INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA

CLASIFICACION DE CLIENTES CON BASE EN SU HISTORIAL DE PAGOS COMBINANDO ALGORITMOS ENSAMBLADOS Y DEEP LEARNING

TESIS

PARA OBTENER EL GRADO DE

MAESTRO EN SISTEMAS COMPUTACIONALES

P R E S E N T A

ISC. ELVIS JAVIER RECIO CAMPOS

DIRECTOR DE TESIS:
DR. EDDY SÁNCHEZ DE LA CRUZ

CO-ASESOR:
DR. RAJESH ROSHAN BISWAL

MISANTLA, VERACRUZ **FEBRERO 2018**

INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA
DIVISIÓN DE ESTUDIOS PROFESIONALES
AUTORIZACIÓN DE IMPRESIÓN DE TRABAJO DE TITULACIÓN MAESTRÍA

FECHA: 19 de Febrero de 2018.

ASUNTO: **AUTORIZACIÓN DE IMPRESIÓN DE TESIS.**

A QUIEN CORRESPONDA:

Por medio de la presente se hace constar que el (la) C:

ELVIS JAVIER RECIO CAMPOS

estudiante de la maestría en SISTEMAS COMPUTACIONALES con No. de Control 152T0737 ha cumplido satisfactoriamente con lo estipulado por el **Lineamiento de Posgrado para la obtención del grado de Maestría** mediante **Tesis.**

Por tal motivo se **Autoriza** la impresión del **Tema** titulado:

CLASIFICACIÓN DE CLIENTES CON BASE EN SU HISTORIAL DE PAGOS COMBINANDO ALGORITMOS ENSAMBLADOS Y DEEP LEARNING

Dándose un plazo no mayor de un mes de la expedición de la presente a la solicitud del examen para la obtención del grado de maestría.

SEP INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA ITSM
DIVISIÓN DE ESTUDIOS PROFESIONALES

ATENTAMENTE

M.S.C. Eddy Sánchez de la Cruz
Presidente

M.S.C. Galdino Martínez Flores
Secretario

Dr. Rajesh Roshan Biswal
Vocal

Archivo.

DEDICATORIA

A mi querida esposa:

Por su dedicación y apoyo en momentos
difíciles, por confiar y alentarme a seguir con este trabajo.

A mis padres:

Por su apoyo, confianza y motivación para iniciar
y culminar esta etapa de la vida. Por estar pendientes día a día.

A mi hermana:

Por sus palabras de aliento y cariño para motivarme
a culminar, gracias hermanita.

A mis hijas:

Gracias princesas por tratar de entender y ceder
un poco de su tiempo para que lo dedicará a este trabajo.

AGRADECIMIENTOS

Dr. Eddy Sánchez de la Cruz:

Por aceptar ser mi director de tesis, brindar la confianza y dedicación para realizar este trabajo, así como compartir sus conocimientos para tener claro la meta y guiar paso a paso para poder culminar este trabajo. Por brindar su amistad y apoyo incondicional.

Dr. Sídney René Toledo:

Por brindar su amistad, su tiempo y conocimiento, por tu asesoramiento en esta investigación, fuiste un gran pilar junto al Dr. Eddy. Un abrazo amigo.

A mis revisores de tesis:

Maestro Galdino Martínez Flores y Dr. Rajesh Roshan Biswal, por su tiempo y comentarios que tuvieron para que esta tesis se elaborará con los mínimos errores.

A mis compañeros:

Que durante dos años que estuvimos en las aulas, compartimos el conocimiento para que cada uno saliera lo mejor preparado. Un abrazo.

RESUMEN

Este trabajo se origina a partir de la problemática de disminuir el número de clientes morosos de una empresa dedicada a ofrecer servicios de instalación, monitoreo y seguridad tanto residencial como comercial. Para esto, se propone como objetivo clasificar correctamente a sus clientes de acuerdo al historial que cuenta la empresa. El proceso se realiza de la siguiente manera: primero se realizan las diferentes etapas de análisis y pre-proceso de la base de datos aplicada a la información histórica, en la etapa de análisis se toma en cuenta las características que la base de datos muestra originalmente y los criterios para llevar a cabo el proceso de refinamiento de la información (extracción de información). Como siguiente paso se tiene la etapa de pre-procesamiento de datos, en la cual se completa o eliminan valores incompletos o faltantes, se corrigen inconsistencias y se crean nuevos atributos), lo que permite obtener un almacén de datos limpio y libre de errores para su uso en futuras aplicaciones. Posteriormente se aplican los algoritmos de selección de atributos como son: CFS, Chi-Cuadrada, Information gain, random forest y consistencia, para determinar qué atributos son relevantes para definir un modelo de clasificación de clientes, una vez definidos los atributos, se seleccionan los que presentan mejor relevancia basándose en el ranking[1] de atributos generados por los algoritmos de selección de atributos. Una vez seleccionado los atributos, se implementa un modelo de clasificación de clientes aplicando combinatoriamente algoritmos ensamblados con un algoritmo de aprendizaje profundo, para obtener el porcentaje más alto de clasificación, cada uno con tres métodos de evaluación: Use training set, Cross-validation y Percentage Split.

[1] Lista o relación ordenada de atributos con arreglo a un criterio determinado

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ANEXOS

CAPÍTULO 1.
GENERALIDADES

1.1 Introducción

En estos últimos años, las grandes empresas tecnológicas están apostando por el desarrollo y la mejora de algoritmos de reconocimiento de voces, imágenes y textos, así como del autoaprendizaje en sectores como internet, las finanzas, el transporte, el diagnóstico médico o las telecomunicaciones, es decir, el campo de aplicación del Aprendizaje profundo o automático (del inglés *Machine Learning*) es cada vez más elevado. Si bien muchas de estas aplicaciones las utilizamos a menudo no conocemos que realmente disponen de algoritmos de aprendizaje automático (*Back End*). Así, están presentes en la gran mayoría de motores de búsqueda de internet que se personalizan automáticamente según las preferencias del usuario. El Aprendizaje automático, nació como una idea ambiciosa de la Inteligencia Artificial (IA) en la década de los 60 o como una subdisciplina de la IA, producto de las ciencias de la computación y las neurociencias. Lo que esta rama pretendía estudiar era el reconocimiento de patrones (en los procesos de ingeniería, matemáticas, computación, etc.) y el aprendizaje por parte de las computadoras. En los albores de la IA, los investigadores estaban ávidos por encontrar una forma en la cual las computadoras pudieran aprender únicamente basándose en datos. El propósito del Machine Learning es que las personas y las máquinas trabajen de la mano y ser capaces de aprender como un humano. Precisamente esto es lo que hacen los algoritmos, permiten que las máquinas ejecuten tareas, tanto generales como específicas.

En la presente investigación se necesita saber cuáles son las características de los clientes de una empresa que brinda servicio de monitoreo (seguridad) basados en el sistema o modelo automatizado que tiene la empresa para, posteriormente, elegir adecuadamente a sus nuevos clientes y poder tener un estimado de pagos puntuales para realizar la inversión y seguir creciendo. Para ello, se exploran con dos enfoques de la IA: algoritmos ensamblados y las redes neuronales artificiales profundas (RNA). Los resultados muestran porcentajes aceptables de correcta clasificación.

1.2 Descripción del problema

Una empresa que brinda servicio de monitoreo (seguridad) quiere saber cuáles son las características de sus clientes para, posteriormente, elegir adecuadamente a sus nuevos clientes y poder tener un estimado de pagos puntuales para realizar la inversión y seguir creciendo. Para ello, se tiene información histórica en relación a los clientes; sin embargo, no existe un sistema o modelo automatizado que implemente un algoritmo de Aprendizaje automático y clasifique a los clientes potenciales, es decir, esta predicción se hace actualmente de manera empírica por parte del personal de la empresa encargado para tal fin.

1.3 Justificación

El aporte principal del presente trabajo a las ciencias de la computación, es definir de manera específica, mediante la metodología propuesta, un esquema de trabajo ordenado y simplificado, que permita la generación de almacenes de datos limpios y depurados, adecuando la información para la implementación de algoritmos de selección de atributos, generando un conjunto de datos que garantice mejores resultados al momento de la clasificación. Se definirán también, las características que debe cumplir el Dataset, para el correcto funcionamiento de los algoritmos de selección, mostrando el rendimiento que se obtiene al evaluarlos en un modelo de clasificación de clientes utilizando de forma combinada los algoritmos incorporados en *Waikato Enviroment Kmowledge Analysis (WEKA)*.

