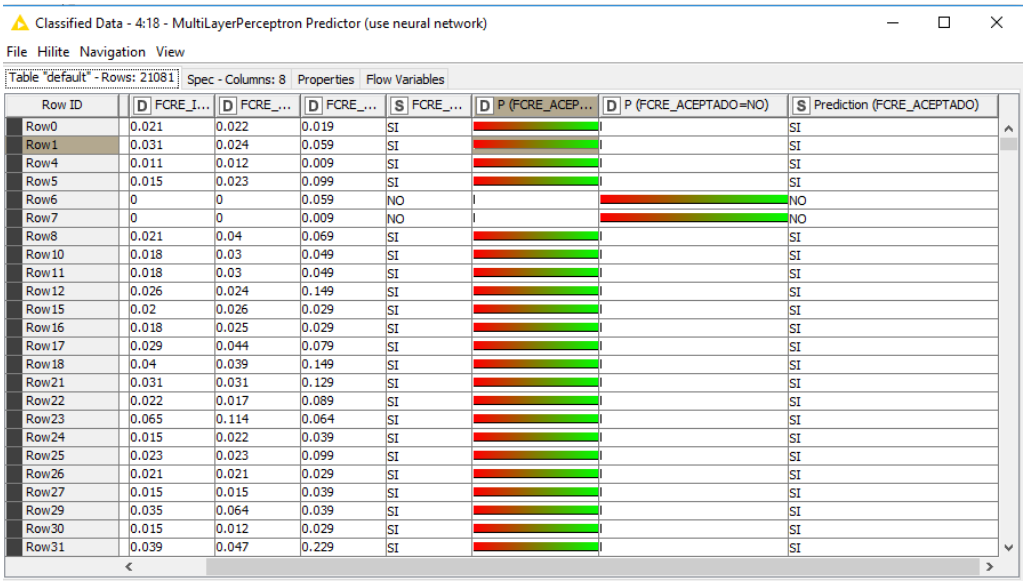


Figura 44. Gráfica de resultados modelo de Neural Network



Row ID	FCRE_I...	FCRE_...	FCRE_...	FCRE_...	P (FCRE_ACEP...	P (FCRE_ACEPTADO=NO)	Prediction (FCRE_ACEPTADO)
Row0	0.021	0.022	0.019	SI			SI
Row1	0.031	0.024	0.059	SI			SI
Row4	0.011	0.012	0.009	SI			SI
Row5	0.015	0.023	0.099	SI			SI
Row6	0	0	0.059	NO			NO
Row7	0	0	0.009	NO			NO
Row8	0.021	0.04	0.069	SI			SI
Row10	0.018	0.03	0.049	SI			SI
Row11	0.018	0.03	0.049	SI			SI
Row12	0.026	0.024	0.149	SI			SI
Row15	0.02	0.026	0.029	SI			SI
Row16	0.018	0.025	0.029	SI			SI
Row17	0.029	0.044	0.079	SI			SI
Row18	0.04	0.039	0.149	SI			SI
Row21	0.031	0.031	0.129	SI			SI
Row22	0.022	0.017	0.089	SI			SI
Row23	0.065	0.114	0.064	SI			SI
Row24	0.015	0.022	0.039	SI			SI
Row25	0.023	0.023	0.099	SI			SI
Row26	0.021	0.021	0.029	SI			SI
Row27	0.015	0.015	0.039	SI			SI
Row29	0.035	0.064	0.039	SI			SI
Row30	0.015	0.012	0.029	SI			SI
Row31	0.039	0.047	0.229	SI			SI

Figura 45. Modelo de Neural Network

Este modelo indica que la mayoría de las predicciones son a que el crédito será aprobado.

NaiveBayes

La Fig. 46 muestra el workflow creado para generar el modelo mediante NaiveBayes, donde se usa un procedimiento similar a los anteriores. En la Fig. 47 se observa el modelo creado.

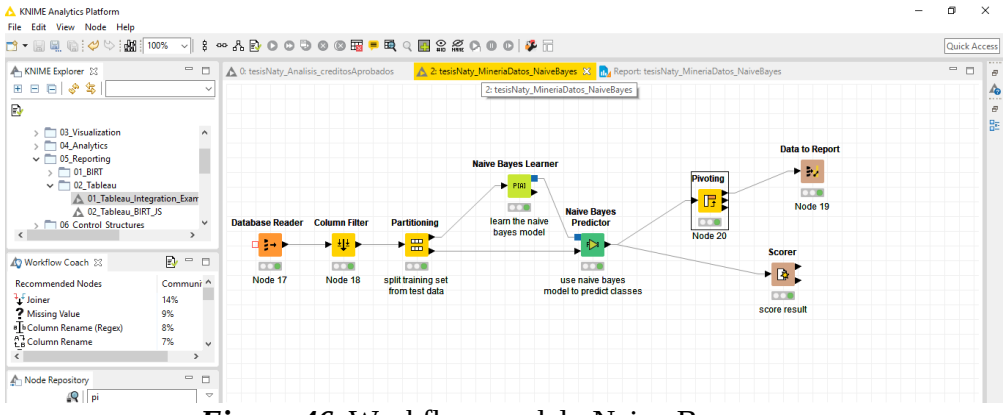
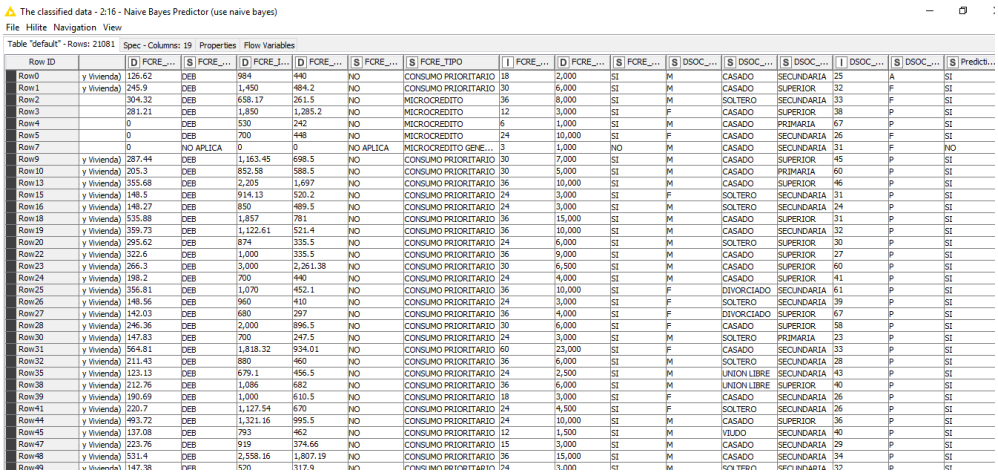