En el contexto de la empresa, la principal contribución del trabajo realizado, es la generación de un modelo de datos presentado en diversas formas; la primera es el resultado del análisis exploratorio de la base de datos, en este, se integra una nueva base de datos con las tablas que establece una relación entre clientes y los pagos realizados de los diferentes servicios que proporciona la empresa, esto permitirá al negocio, tener claridad de los datos relevantes para el negocio y discriminar información redundante o innecesaria para la empresa. La segunda utilidad de los resultados de esta investigación, es la de proporcionar un almacén de datos integrado, limpio y depurado en formato estándar (.csv) el cual puede ser utilizado por la empresa para distintas aplicaciones computacionales, estadísticas

y/o administrativas, teniendo como base de conocimiento la información de sus clientes y su relación con los pagos; por último, los resultados obtenidos de la implementación de los algoritmos de selección de atributos y su evaluación en un modelo de clasificación de clientes, permitirá a la empresa tener una visión del funcionamiento esperado al realizar aplicaciones futuras que puedan incorporarse al negocio dando la certeza financiera que el negocio necesita.

1.4 Objetivos:

1.4.1 Objetivo general

Distinguir el tipo de cliente según sea su historial de pagos, combinando clasificadores ensamblados y redes neuronales artificiales profundas, para determinar la capacidad crediticia de los clientes con un nivel de confianza aceptable.

1.4.2 Objetivos específicos

- Crear u obtener una base de datos con información histórica de los pagos de los clientes en relación al servicio que la empresa les proporciona.
- Pre-procesar la base de datos.
- Aplicar selección de atributos basados en algoritmos de inteligencia artificial.
- Creación de Datasets para experimentos
- Seleccionar la combinación del mejor clasificador ensamblado con una red neuronal artificial profunda.
- Analizar los resultados con las métricas existentes de Aprendizaje automático.

1.5 Hipótesis

La combinación de clasificadores ensamblados con algoritmo de red neuronal artificial profunda genera buenos porcentajes de clasificación de los tipos de clientes registrados en un Dataset histórico de pagos (±90).

1.6 Propuesta de solución

La metodología propuesta para el presente trabajo consta de 5 pasos, los cuales son descritos a continuación:

1. **Base de datos original**: Se realiza en esta primera etapa el análisis exploratorio de la base de datos para definir un almacén de datos con las tablas que tienen relación con los clientes y los pagos realizados en la empresa. El resultado de este primer análisis, es un conjunto de tablas reducido de tablas.
2. **Pre-procesamiento:** Consiste en la generación un almacén de datos depurado, limpio y sin inconsistencias, para aplicar algoritmos de selección de atributos. El resultado de esta etapa es un archivo con extensión .CSV que contiene los datos adecuados para la etapa de selección de atributos relevantes.
3. **Algoritmos de Selección de Atributos**: En esta etapa se aplican los algoritmos de selección de atributos: CFS, Chi-Cuadrada, Information gain, random forest y consistencia, para determinar los atributos más relevantes para definir un modelo de clasificación de clientes. Se utiliza el paquete Fselector del software R Studio.
4. **Nuevo Dataset rankeado:** Se seleccionarán los atributos que presenten la mejor relevancia basándose en el ranking[2] de atributos generados por los algoritmos de selección de atributos de la etapa 3.
5. **Clasificación con algoritmos de Aprendizaje Profundo:** Se implementa un modelo de clasificación de clientes aplicando combinatoriamente algoritmos de aprendizaje profundo y algoritmos Meta-Clasificadores, obtener el porcentaje más alto de clasificación. Se utiliza el paquete WEKA con el algoritmo Dl4jMlpClassifier y algoritmos Meta-Clasificadores.

[2] Lista o relación ordenada de atributos con arreglo a un criterio determinado

En la figura 1.1, se muestra de manera gráfica la metodología utilizada para la realización de este trabajo de investigación:

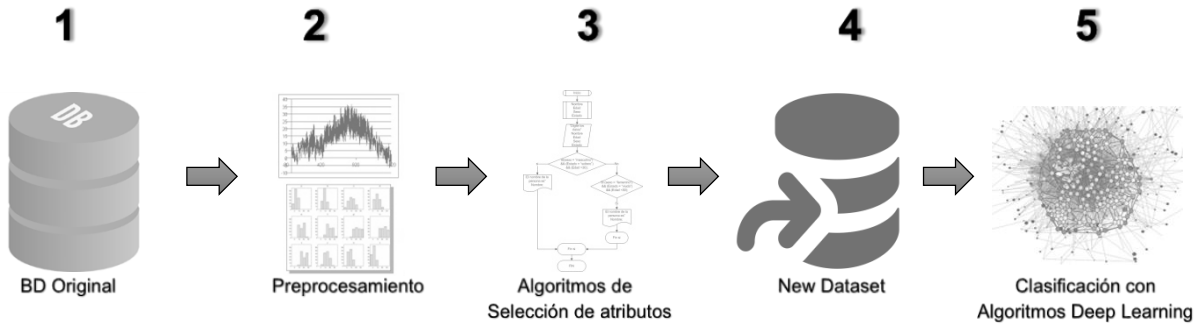


Figura 1.1 Metodología de trabajo de investigación.

1.7 Alcances y limitaciones

Alcances

- Definir una nueva base de datos con la información proveniente del cliente.
- Generar almacenes de datos con diferentes números de atributos y registros.
- Implementar algoritmos de selección de atributos en los almacenes de datos generados, para seleccionar el conjunto de atributos mejor rankeados y eliminar atributos con el menor ranking.
- Definir el modelo de clasificación para realizar análisis de resultados utilizando algoritmos de aprendizaje profundo
- Con los resultados obtenidos se realiza una comparación con los resultados de Toledo, S. (2018).

Limitaciones

- Se utilizan los almacenes de datos generados en la investigación de Toledo, S. (2018), partiendo de la base de datos original de la empresa.
- No se desarrolla software *FrontEnd*.

1.8 Estructura de la tesis

El presente documento se encuentra organizado en cinco capítulos, mismos que a continuación describiremos:

Capítulo I, está compuesto por todos los puntos que rigen la investigación, comenzando por una breve introducción para posteriormente definir el problema, desglosando objetivos concretos que nos ayudaran a limitar y obtener un alcance para poder brindar una solución.

Capítulo II, presentamos los trabajos, documentos e investigaciones relacionados a este trabajo, que nos ayudaran para realizar este proyecto.

Capítulo III, se presenta la metodología que se utiliza para realizar el proyecto, iniciando en la obtención o extracción de datos de un dataset, el pre-procesamiento de la información, la aplicación de algoritmos de selección, hasta llegar a la obtención de resultados.

Capítulo IV, en este capítulo se presenta el análisis e interpretación de los datos de acuerdos a los tres métodos de evaluación que se utilizaron: Use training set, Cross-validation y Percentage Split, con el fin de validar el modelo que se está proponiendo para obtener mejores resultados.

Capítulo V, se presentan las conclusiones que se obtuvieron al utilizar esta técnica y se hacen las recomendaciones a trabajos futuros.

CAPÍTULO 2.
MARCO TEÓRICO

2.1 Pre-procesamiento de datos

El pre-procesamiento de datos es el primer paso en muchos procesos de toma de decisión y de algoritmos de minería de datos, este puede ser necesario o simplemente mejora el rendimiento del algoritmo. Sin embargo, en muchos ejemplos, no recibe la atención adecuada. Las operaciones realizadas durante la fase de pre-procesamiento pueden agruparse en dos categorías. Por un lado, están aquellas técnicas destinadas a detectar y manipular datos considerados imperfectos; y, por otro lado, se consideran aquellas técnicas cuya finalidad es transformar los datos para hacerlos más manejables.

2.1.1 Agrupamiento de Datos

El agrupamiento de datos es una acción que permite reunir todos los datos disponibles para la resolución del problema, en este paso se agrupan datos provenientes de distintas fuentes.

2.1.2 Integración de Datos

El objetivo de la integración de los datos es agrupar juntos todos los datos provenientes de diferentes fuentes, estos pueden tener diferentes formatos. Frecuentemente esta integración de datos se realiza en una base de datos.

2.1.3 Limpieza de Datos

La limpieza de datos consiste en detectar los datos erróneos o irrelevantes y descartarlos. Una de las actividades dentro de la limpieza de datos es el tratamiento de datos ausentes. Esto sucede cuanto falta el valor de un atributo. Para rellenar este valor se pueden tomar diversas estrategias, algunas de las cuales son: utilizar la media o la moda de los valores del entorno, generar un valor aleatorio basándose en una distribución gausiana, algún tipo de interpolación, etc.

Un problema más difícil es la eliminación de los datos ruidosos. Estos casos corresponden con ejemplos que son significativamente diferentes o son inconsistentes con el conjunto de datos.