Figura 46. Workflow modelo Naive Bayes



Row ID		D FCRE_...	S FCRE_...	D FCRE_J...	D FCRE_...	S FCRE_...	S FCRE_TIPO	I FCRE_...	D FCRE_...	S FCRE_...	S DSOC_...	S DSOC_...	S DSOC_...	I DSOC_...	S DSOC_...	S Predicti...
Row0	y Vivienda)	126.62	DEB	984	440	NO	CONSUMO PRIORITARIO	18	2,000	SI	M	CASADO	SECUNDARIA	25	A	SI
Row1	y Vivienda)	245.9	DEB	1,450	484.2	NO	CONSUMO PRIORITARIO	30	6,000	SI	M	CASADO	SUPERIOR	32	F	SI
Row2		304.32	DEB	658.17	261.5	NO	MICROCREDITO	36	8,000	SI	M	SOLTERO	SECUNDARIA	33	F	SI
Row3		281.21	DEB	1,850	1,285.2	NO	MICROCREDITO	12	3,000	SI	F	CASADO	SUPERIOR	38	P	SI
Row4		0	DEB	530	242	NO	MICROCREDITO	6	1,000	SI	M	CASADO	PRIMARIA	67	P	SI
Row5		0	DEB	700	448	NO	MICROCREDITO	24	10,000	SI	F	CASADO	SECUNDARIA	26	F	SI
Row7		0	NO APLICA	0	0	NO APLICA	MICROCREDITO GENE...	3	1,000	NO	M	CASADO	SECUNDARIA	31	F	NO
Row9	y Vivienda)	287.44	DEB	1,163.45	698.5	NO	CONSUMO PRIORITARIO	30	7,000	SI	M	CASADO	SUPERIOR	45	P	SI
Row10	y Vivienda)	205.3	DEB	852.58	588.5	NO	CONSUMO PRIORITARIO	30	5,000	SI	M	CASADO	PRIMARIA	60	P	SI
Row13	y Vivienda)	355.68	DEB	2,205	1,697	NO	CONSUMO PRIORITARIO	36	10,000	SI	M	CASADO	SUPERIOR	46	P	SI
Row15	y Vivienda)	148.5	DEB	914.13	520.2	NO	CONSUMO PRIORITARIO	24	3,000	SI	F	SOLTERO	SECUNDARIA	31	P	SI
Row16	y Vivienda)	148.27	DEB	850	489.5	NO	CONSUMO PRIORITARIO	24	3,000	SI	M	SOLTERO	SECUNDARIA	24	P	SI
Row18	y Vivienda)	535.88	DEB	1,857	781	NO	CONSUMO PRIORITARIO	36	15,000	SI	M	CASADO	SUPERIOR	31	P	SI
Row19	y Vivienda)	359.73	DEB	1,122.61	521.4	NO	CONSUMO PRIORITARIO	36	10,000	SI	M	CASADO	SECUNDARIA	32	P	SI
Row20	y Vivienda)	295.62	DEB	874	335.5	NO	CONSUMO PRIORITARIO	24	6,000	SI	M	SOLTERO	SUPERIOR	30	P	SI
Row22	y Vivienda)	322.6	DEB	1,000	335.5	NO	CONSUMO PRIORITARIO	36	9,000	SI	M	CASADO	SUPERIOR	27	P	SI
Row23	y Vivienda)	266.3	DEB	3,000	2,261.38	NO	CONSUMO PRIORITARIO	30	6,500	SI	M	CASADO	SUPERIOR	60	P	SI
Row24	y Vivienda)	198.2	DEB	700	440	NO	CONSUMO PRIORITARIO	24	4,000	SI	M	CASADO	SUPERIOR	41	P	SI
Row25	y Vivienda)	356.81	DEB	1,070	452.1	NO	CONSUMO PRIORITARIO	36	10,000	SI	F	DIVORCIADO	SECUNDARIA	61	P	SI
Row26	y Vivienda)	148.56	DEB	960	410	NO	CONSUMO PRIORITARIO	24	3,000	SI	F	SOLTERO	SECUNDARIA	39	P	SI
Row27	y Vivienda)	142.03	DEB	680	297	NO	CONSUMO PRIORITARIO	36	4,000	SI	F	DIVORCIADO	SUPERIOR	67	P	SI
Row28	y Vivienda)	246.36	DEB	2,000	896.5	NO	CONSUMO PRIORITARIO	30	6,000	SI	F	CASADO	SUPERIOR	58	P	SI
Row30	y Vivienda)	147.83	DEB	700	247.5	NO	CONSUMO PRIORITARIO	24	3,000	SI	M	SOLTERO	PRIMARIA	23	P	SI
Row31	y Vivienda)	564.81	DEB	1,818.32	934.01	NO	CONSUMO PRIORITARIO	60	23,000	SI	F	CASADO	SECUNDARIA	33	P	SI
Row32	y Vivienda)	211.43	DEB	880	460	NO	CONSUMO PRIORITARIO	36	6,000	SI	M	SOLTERO	SECUNDARIA	28	P	SI
Row35	y Vivienda)	123.13	DEB	679.1	456.5	NO	CONSUMO PRIORITARIO	24	2,500	SI	M	UNION LIBRE	SECUNDARIA	43	P	SI
Row38	y Vivienda)	212.76	DEB	1,086	682	NO	CONSUMO PRIORITARIO	36	6,000	SI	M	UNION LIBRE	SUPERIOR	40	P	SI
Row39	y Vivienda)	190.69	DEB	1,000	610.5	NO	CONSUMO PRIORITARIO	18	3,000	SI	F	CASADO	SECUNDARIA	26	P	SI
Row41	y Vivienda)	220.7	DEB	1,127.54	670	NO	CONSUMO PRIORITARIO	24	4,500	SI	F	SOLTERO	SECUNDARIA	26	P	SI
Row44	y Vivienda)	493.72	DEB	1,321.16	995.5	NO	CONSUMO PRIORITARIO	24	10,000	SI	M	CASADO	SUPERIOR	36	P	SI
Row45	y Vivienda)	137.08	DEB	793	462	NO	CONSUMO PRIORITARIO	12	1,500	SI	M	VIUDO	SECUNDARIA	40	P	SI
Row47	y Vivienda)	223.76	DEB	919	374.66	NO	CONSUMO PRIORITARIO	15	3,000	SI	M	CASADO	SECUNDARIA	29	P	SI
Row48	y Vivienda)	531.4	DEB	2,558.16	1,807.19	NO	CONSUMO PRIORITARIO	36	15,000	SI	M	CASADO	SECUNDARIA	34	P	SI
Row49	y Vivienda)	147.38	DEB	530	317.9	NO	CONSUMO PRIORITARIO	24	3,000	SI	M	SOLTERO	SECUNDARIA	32	P	SI