2.1.4 Selección de Variables y Atributos

En esta fase del pre-procesado de datos, se descartan atributos que no son relevantes para la toma de decisión. En general, el volumen de datos original suele exceder de lo deseable y de lo práctico para su aplicación en la minería de datos. Además, está bien estudiado que gran parte de la información es redundante, principalmente debido a que muchas variables están correlacionadas. Por lo tanto, una importante reducción de atributos puede aplicarse a los datos sin pérdida significativa de información.

El objetivo de la selección de atributo es encontrar el conjunto mínimo de atributos de forma que la distribución resultante de probabilidad de las clases de datos es tan próxima como sea posible a la distribución original usando todos los atributos

2.1.5 Reducción de la Dimensionalidad

En la reducción de la dimensionalidad se aplica una transformación para obtener una representación reducida o comprimida de los datos originales. Si los datos originales pueden ser reconstruidos desde los datos comprimidos sin pérdida de información, entonces la reducción se denomina sin pérdida de datos. Por el contrario, si se puede reconstruir los datos solo de forma aproximada entonces se denomina lossy.

2.1.6 Filtrado de Datos

Durante el filtrado de datos un subconjunto de datos es usado para representar un conjunto de datos más amplio y frecuentemente inmanejable. De forma similar a la selección de atributos, el filtrado de datos trata de eliminar información redundante para obtener buenos modelos con un volumen de datos manejable.

2.1.7 Transformación de Datos

En este paso se construyen nuevos atributos a partir de los atributos originales. Esta transformación puede facilitar una mejor interpretación de la información.

2.2 Aprendizaje automático

El Aprendizaje automático es el subcampo de las ciencias de la computación y una rama de la IA cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento. El aprendizaje automático también se centra en el estudio de la complejidad computacional de los problemas. El aprendizaje automático puede ser visto como un intento de automatizar algunas partes del método científico mediante métodos matemáticos. El aprendizaje automático, además, tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica.

2.2.1 Tipos de Machine learning

- Supervised learning. Depende de datos previamente etiquetados, como podría ser el que una computadora logré distinguir imágenes de coches, de las de aviones. Para esto, lo normal es que estas etiquetas o rótulos sean colocadas por seres humanos para asegurar la efectividad y calidad de los datos. En otras palabras, son problemas que ya hemos resuelto, pero que seguirán surgiendo en un futuro. La idea es que las computadoras aprendan de una multitud de ejemplos, y a partir de ahí puedan hacer el resto de cálculos necesarios para que nosotros no tengamos que volver a ingresar ninguna información. Ejemplos: reconocimiento de voz, detección de spam, reconocimiento de escritura, entre otros.

- Unsupervised learning. En esta categoría lo que sucede es que al algoritmo se le despoja de cualquier etiqueta, de modo que no cuenta con ninguna indicación previa. En cambio, se le provee de una enorme cantidad de datos con las características propias de un objeto (aspectos o partes que conforman a un avión o a un coche, por ej.), para que pueda determinar qué es, a partir de la información recopilada. Ejemplos: detectar morfología en oraciones, clasificar información, etc.

- Reinforcement learning. En este caso particular, la base del aprendizaje es el refuerzo. La máquina es capaz de aprender con base a pruebas y errores en un número de diversas situaciones. Aunque conoce los resultados desde el principio, no sabe cuáles son las mejores decisiones para llegar a obtenerlos. Lo que sucede es que el algoritmo progresivamente va asociando los patrones de éxito, para repetirlos una y otra vez hasta perfeccionarlos y volverse infalible. Ejemplos: navegación de un vehículo en automático, toma de decisiones, etc.

Nuestro trabajo corresponde al tipo de aprendizaje supervisado, ya que depende de datos previamente etiquetados y guardados en la base de datos. La idea es que las computadoras aprendan y reconozcan la información proporcionada y a partir de ahí puedan hacer los cálculos y relaciones necesarias para que nos muestre un resultado sobre los nuevos clientes.

2.3 Aprendizaje profundo

En los albores del siglo XXI, la IA, en sus diversas disciplinas que la integran, comienza a emular más fielmente el comportamiento y razonamiento humano, por lo que se han logrado notables avances en diversas disciplinas del conocimiento que se valen de la IA. Entre las distintas disciplinas que la integran, el aprendizaje profundo es una novedosa alternativa (de hace algunas décadas) que se perfila

para lograr que un autómata[3] sea capaz de tener decisiones propias, algo que actualmente sólo es posible en la ciencia ficción.

El aprendizaje profundo es una disciplina para búsqueda de patrones mediante abstracciones profundas que se logran con múltiples capas ocultas de una RNA. Cárdenas, A. (2012). Para lograr la abstracción, en una capa oculta se selecciona, por ejemplo, el borde de la zona de interés en una imagen (Figura 2.1); la profundidad se alcanza repitiendo la abstracción en tantas capas ocultas se desee.

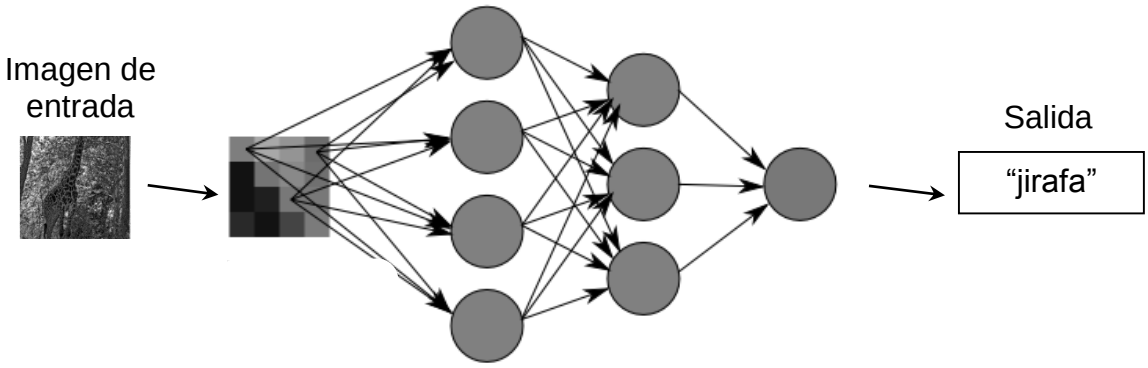


Figura 2.1 Abstracciones en aprendizaje profundo, LeCun, Y. et al. (2015).

La salida de nuestra RNA será la respuesta que buscamos, que puede corresponder a una clasificación que indique que en una imagen de entrada como la de la Figura 2.1, existe la representación de una “jirafa”. En la Figura 2.2 vemos cómo serían las imágenes de entrada y salida en aprendizaje Profundo.

[3] Modelo matemático que, dada una entrada de símbolos, “salta” a través de una serie de estados de acuerdo a una función de transición.

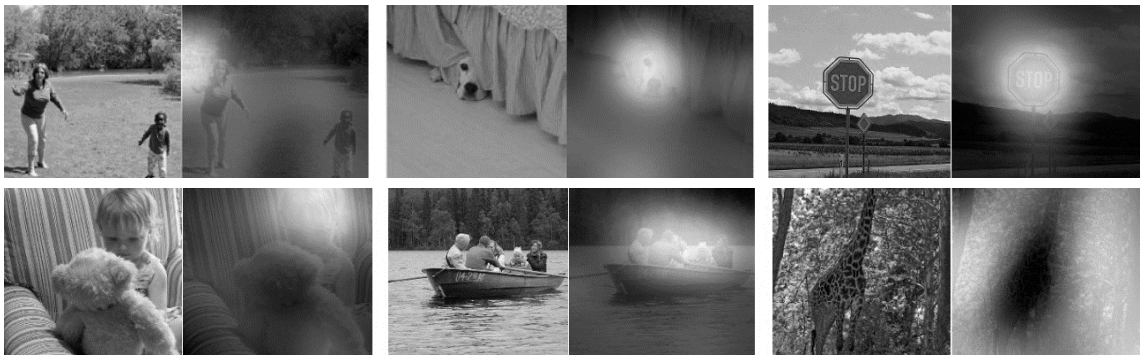


Figura 2.2 Imágenes de entrada y salida en aprendizaje profundo, LeCun, Y. et al. (2015).

Distintos enfoques de RNA son usados en aprendizaje profundo para lograr que un autómata sea equiparable a un humano; entre estos enfoques están: redes neuronales profundas, redes neuronales profundas convolucionales y redes de creencia profundas. A continuación, se expone brevemente el tópico de RNA, el cual nos brinda una comprensión de cómo se lleva a cabo el aprendizaje profundo.

2.3.1 Redes Neuronales Artificiales

Las RNA están inspiradas en la forma como funciona el sistema nervioso biológico y sus primeros inicios datan de los trabajos de Warren McCulloch & Walter Pitts (1943), quienes propusieron el primer modelo formal de neurona biológica. Formalmente se define una RNA como un elemento de procesamiento que recibe un conjunto de entrada $X = \{x_1, x_2, x_3, ..., x_n\}$, y que son modificadas respectivamente por una serie de pesos $W = \{w_1, w_2, w_3, ..., w_n\}$. Los diferentes valores modificados por los pesos se suman en lo que se denomina la entrada neta (la entrada neta es el resultado de la sumatoria de la multiplicación de los valores de entrada por su correspondiente peso, además de un valor *bias* o umbral de la neurona, que se determina cuando se activa la misma). Seguidamente ocurre la activación de la neurona, la cual depende de la función de activación que actúa sobre la entrada neta y se encarga de regular la salida de la neurona.

Como se observa en el diagrama de la Figura 2.3, la sumatoria y la función de activación representan el cuerpo celular de la neurona y allí se realizan los cálculos correspondientes para transmitir el resultado a la salida *y*. Cárdenas, A. (2012).

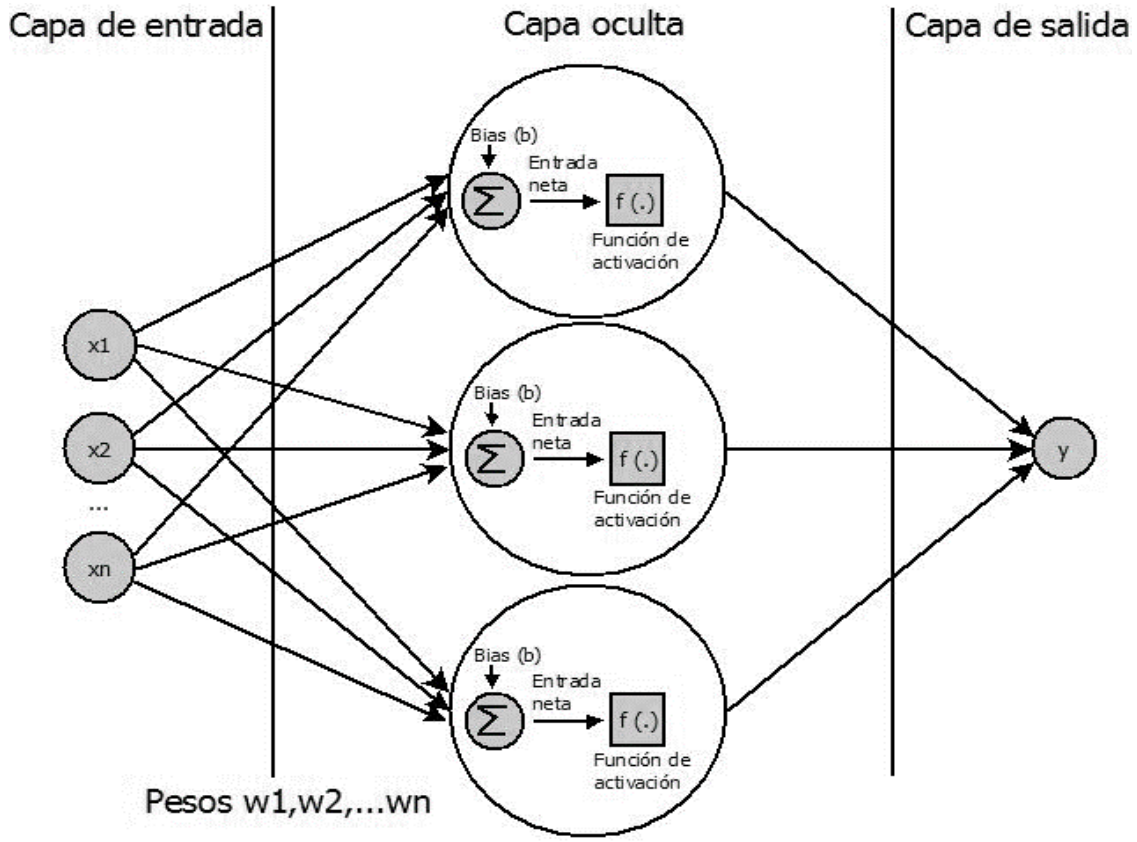


Figura 2.3 Diagrama básico de una RNA.

Una RNA se divide generalmente en tres capas: capa de entrada, capa oculta y capa de salida, comeo se muestra en la Figura 2.3.

2.3.2 Herramientas de implementación para aprendizaje profundo

Existe una gran variedad de herramientas en las que se puede implementar el Aprendizaje profundo. En la Tabla 2.1 se enlistan las que, a conocimiento del autor, son representativas.

Tabla 2.1 *Aplicaciones de Aprendizaje profundo (Deep Learning).*

SOFTWARE	DESARROLLADOR	OPEN SOURCE	ENLACE
Computational Network Toolkit (CNTK)	Microsoft Research	Si	cntk.ai
Deeplearning4j	Adam Gibson	Si	deeplearning4j.org
Caffe	Berkeley Vision and Learning Center	Si	caffe.berkeleyvision.org
Theano	University of Montreal's LISA group	Si	deeplearning.net/software/theano
Thorch	Ronan Collobert, Koray Kavukcuoglu y Clement Farabet	Si	torch.ch
TensorFlow	Google Brain team	Si	tensorflow.org
WEKA	University of Waikato	Si	https://www.cs.waikato.ac.nz/~ml/weka/

2.4 Algoritmos ensamblados (meta-clasificadores)

Los meta-clasificadores, surgen con el objetivo de mejorar la precisión de las predicciones, Witten y Frank (2005), ha surgido un interés creciente en la definición de métodos que combinan varios modelos clasificatorios. Se trata, entonces, de combinar las predicciones del conjunto de modelos, normalmente por votación, para clasificar nuevos ejemplos. La precisión obtenida por esta combinación supera, generalmente, la precisión de cada componente individual del conjunto.

2.4.1 Combinación de meta-clasificadores y aprendizaje profundo

En la presente investigación se combinan todos los clasificadores ensamblados contenidos en la herramienta WEKA, con el algoritmo de aprendizaje profundo Multilayer perceptron Dl4jMlpClassifier (ver Tabla 2.2).

Tabla 2.2 *Combinaciones propuestas de algoritmos META y algoritmos de aprendizaje profundo*.

PRUEBA	META	FUNCION
1		Dl4jMlpClassifier
2	AdaBoostM1	Dl4jMlpClassifier
3	AttributeSelectedClassifier	Dl4jMlpClassifier
4	AutoWEKAClassifier	Dl4jMlpClassifier
5	Bagging	Dl4jMlpClassifier
6	ClassificationViaRegression	Dl4jMlpClassifier
7	CostSensitiveClassifier	Dl4jMlpClassifier
8	CVParameterSelection	Dl4jMlpClassifier
9	FilteredClassifier	Dl4jMlpClassifier
10	GridSearch	Dl4jMlpClassifier
11	IterativeClassifierOptimizer	Dl4jMlpClassifier
12	LogitBoost	Dl4jMlpClassifier
13	MetaCost	Dl4jMlpClassifier
14	MultiClassClassifier	Dl4jMlpClassifier
15	MultiClassClassifierUpdateable	Dl4jMlpClassifier
16	MultiScheme	Dl4jMlpClassifier
17	MultiSearch	Dl4jMlpClassifier
18	OneClassClassifier	Dl4jMlpClassifier
19	OrdinalClassClasifier	Dl4jMlpClassifier
20	RandomCommittee	Dl4jMlpClassifier
21	RandomizableFilteredClassifier	Dl4jMlpClassifier
22	RandomSubSpace	Dl4jMlpClassifier
23	RegressionByDiscretization	Dl4jMlpClassifier
24	Stacking	Dl4jMlpClassifier
25	Vote	Dl4jMlpClassifier
26	WeightedInstancesHandlerWrapper	Dl4jMlpClassifier

2.5 Herramientas computacionales

A continuación, se describen en esta sección las herramientas (o software) utilizados en esta investigación para el pre-procesamiento y clasificación de la información de la base de datos.

Excel. Es una aplicación de hojas de cálculo que forma parte de la suite de oficina Microsoft Office. Es una aplicación utilizada en tareas financieras y contables, con fórmulas, gráficos y un lenguaje de programación.

MySQL WorkBench. Es una herramienta visual de diseño de bases de datos que integra desarrollo de software, Administración de bases de datos, diseño de bases de datos, creación y mantenimiento para el sistema de base de datos MySQL.

R-Studio. Es un conjunto de utilidades de recuperación de datos completamente funcional. Incluye versiones tanto de Windows OS y de Mac OS Linux. Puede recuperar datos de discos duros (HDD), unidades de estado sólido (SSD), memoria flash, y otros dispositivos de almacenamiento de datos internos y externos. Los programas están pensados para especialistas en la recuperación de datos, pero los profesionales de las TI y los usuarios de ordenador normales pueden utilizarlos también para recuperar por sí mismos los archivos perdidos.