Figura 47. Modelo NaiveBayes

3.13. Fase de evaluación

En cada uno delos wokflows creados por cada técnica se agregó un operando para generar la matriz de confusión y de esta manera evaluar cada modelo. En las Fig48, 49, 50 se encuentran las matrices de confusión de cada modelo.

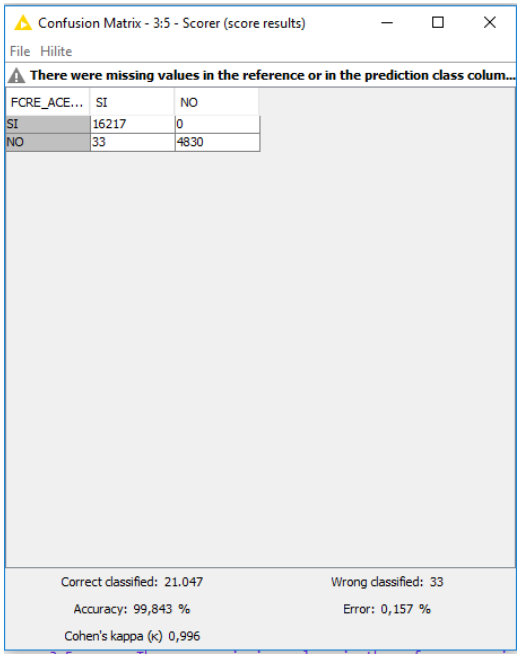


Figura 48. Matriz de confusión modelo Decisión Tree

:

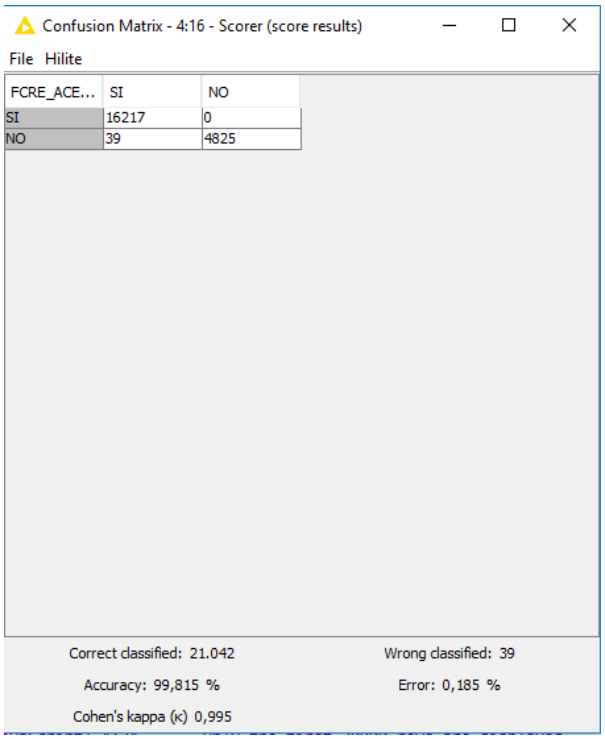


Figura 49. Matriz de confusión modelo Neural Network

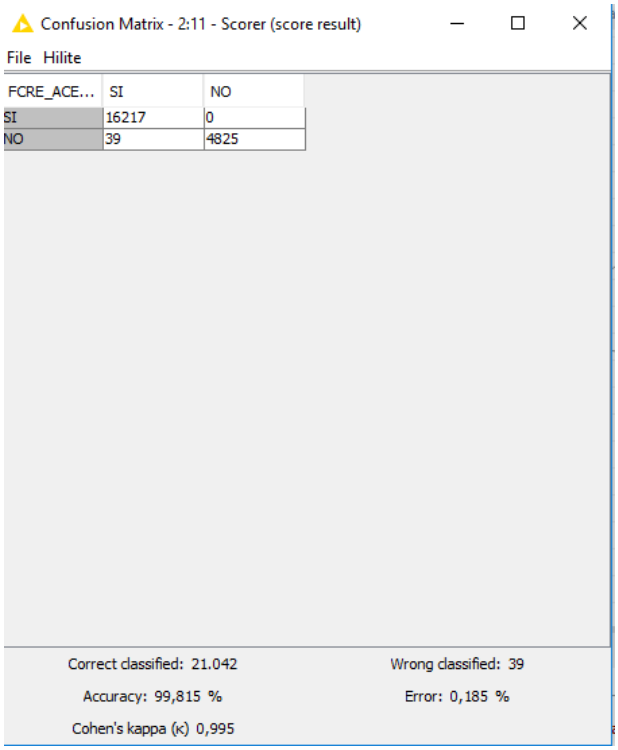


Figura 50. Matriz de confusión modeloNaiveBayes

De acuerdo a las anteriores figuras, se pude resumir lo siguiente:

Tabla 12
Resumen resultados matriz de Confusión

Técnica	Clasificados correctamente	Exactitud	Coeficiente Kappa	Clasificados Incorrectamente	Error
Árbol de decisión	21047	99,843	0,996	33	0,157
Red Neuronal	21042	99,815	0,995	39	0,185
NaiveBayes	21042	99,815	0,995	39	0,185

De acuerdo a la Tabla 6, el número de datos clasificados para las tres técnicas es bastante alta consiguiendo porcentajes exactitud admisibles, de la misma manera porcentajes de error bajos. También se muestran los valores del coeficiente Kappa, que indica que el porcentaje de concordancia de la variable predicha y real son altas. Finalmente, se puede concluir que con un pequeño valor la técnica de Árbol de Decisión es la más exacta.

4. CAPÍTULO IV

RESULTADOS Y CONCLUSIONES

4.1.Resultados

Para presentar los resultados obtenidos mediante este trabajo se realizaron algunos reportes del comportamiento de los datos de acuerdo al valor predicho de FCRE_ACEPTADO que indica si un crédito es aprobado o rechazado.

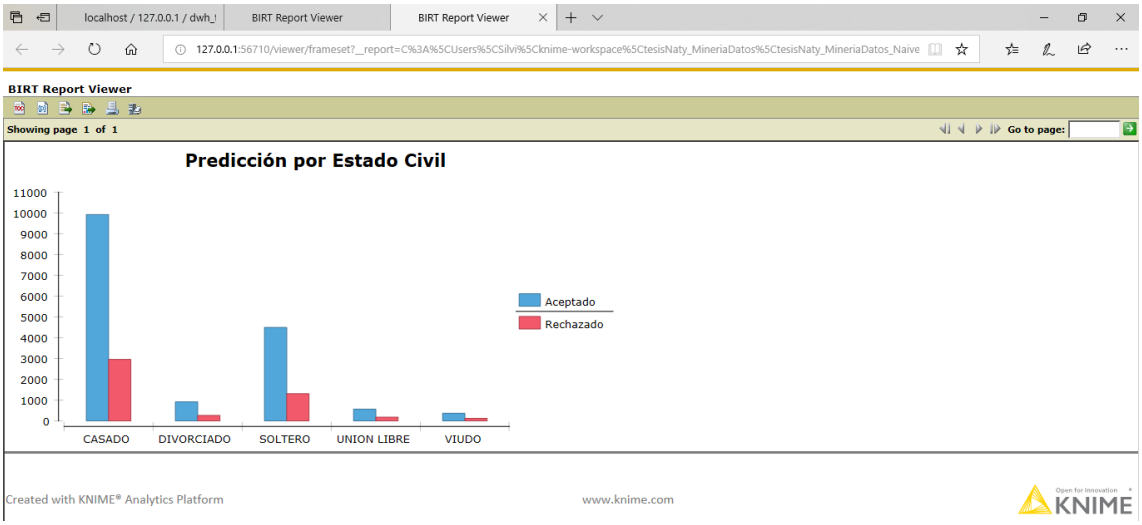


Figura 51. Reporte: Predicción por estado civil

Este reporte indica que mediante el modelo creado una persona en estado civil “soltero” tiene una alta probabilidad de que su crédito sea aprobado mientras que una persona “viuda”, “unión libre” y divorciado” tiene la menor probabilidad de que su crédito sea aprobado. De igual manera una persona “casada” tiene la mayor probabilidad que su crédito sea rechazado.