WEKA, en español «entorno para análisis del conocimiento de la Universidad de Waikato», es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. WEKA es software libre distribuido bajo la licencia GNU-GPL. WEKA contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades.

En la tabla 2.3 se muestran los paquetes instalados en WEKA para el presente estudio.

Tabla 2.3 *Paquetería instalada en WEKA.*

No.	PAQUETES INSTALADOS		
1	CHIRP	WekaPyScript(Necesito instalar phyton	multilayerPerceptronCS
2	DTNB	alternatingDecisionTrees	multisearch
3	EBMC	bestFirstTree	naiveBayesTree
4	GPAttributeGeneration	classificationViaClustering	oneClassClassifier
5	HMM	clojureClassifier	ordinalClassClassifier
6	IBkLG	complementNaiveBayes	ordinalLearningMethod
7	J48Consolidated	conjunctiveRule	ordinalStochasticDominance
8	J48graft	discriminantAnalysis	ridor
9	JCDT	extraTrees	scriptingClassifiers(instalo kfGroovy y tigerjython)
10	LibLINEAR	functionalTrees	simpleCART
11	LibSVM	fuzzyLaticeReasoning	simpleEducationalLearningSchemes
12	MODLEM	fuzzyUnorderedRuleInduction	thresholdSelector
13	MultiObjectiveEvolutionaryFuzzyClassifier	gridSearch	userClassifier
14	NNge	hiddenNaiveBayes	vfdr(marco error)
15	OpenmlWeka	hyperPipes	votingFeatureIntervals
16	OpenmlWeka(No se pudo instalar porque pide cuenta en openWeka)	kernelLogisticRegression	wekaDeeplearning4jCPU(instalo CPULibs y Core)
17	RBFNetwork	lazyAssociativeClassifier	winnow
18	RankerByDTClassification	lazyBayesianRules	**ForestPA**
19	Rseslib	metaCost	SysFor
20	SPegasos	multiLayerPerceptrons	classificationViaClustering

2.6 Introducción de la Empresa

La empresa SISCOM, ubicada en la ciudad de Boca del Rio, Veracruz, con dirección en la calle 2 de abril del Fraccionamiento Reforma, enfocada en el sector de seguridad tecnológica. Ofrece el nivel más óptimo de soluciones, productos y servicios en el sector de la seguridad con tecnología, dentro de los que se encuentran Central de alarmas, alarmas, cámaras, CCTV, GPS, cercos electrificados, además del servicio de monitoreo de seguridad.

2.7 Estado del Arte

Martín, R., (2017), propone técnicas para la selección y clasificación, desarrollando cuatro variantes de métodos de selección de atributos. Para los experimentos utiliza los algoritmos de selección de atributos CFS-Subset, ReliefF y Chi-Square incluidos en WEKA aplicados a datos históricos del mercado de valores.

Aguilar, J., & Diaz, N., (2005), definen y clasifican los algoritmos de selección de atributos en filter, wrapper e hibrido, y proponen un algoritmo utilizando técnicas de bootstrapping, dividido en cuatro fases: generación de subconjuntos de atributos (generación), evaluación de cada subconjunto (evaluación), actualización los pesos de cada atributo (actualización) y ordenación de atributos por su peso (ordenación). Comparan los resultados obtenidos con las técnicas de bootstrapping y la técnica de búsqueda exhaustiva, definiendo un algoritmo para la generación de ranking de atributos. El resultado de este trabajo concluyó que el método propuesto produce resultados similares al método exhaustivo con la diferencia de realizarlos con un número reducido de pasos.

Liu, H., & Setiono, R., (2005), utilizan el método estadístico de chi cuadrada y la discretización de atributos numéricos, comprueban que esta conversión elimina atributos irrelevantes. Describe el algoritmo utilizando Chi cuadrada para discretizar los atributos numéricos repetidamente hasta que se encuentran algunas inconsistencias en los datos, y logra la selección de características mediante discretización. Los resultados empíricos demuestran que Chi cuadrada es eficaz en la selección de características y la discretización de los atributos numéricos y ordinales. Compara los atributos originales con los atributos generados por el algoritmo determinando que el método chi cuadrada es útil y confiable para la discretización y selección de atributos numéricos.

Ruiz, R., et al., (2005), en este trabajo se presentan distintas formas de comparar los rankings generados por algoritmos de selección de atributos, mostrando la diversidad de interpretaciones posibles en función del enfoque dado al estudio que se realice. Parten de la premisa de la no existencia de un único subconjunto ideal para todos los casos. La finalidad de este algoritmo es reducir el conjunto de datos a los primeros atributos de cada ranking sin perder predicción

frente a los conjuntos de datos originales. Este trabajo propone un método que mide el comportamiento de un ranking de atributos generado y a su vez es válido para comparar diferentes algoritmos de listas de atributos. Para ello, se basa en un nuevo concepto, el área bajo la curva de comportamiento al clasificar un ranking de atributos (AURC). En este trabajo comparan los métodos Chi Cuadrada, Ganancia de Información (IG), RELIEF (RL) y SOAP. Como conclusión de los trabajos recomiendan el uso de los algoritmos SOAP e IG y del clasificador C4.5 cuando se desee clasificar con pocos atributos, y RL y el clasificador 1NN en los demás casos.

Ruiz, R., et al., (2005), combinan la velocidad de los algoritmos de ranking con un método rápido de búsqueda sobre la lista de atributos. El método denominado IRU (Incremental Ranked Usefulness) se basa en la idea de relevancia y redundancia, en el sentido de que un atributo ordenado se escoge si añade información al incluirlo en el subconjunto. Una extensa comparativa con otros métodos de selección, utilizando bases de datos de mediana y alta dimensionalidad, demuestran la eficiencia y la eficacia. Definen el concepto de relevancia y redundancia y realizan la comparación con los métodos de Information Gain (IG), RELIEF (RL) y Selection of Attributes by Projection (SOAP). Concluyeron demostrando que la técnica extrae los mejores atributos no consecutivos del ranking, intentando evitar estadísticamente la influencia de los atributos innecesarios en la clasificación posterior. Esta nueva heurística, denominada IRU, muestra un excelente comportamiento comparado con la técnica tradicional de búsqueda secuencial hacia adelante, no sólo considerando la exactitud de la clasificación en relación con la aproximación filtro, sino también en relación con el coste computacional de la aproximación wrapper

Cruz, R., et al., (2017), presentan resultados del estudio de los algoritmos Naive Bayes, C4.5, Perceptrón multicapa y Kvecinos aplicados a 38 conjuntos de datos con diferentes características, de lo cual resultaron algunas reglas que describen patrones de comportamiento en correspondencia con la población tratada. Los resultados de este trabajo proporcionan una alternativa para decidir qué clasificadores son los mejores para ser utilizados para un conjunto de datos con unas características particulares.

Aboobyda J. & Taring M., (2016), en este documento, se propone un nuevo modelo para clasificar el riesgo de préstamo en el sector bancario mediante el uso de extracción de datos. El modelo se ha construido utilizando datos del sector bancario para predecir el estado de los préstamos. Se han utilizado tres algoritmos para construir el modelo propuesto: j48, bayesNet y naiveBayes. Al usar la aplicación WEKA, el modelo ha sido implementado y probado. Los resultados se discutieron y se realizó una comparación completa entre algoritmos. J48 fue seleccionado como el mejor algoritmo basado en la precisión. Los resultados obtenidos en esta investigación; en la categoría de exactitud de clasificación dicen que con el uso del algoritmo j48 se logró un 78.38%, redes bayesianas 77.47% y bayes ingenuo 73.87%; para obtener estos resultados los autores hacen uso de distintos datos para entrenar los algoritmos, llegando a la conclusión que el algoritmo j48 es el mejor para su modelo predictivo ya que tiene mayor exactitud y un error mínimo en la fase de entrenamiento. En base a esto, la investigación nos muestra que debemos tener un buen conjunto de datos para entrenamiento y posteriormente normalizarlos, para que, en la fase de aprendizaje con el uso del modelo de clasificación de Machine Learning, tenga un alto porcentaje de exactitud y un error mínimo en predicción.

Rivero, J., et al., (2016), aplican la metodología CRISP-DM que define una secuencia de seis pasos los cuales permiten construir e implementar modelos a ser usados en entornos reales, sirviendo así como soporte para la toma de decisiones en los negocios, con lo que implementan de flujos de datos distribuidos permitiendo recopilar información correspondientes a clientes individuales o prospectos, como información geográfica y demográfica (edad, ingresos, miembros de la familia, cumpleaños); psicográfica (actividades, intereses, opiniones) y de comportamiento. Estas bases de datos dan a las compañías y empresas una panorámica de 360 grados de sus clientes atendiendo a su comportamiento. Para las etapas de pre-procesamiento y análisis utilizan las herramientas WEKA, MOA y SAMOA.

CAPÍTULO 3.
METODOLOGÍA

En este capítulo, se describen los trabajos realizados de acuerdo a la metodología utilizada para la presente investigación descrita en la Figura 1.1, la cual consta de 5 etapas: Base de datos original, pre-procesamiento, algoritmos de selección de atributos, integración del nuevo Dataset y la clasificación utilizando algoritmos de Deep Learning. Los primeros cuatro etapas están basadas en Toledo, S. (2018).

3.1 Base de datos original

La fuente de la información que forma el almacén de datos, fue proporcionada por la empresa., misma que está formada por un total de 126 tablas dentro una base de datos llamada ISSAI, la cual cuenta en su totalidad con 1181 atributos y 130446 registros, conteniendo información histórica de sus clientes del año 2010 al año 2017, la información de la base de datos se exploró utilizando MySQL Workbench (Anexo 1).

3.1.1 Extracción y selección de información de la Base de Datos Original

Para la selección de las tablas que formarán las propuestas de almacén de datos, se tomaron en cuenta dos criterios, el primero es, incluir las tablas que contengan información que relacione a los clientes y el comportamiento que tiene con los pagos que realiza por su servicio. El resultado de la aplicación del primer criterio, se muestra en la Figura 3.1.

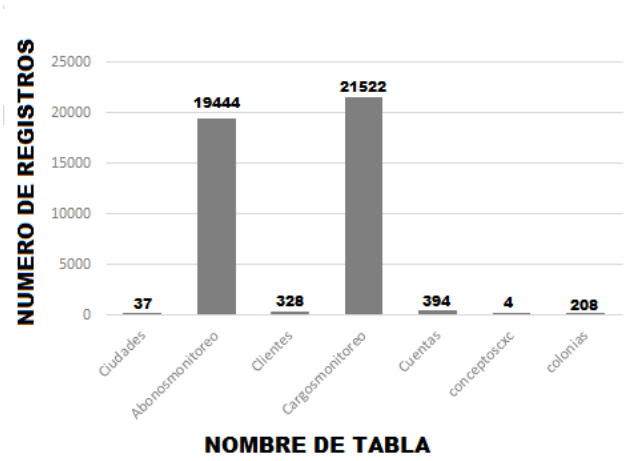


Figura. 3.1 Gráfico resultado de la selección aplicando el criterio 1.

El segundo criterio es, incluir las tablas que contengan la mayor cantidad de registros posibles después de aplicar el criterio 1 antes mencionado, excluyendo todas las tablas que tengan el menor número de registros. Por esto, que se decidió dividir en dos segmentos la información; en el primer segmento, se encuentran las tablas clientes, cuenta y colonias, con un número de registros comprendidos entre 208 y 394 registros (ver Figura 3.2) al cual nos referiremos como segmento 1; mientras que el siguiente segmento, llamado segmento 2, se encuentran las tablas abonosmonitoreo y cargosmonitoreo que representan las tablas con mayor número de registros comprendidos entre 19444 y 21522 respectivamente (ver Figura 3.3).

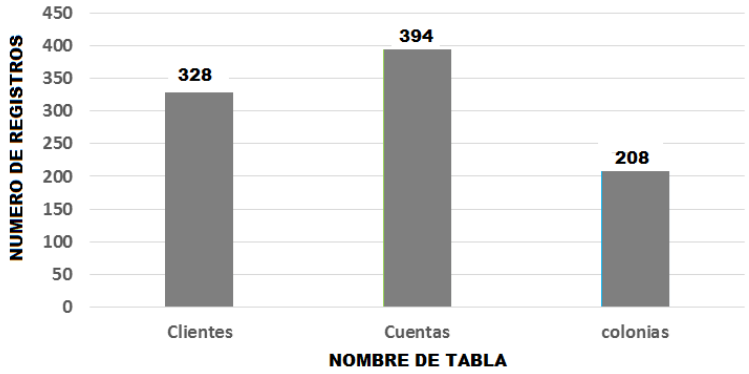


Figura 3.2 Gráfica aplicando el Criterio 2 – Segmento 1.

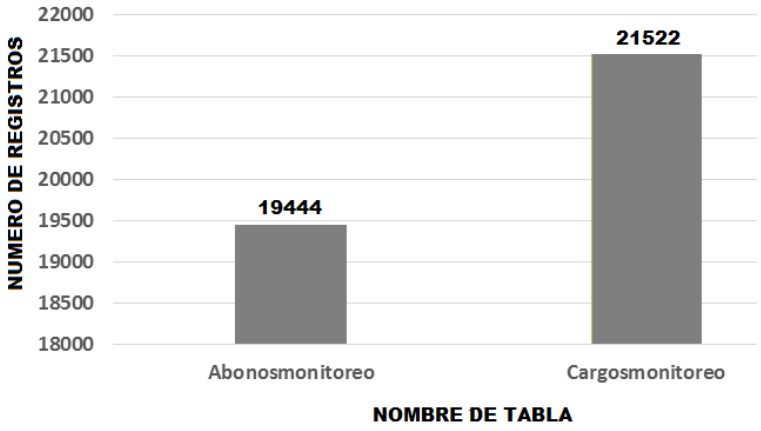


Figura 3.3 Gráfica aplicando el Criterio 2 – Segmento 2.

Con estos segmentos, se excluyen del almacén de datos, las tablas *ciudades* con 37 registros y *conceptoscxc* con 4 registros, ya que representan las tablas con menor número de registros.

3.1.2 Integración de los segmentos de datos

Una vez identificado los segmentos importantes de registros en el almacén de datos, se integraron las tablas correspondientes a cada segmento, quedando conformados como se muestra en la Tabla 3.1.

Tabla 3.1 *Integración de Segmento de la base de datos aplicando el criterio 2.*

Segmento 1

TABLA	NO. REGISTRO	NO. ATRIBUTOS
Clientes	328	29
Cuentas	394	18
Colonias	208	4
	TOTAL ATRIBUTOS	51

Segmento 2

TABLA	NO. REGISTRO	NO. ATRIBUTOS
Abonosmonitoreo	328	29
Cargosmonitoreo	394	18
	TOTAL ATRIBUTOS	47

Cada uno de los segmentos fueron realizados en Microsoft Excel, en el caso del primer segmento, quedo formado por 51 atributos con 394 registros, mientras que el segundo segmento, queda formado por un total de 42 atributos y 21520 registros.

Con este primer filtrado de la información de la base de datos, se observó que la información no cuentea con una correspondencia lineal con cada uno de los registros, esto debido a que la tabla *cuentas* tiene una relación establecida con la tabla *clientes*, por lo que la integración lineal, ocasiona que los registros queden desordenados al momento de juntar las tablas de cada segmento ; además con esta operación, se reduce el número de tablas, en función de la relación *cliente-cargos-abonos* y de la cantidad de atributos y registros del almacén de datos. Identificando la relación existente entre las tablas *cargosmonitoreo*, *abonosmonitoreo*, *cuentas* y *clientes*, por lo que se tuvo que realizar un seguimiento de los datos en todas las tablas obteniendo lo siguiente información:

1. La tabla *cargosmonitoreo* por medio del atributo CMOClave tiene relación con el atributo AMOCargo de la tabla *abonosmonitoreo*.
2. La tabla *cargosmonitoreo* por medio del atributo CMOCuenta tiene relación con el atributo CTAClave de la tabla cuentas.
3. La tabla *cuentas* por medio del campo CTACliente tiene relación con el atributo CLIClave de la tabla clientes.

Basados en este análisis realizado, se toma la determinación de integrar un único Dataset, que contenga los atributos de las tablas *abonosmonitoreo, cargosmonitoreo, cuentas y clientes* (Tabla 3.2), ya que estas representan la información central de los clientes y su comportamiento con respecto a los pagos de sus servicios en la empresa.

Tabla 3.2 *Tablas seleccionadas para el almacén de datos.*

TABLA	NO. REGISTROS
Abonosmonitoreo	19444
Clientes	328
Cargosmonitoreo	21522
Cuentas	394

Para realizar la integración del almacén de datos, se realizó una base de datos llamada siscomBD, utilizando MySQL Workbench que contiene solamente las tablas que se obtienen como resultado de la extracción y selección de la base de datos original, estas tablas son *abonosmonitoreo*, *cargosmonitoreo*, *clientes* y *cuentas*. El modelo lógico de la base de datos, después del proceso de extracción y selección exploratoria de la base de datos original de la empresa, queda conformado como se muestra en la Figura 3.4.