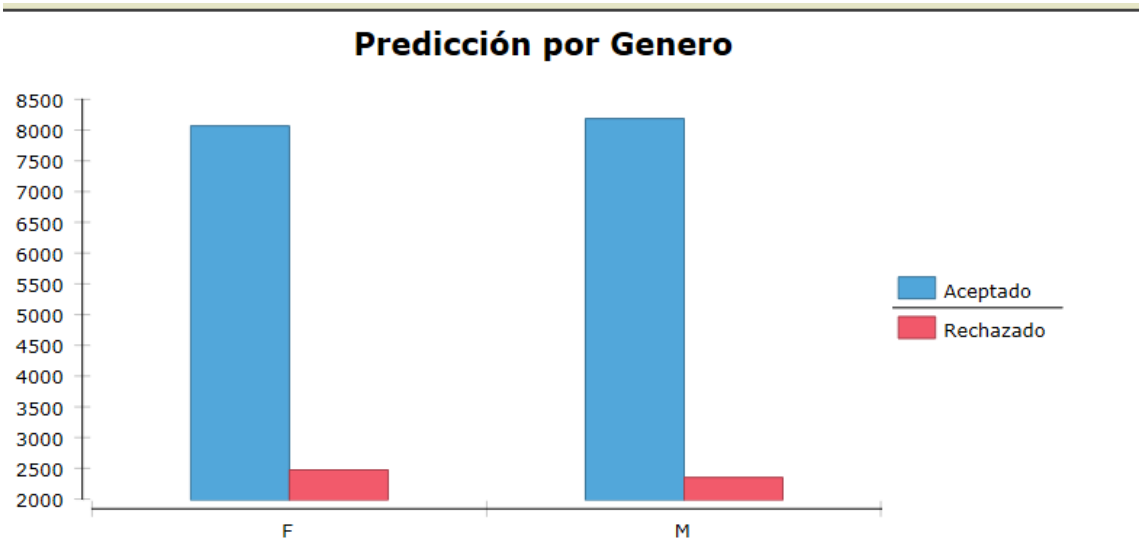


Figura 52. Reporte: Predicción por género

De acuerdo a este reporte es más alta la probabilidad que a una persona le aprueben un crédito a que le rechacen.

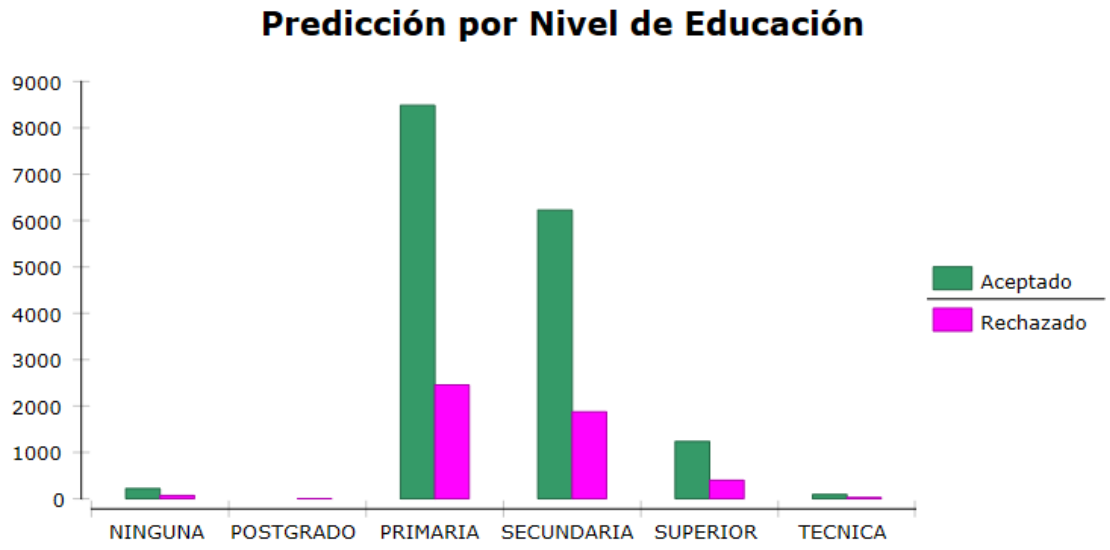


Figura 53. Reporte: Predicción por nivel de educación

Aquí se puede visualizar como las personas con nivel de educación primaria y secundaria tienen mayor probabilidad de que sus créditos sean aprobados. Además, las personas de postgrado tienen la menor probabilidad que les rechacen el crédito.