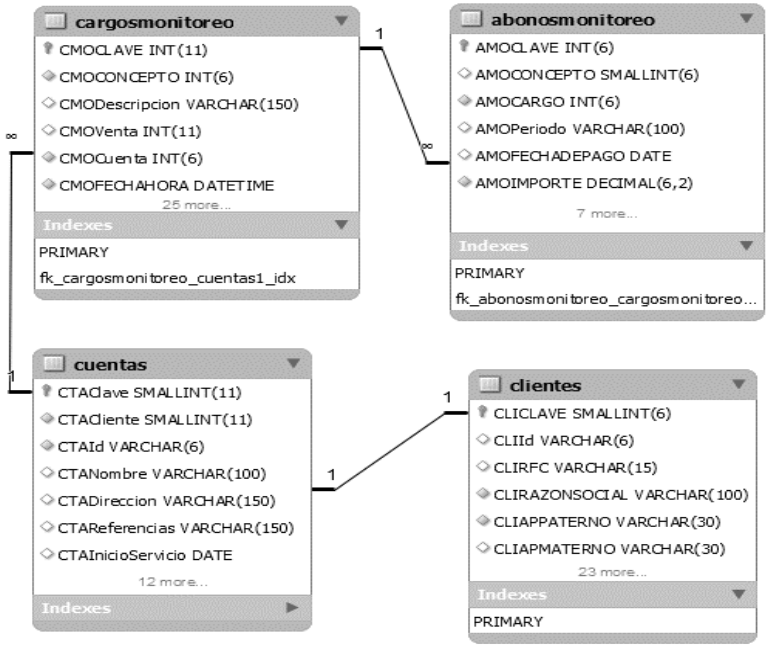


Figura 3.4 Modelo lógico de la Base de datos SISCOM generado.

3.1.3 Integración de almacenes de datos

Con la base de datos implementada en la etapa de extracción mostrada en la Figura 3.4, se realizaron dos propuestas de almacén de datos, las cuales se diferencian por el tipo de relación que se establece para integrar los almacenes de datos. Para ambas propuestas es necesario tener presente que el procedimiento de la empresa establece que, para cada uno de los servicios prestados (venta, instalación o monitoreo mensual), se genera un cargo de manera automática en la tabla *cargomonitoreo*, una vez que este cargo es pagado, se realiza un registro en la tabla abonosmonitoreo el cual relaciona al cargo generado y esta a su vez con la cuenta y datos del cliente. Teniendo en cuenta este funcionamiento interno de la empresa, las propuestas se integran de la siguiente manera:

- Propuesta 1. Relación Abono-Cargo (abonomonitoreo-cargosmonitoreo): Esta propuesta de almacén de datos, contiene los atributos que el experto de la empresa considera que son necesarios para integrar la aplicación, se maneja la tabla abonosmonitoreo como principal, para definir el conjunto de registros que formaran el Dataset, tomando en cuenta que esta tabla, almacena los pagos reales realizados por los clientes para cada uno de los cargos almacenados en la tabla cargosmonitoreo. Es por ello que el número de registros es similar al tamaño de la tabla abonosmonitoreo. Esta propuesta de Dataset cuenta con 19443 registros un total de 25 atributos (ver Anexo 2).
- Propuesta 2. Relación Cargo-Abono (*cargosmonitoreo-abonosmonitoreo*)**:** Esta propuesta de almacén de datos, surge como resultado del análisis de los datos, para ello, se toma en cuenta para su integración, la importancia que recae sobre la tabla *cargosmonitoreo*, ya que en esta tabla como se menciona anteriormente, se registran cada uno de los pagos que deben realizar los clientes, mientras que en la tabla *abonosmonitoreo* se registran los pagos una vez realizados por los clientes, con respecto al cargo generado, es por ello que en la tabla *cargosmonitoreo* existen una cantidad mayor de registros que en la tabla *abonosmonitoreo*, ya que estos faltantes de datos en la tabla de *abonosmonitoreo*, representan pagos que no se han realizado y que por lo tanto no se registran en esta tabla. Por lo anterior esta propuesta contiene todos los atributos de las tablas *cargosmonitoreo*, *abonosmonitoreo*, *cuentas* y *clientes*. Esta propuesta de Dataset cuenta con 89 atributos y 22245 registros (ver Anexo 3).

La integración de la consulta de la propuesta 1 y propuesta 2 implementadas en Mysql WorkBench, se muestran en el Anexo 4 y 5 respectivamente.

Con estas actividades, culmina la etapa 1 de “Análisis de la base de datos original”, teniendo los Dataset que se utilizan para la etapa 2 de la metodología (Figura 1.1), correspondiente al “Pre-procesamiento de datos”.

3.2 Pre-procesamiento de datos.

Una vez que se integraron las propuestas de Dataset, el siguiente paso es realizar la etapa de pre-procesamiento de acuerdo a la metodología de la investigación, para esto se realizó en cada uno de los almacenes de datos propuestos, lo siguiente: completar o eliminar valores incompletos o faltantes, detectar valores atípicos (outliers), corregir datos inconsistentes, creación de atributos nuevos, eliminación de atributos no aceptados, transformar atributos categóricos a numéricos y fragmentar fechas.

3.2.1 Paso 1: Completar o eliminar valores incompletos o faltantes

Esta técnica permite asignar valores al atributo o conjunto de atributos que no tengan asignado valor, las cuales son necesarios para el proceso de integrar el almacén de datos para el proceso de clasificación, dicha ausencia de valores en los atributos, puede ser originada porque no fueron ingresados al momento de la captura, errores técnicos al momento de extraer la información o por no ser considerados relevantes para el sistema en cuestión. El procedimiento para completar estos valores faltantes, implica alguno de las siguientes acciones: ignorar la tupla, completar el dato faltante a mano, usar una constante global, usar el valor medio del atributo moda para el atributo faltante o que pertenezcan a la misma clase, usar el valor más probable basado en inferencia (formulas bayesianas o arboles de decisión).

Para determinar el número de valores faltantes, se realizó un filtrado de los atributos de cada uno de los Dataset propuestos, especificando el número de valores NULL y el número de datos que tenían valores faltantes, utilizando las siguientes fórmulas de Microsoft Excel: =CONTAR.SI(rango,"NULL") y =CONTAR.BLANCO(rango). Con lo que se obtuvo lo siguiente:

- Para el Dataset 1, no se encontró ningún atributo con espacios en blanco, ni tampoco valores nulos en su contenido. Por lo que no se tuvo que realizar ningún proceso de sustitución ni de eliminación (ver Anexo 6).

- Para el Dataset 2, se encontró un total de 163927 espacios en blanco, que correspondían a omisiones al momento de la captura del registro y, un total de 194244 valores con el valor NULL, correspondiente a errores de omisión al momento de la captura y porque no existían abonos para todos los cargos que estaban generados. Dichos campos vacíos, fueron sustituidos por el valor 0 para el caso de los atributos numéricos, con el valor S/D (Sin Definir) para el caso de los atributos tipo texto y en algunos casos, con el valor que representaba el valor de la moda en ese atributo (ver Anexo 7). El detalle de las acciones realizadas en el Dataset 2, se pueden observar en el Anexo 8.

3.2.2 Paso 2: Detectar valores atípicos (outliers)

Por medio de filtros en Microsoft Excel y de manera exploratoria, se determinó que ninguno de los Dataset, tenían valores atípicos en sus atributos, esto debido a que la información correspondía a claves integradas por números secuenciales, fechas y descripciones o conceptos de tipo texto, además de contar con campos categorizados de manera numérica, solo en el caso del atributo CMOIMPORTE se encontraban datos que salían de los parámetros normales de información del atributo, esto debido a que existen servicios que tienen un mayor importe, comparado con el parámetro moda de este atributo, correspondiente al pago de monitoreo mensual.

3.2.3 Paso 3: Corregir inconsistencias.

Se modificó la información del Dataset 1 y Dataset 2, debido a que la información contenida en estos campos presentaba problemas en la codificación del texto, no apareciendo la letra ñ ni las letras acentuadas de manera correcta por lo que, se modificación por la letra n y por las letras sin acentos. Se modificó para ambos almacenes el formato de los atributos tipo fecha, esto con la finalidad de unificar los campos y poder utilizarlos en operaciones futuras, ya que se encontraban con el formato dd/mm/yyyy hr:min:seg y, para hacer las operaciones solo se necesitaba el formato dd/mm/yyyy.

En resumen, para el almacén propuesto 1, se realizó un total de 47 valores inconsistentes y 1563 reemplazos (ver Anexo 9), mientras que para el almacén 2, se realizaron un total de 69 valores inconsistentes y 4331 reemplazos en la información almacenada (ver Anexo 10).