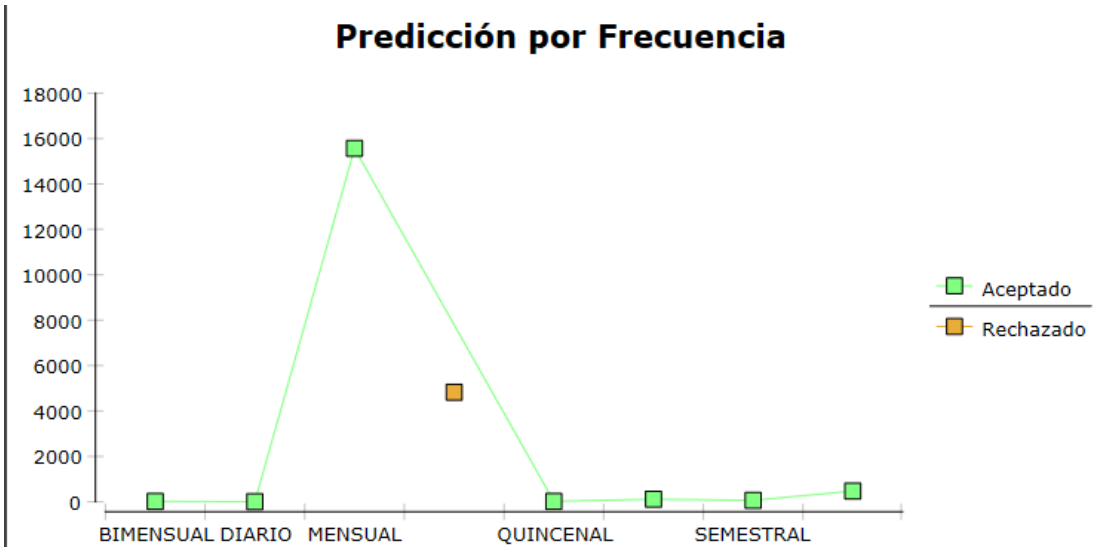


Figura 54. Reporte: Predicción por frecuencia

Con respecto a la frecuencia de pago de las cuotas este reporte indica que la frecuencia mensual tiene mayor probabilidad de ser aceptado.

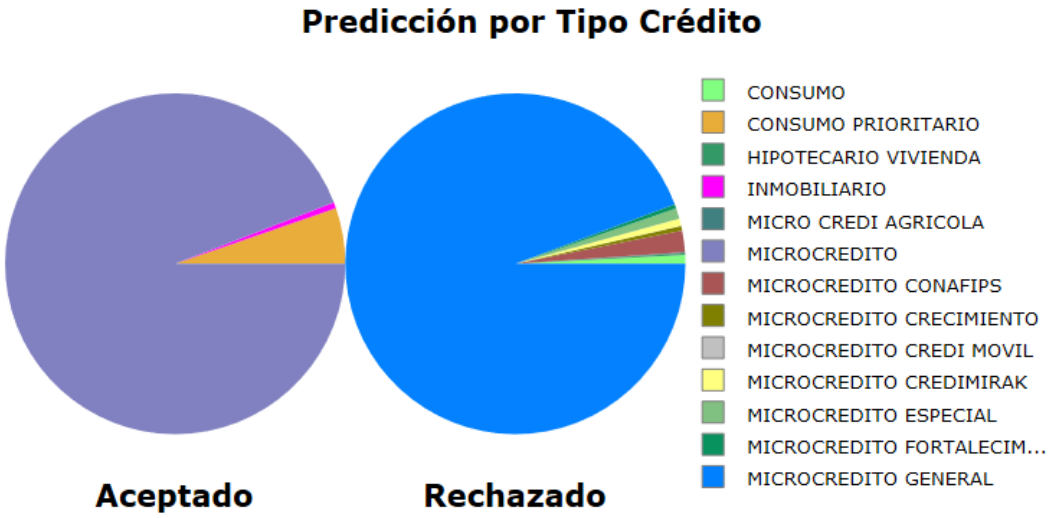


Figura 55. Reporte: Predicción por tipo de crédito

De acuerdo al Tipo de créditos el modelo indica que los créditos “Microcréditos” tiene patrones de ser aceptados, por otro lado, el “Microcrédito General” tienden a ser rechazados.

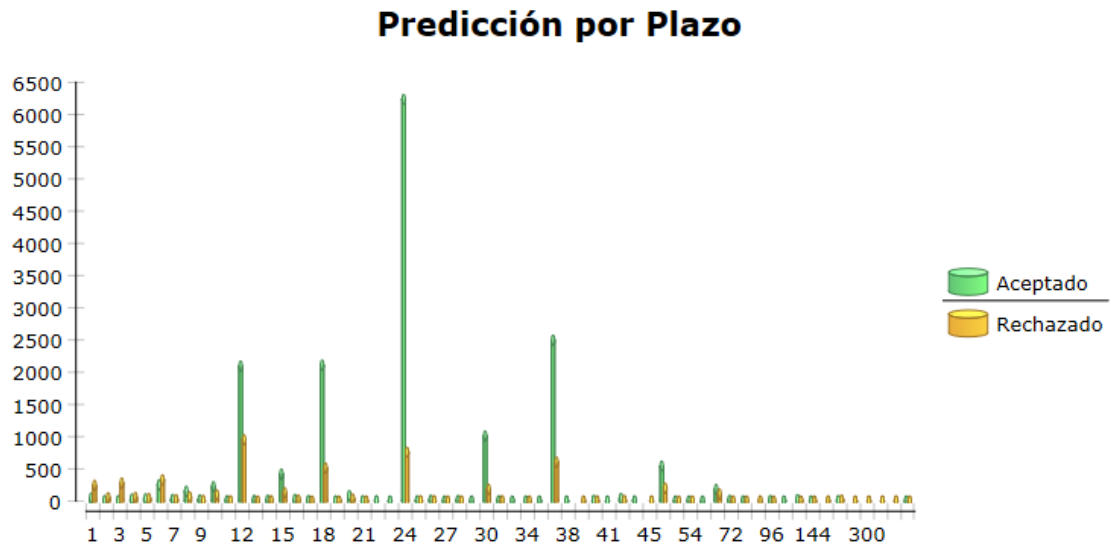


Figura 56. Reporte: Predicción por plazo

Finalmente, mediante un análisis de plazos, los créditos generados para 24 meses tienen más probabilidad de ser aceptados mientras que los de 12 meses tiene mayor predicción para ser rechazados.

4.2.Conclusiones

C1-OE1: La gran cantidad de información de datos generada por las empresas, negocios y en este caso entidades financieras pueden convertirse en valiosas fuentes de conocimiento para la toma de decisiones de los directivos. Este proyecto ha logrado encontrar los patrones de comportamiento de los socios de una cooperativa de ahorro y crédito que permitan clasificarlo en un buen o mal pagador, en base a la información almacenada en bases de datos de la cooperativa, que permitirá reducir tiempos en el área crediticia y asegurar la selección de buenos pagadores de créditos.

C2-OE1: Para la realización del proceso ETL fue necesario determinar los requerimientos de la cooperativa de ahorro y crédito específicamente del área crediticia, donde se encontró los siguientes problemas: créditos mal otorgados, falta de competitividad, deficiencia en procesos

crediticios, y que cliente sin historial crediticio no podían acceder a un crédito. Es por esto que analizando esta problemática se definieron los factores comunes de un socio que accede a un crédito y se realizó un modelo multidimensional, que brindó la posibilidad de establecer relaciones entre los datos recibidos.Además, este modelo permitió reconocer los datos a cargar en la bodega de datos y la estructura para el proceso ETL, puesto que se pudo seleccionar las fuentes de datos para cada ETL, las transformaciones requeridas y nodos necesarios de la herramienta.

C1-OE2: En el proceso de búsqueda de herramientas ETL, minería de datos y gestor de base de datos, se pudieron encontrar una gran variedad de opciones, algunas gratuitas, otras comerciales, otras opensource, con interfaz amigable, multiplataforma, entre otros. Se seleccionó la herramienta KMINE debido a que presentaba un área de trabajo bastante intuitivo y fácil de usar, pero sobre todo permitió realizar el proceso ETL y minería en una misma herramienta. Mientras que para gestor de base de datos se usó *MySQL* por su facilidad de uso, velocidad, rendimiento y bajo costo.

C2-OE2: Para a verificación y obtención de los mejores resultados se realizó tres modelos de minería de datos con las técnicas Decisión Tree, NaiveBayes y Neural Network, que permitió encontrar resultados muy satisfactorios, debido a que la precisión de los tres modelos es bastante aceptable, sobresaliendo por muy poco la técnica Decisión Tree.

C1-OE3: La herramienta KNIME permitió no solo realizar el proceso ETL y minería de datos, sino que fue posible crear reportes, reglas, modelos, gráficos, entre otros para visualizar de una manera más fácil los resultados obtenidos. Además, la herramienta posee características que la hacen óptima para trabajos como el presentado.

C2-OE3: El modelo Decisión Tree indicó que uno de los factores que más predice el comportamiento del socio es el tipo de crédito que se solicita. Por otro lado, el modelo de Neural

Network predice mayormente que los créditos serán aprobados. Además, mediante el modelo NaiveBayes se determinó que más alta es la probabilidad que a una persona le aprueben un crédito a que le rechacen y que el nivel de educación primaria y secundaria tienen mayor probabilidad de que sus créditos sean aprobados. Con respecto a la frecuencia de pago de las cuotas la frecuencia mensual tiene mayor probabilidad de ser aceptado. También se encontró que el tipo de créditos "Microcréditos" tiene patrones de ser aceptados, por otro lado, el "Microcrédito General" tienden a ser rechazados. Finalmente, mediante un análisis de plazos, los créditos generados para 24 meses tienen más probabilidad de ser aceptados mientras que los de 12 meses tiene mayor predicción para ser rechazados.

C1-OE4: La evaluación del modelo predictivo se lo realizó en la misma herramienta donde mediante un operador y el particionamiento de la data en entrenamiento y testeo se pudo obtener la matriz de confusión de cada modelo creado, donde los valores clasificados correctamente fueron rotundamente superior a los clasificados incorrectamente, obteniendo una precisión promedio del 99,8%.

4.3.Recomendaciones

Se recomienda implementar un sistema predictivo basándose en los modelos creados para que sean utilizados por las personas encargadas del análisis de otorgamiento o no de créditos, permitiendo el ahorro de tiempo y confianza en las decisiones tomadas.

Mediante los resultados obtenidos, la cooperativa de Ahorro y Crédito puede segmentar sus productos a clientes específicos en base a los patrones de buenos pagadores, para de esta manera aumentar el número de socios y créditos otorgados mejorando la rentabilidad de la cooperativa de Ahorro y Crédito.

Se recomienda mejorar los reportes presentados utilizando las múltiples ventajas de la herramienta KMINE, por ejemplo, reportes en Tableau. Además, considerar el uso de otros operadores de análisis estadístico y automatización para obtener mejores resultados.

Con los mismos datos se recomienda crear un modelo que determine si un crédito aprobado se castigado o no, por mora en los pagos pendientes y así tomar las acciones pertinentes y que la cooperativa no genere índices de morosidad.

BIBLIOGRAFÍA

Alborzi, M., & Khanbabaei, M. (2016). Using data mining and neural networks techniques to propose a new hybrid customer behaviour analysis and credit scoring model in banking services based on a developed RFM analysis method. *International Journal of Business Information Systems*, *23*(1), 1-22. https://doi.org/10.1504/IJBIS.2016.078020

Benalcazar, J., & Vinueza, J. (2017). A*nálisis comparativo de metodologías de minería de datos y su aplicabilidad a la industria de servicios.* Udla.

Biosilveit. (2016). BioSolveIT - KNIME Interfaces. Recuperado 28 de mayo de 2019, de https://www.biosolveit.de/KNIME/

Carisio, Emanuele. (2018, diciembre 17). Herramientas ETL: comparativa y principales categorías. Recuperado 28 de mayo de 2019, de https://blog.mdcloud.es/herramientas-etl-comparativa-y-principales-categorias/

CHristianCH. (2018). Estudio comparativo entre algoritmos de clasificación "Naive Bayes, Desicion Tree, SVM and Neural Network". Recuperado 24 de junio de 2019, de https://www.authorea.com/users/96193/articles/132726-estudio-comparativo-entre-algoritmos-de-clasificaci%C3%B3n-naive-bayes-desicion-tree-svm-and-neural-network/_show_article

Cooperativa, F. D. (s. f.). Cooperativa de Ahorro y CreditoFernando Daquilema. Recuperado 23 de enero de 2019, de https://www.coopdaquilema.com/

Gahlaut, A., Tushar, & Singh, P. K. (2017). Prediction analysis of risky credit using Data mining classification models. En *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-7). https://doi.org/10.1109/ICCCNT.2017.8203982

Gartner. (2018, julio 16). Cuadrante Mágico de Gartner 2018 de Herramientas de Integración de Datos. Recuperado 28 de mayo de 2019, de https://www.denodo.com/es/pagina/cuadrante-magico-de-gartner-2018-de-herramientas-de-integracion-de-datos

Guido H. Poveda-Burgos, Edison A. Erazo-Flores y Gabriel J. Neira-Vera. (2017). Importancia de las cooperativas en el Ecuador al margen de la Economía. Recuperado 28 de mayo de 2019, de http://www.eumed.net/cursecon/ecolat/ec/2017/cooperativas-ecuador.html

ITpedia. (2018, mayo 13). Data Mining, herramientas para la toma de decisiones. Recuperado 28 de mayo de 2019, de https://es.itpedia.nl/2018/05/13/data-mining-tools-voor-besluitvorming/

Khemakhem, S., & Boujelbene, Y. (2018). Predicting credit risk on the basis of financial and non-financial variables and data mining. *Review of Accounting and Finance*, *17*(3), 316-340. https://doi.org/10.1108/RAF-07-2017-0143

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, *27*, 11-23. https://doi.org/10.1016/j.jretconser.2015.07.003

Lara Hernández, A., Monterrubio Hernández, M., Salazar Hernández, J. C., Bautista Monterrubio, E., Núñez Cárdenas, F. de J., & Sánchez Cruz, J. L. (2014). Herramientas de minería de datos. *Ciencia Huasteca Boletín Científico de La Escuela Superior de Huejutla*, *2*(4). https://doi.org/10.29057/esh.v2i4.1076

Lohokare, J., Dani, R., & Sontakke, S. (2017). Automated data collection for credit score calculation based on financial transactions and social media. En *2017 International Conference on Emerging Trends Innovation in ICT (ICEI)* (pp. 134-138). https://doi.org/10.1109/ETIICT.2017.7977024

Martínez, G. (2011). Minería de datos. *04_724_AgujaAnomala.qxp8_Ciencia_*, 11.

Okesola, O. J., Okokpujie, K. O., Adewale, A. A., John, S. N., & Omoruyi, O. (2017). An Improved Bank Credit Scoring Model: A Naïve Bayesian Approach. En *2017 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 228-233). https://doi.org/10.1109/CSCI.2017.36

Parraga, V., & Zaldumbide, J. P. (2018). *Trabajo de titulación previo a la obtención del título de magíster en: gestión de sistemas de información e inteligencia de negocios*. Universidad de las Fuerzas Armadas ESPE, Sangolqui.

PowerData, R. (2015). ¿Qué es el sistema manejador de bases de datos? Recuperado 28 de mayo de 2019, de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/406549/qu-es-el-sistema-manejador-de-bases-de-datos

Rouse, Margaret. (2015). ¿Qué es Base de datos multidimensional (MDB)? - Definición en WhatIs.com. Recuperado 28 de mayo de 2019, de https://searchdatacenter.techtarget.com/es/definicion/Base-de-datos-multidimensional-MDB

Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems*, *89*, 113-122. https://doi.org/10.1016/j.dss.2016.06.014

Shi, Y. (2012). China's National Personal Credit Scoring System: A Real-life Intelligent Knowledge Application. En *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 406–406). New York, NY, USA: ACM. https://doi.org/10.1145/2339530.2339596

Tello, M. L., Eslava, H. J., & Tobías, L. B. (2013). Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técnicas de minería de datos, 14.