3.2.4 Paso 4: Creación de atributos nuevos.

Se agregaron en ambos Dataset tres atributos más que llamados, diasPlazo, diferenciasDias y tipoCliente, mismos que son necesarios para categorizar cada uno de los registros almacenados agregando un criterio de clasificación a la información. El Atributo diasPlazo almacenará los días que tiene de vigencia para realizar el abono para cada uno de los cargos generados, esto atributo tiene como resultado la diferencia entre el atributo CMOFECHADEVENC que tiene la fecha donde vence el cargo realizado y el atributo CMOFECHAHORA que contiene la fecha donde se genera el cargo, el atributo diferenciasDias contiene la suma del atributo días de plazo(diasPlazo) con la diferencia que existe entre la fecha de abono(AMOFechaDePago) y la fecha de vencimiento del plazo(CMOFECHADEVENC)(ver Formula 3.1 y 3.2). El atributo tipoCliente contiene un valor nominal de acuerdo al atributo diferenciasDias (ver Tabla 3.3).

$$diasPlazo = (Fecha\ de\ Vencimiento\ del\ Cargo - Fecha\ de\ Ingreso\ del\ Cargo) \quad 3.1$$

$$diferenciasDias = diasPlazo + (Fecha\ del\ pago - Fecha\ de\ vencimiento\ del\ Cargo) \quad 3.2$$

Tabla 3.3 *Valor nominal para al atributo tipoCliente.*

CONDICION ATRIBUTO diferenciasDias	VALOR
Si valor es negativo(-)	Anticipado
Si el valor es entre 0 y 10	Normal
Si el valor es mayor a 10	Moroso

Para el campo tipoCliente, se utilizó la función SI e Y de Excel, mostrado en la fórmula 3.3.

$$= si\left(AA2 < 0, \text{"ANTICIPADO"}, SI(AA2 > 10, \text{"MOROSO"}, "NORMAL")\right) \qquad 3.3$$

Para realizar esta operación se tuvo que modificar en ambos Dataset, el formato de los campos CMOFECHAHORA, CMOFECHADEVENC, AMOFechaDePago, ya que se encontraban con el formato dd/mm/yyyy hr:min:seg y, para hacer las operaciones solo se necesitaba el formato dd/mm/yyyy. Esto permite categorizar de forma nominal (ANTICIPADO, NORMAL y MOROSO) a cada uno de los registros contenidos en los dos Dataset, tomando como base del valor del atributo diferenciasDias (ver Tabla 3.3), con ello se obtiene la columna de clasificación que es utilizada en la implementación de los algoritmos de selección de atributos y en las pruebas a realizar con los algoritmos de clasificación de WEKA.

Con esta operación queda procesado el Dataset 1 contando con un total de 28 atributos, de los cuales son 8 son de tipo texto, 17 numéricos y los 3 restantes son del tipo fecha y, 19441 registros correspondientes al primer Dataset. Mientras que el Dataset 2 queda formado con 92 atributos, de los cuales 26 son de tipo texto, 53 de tipo numérico y 10 de tipo fecha contando con 22245 registros. Se generan los archivos con extensión .csv separados con comas para su utilización en R. En la Figura 3.5, se muestran cada uno de los almacenes propuestos con la clasificación basada en el atributo tipoCliente generado con el software R.

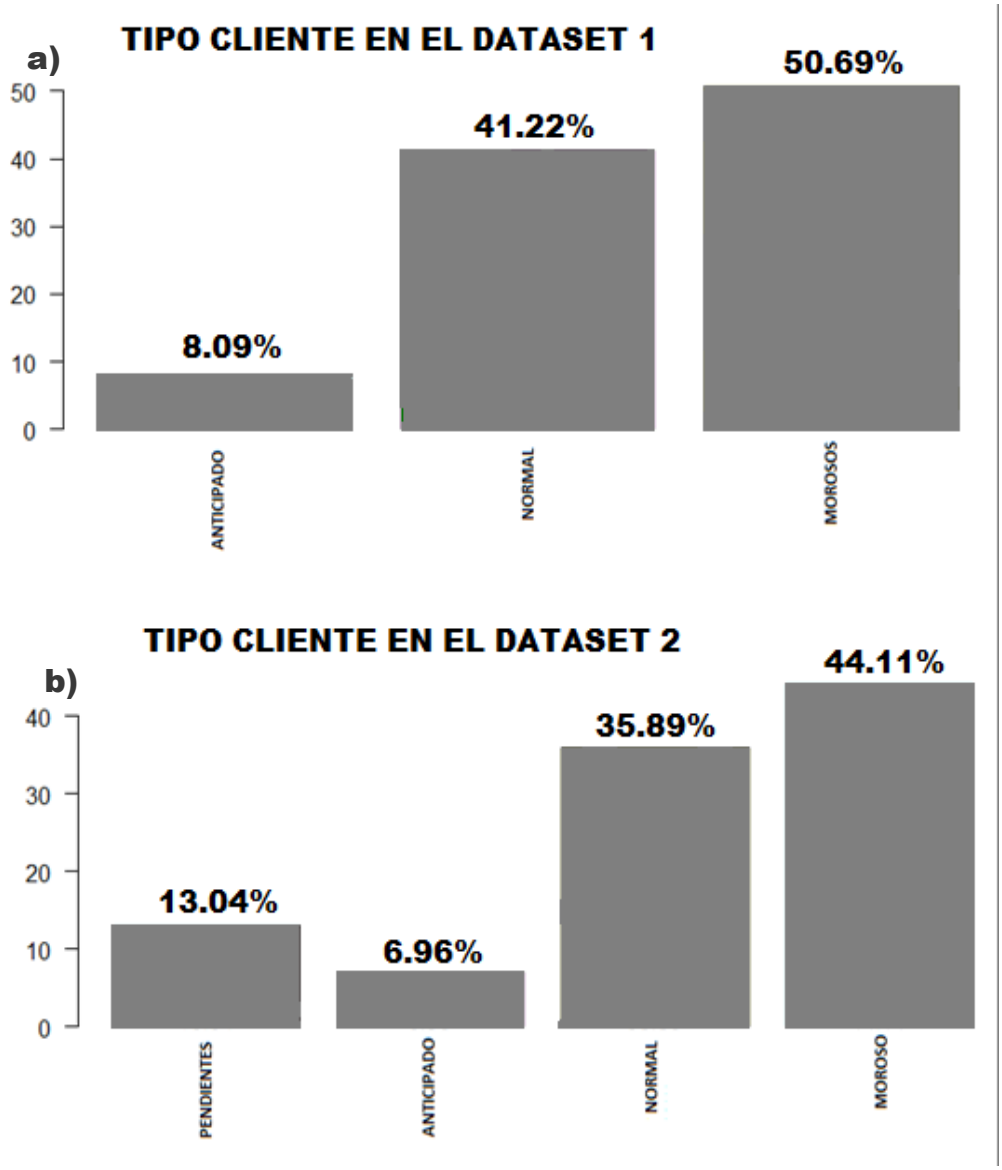


Figura 3.5 Registros categorizados en R usando, a) Dataset 1 y b) Dataset 2.

3.2.5 Paso 5: Eliminación de atributos no aceptados

Una de las operaciones que se realizó a cada uno de los Dataset, es la eliminación de atributos no aceptados por los algoritmos antes mencionados, estos atributos por el tipo de información que contienen representan para los algoritmos un número de categorías no aceptados para su funcionamiento (aproximadamente 53 categorías). Dentro de estos atributos se encuentran las claves, id, direcciones, descripciones como nombres, apellidos o campos con información de edades.

Ante esto se revisaron cada uno de los Dataset y se eliminaron los atributos que contenían este tipo de información y que representa información de poca relevancia para la clasificación.

3.2.6 Paso 6: Transformar atributos categóricos a numérico**s**

En esta etapa se realizó la conversión de datos nominales a datos numéricos para los campos que tenían información de tipo nominal, debido a que los algoritmos de selección de atributos aplicados, requieren valores numéricos en su estructura, provocando errores en la ejecución si no son transformados, por ejemplo, el atributo CMOSTATUS que tenía como valores 3 tipos de información “PAGADO”, “CONDONADO” O “PENDIENTE”, se sustituyó por los valores numéricos 1, 2 y 3 respectivamente. Otro ejemplo es el atributo tipoCliente que tenía como valor nominal “ANTICIPADO”, “NORMAL” O “MOROSO”, se sustituyeron por los valores numéricos 1, 2 y 3 también respectivamente. Algunos atributos solo fue necesario transformar a dos valores numéricos, 1 para valores “Si” y 2 para valores “No”. Para los atributos que no estaban definidos y que se etiquetaron con el valor “S/D” (Sin Definir), se sustituyeron por valores numéricos -1.

3.2.7 Paso 7: Fragmentar fechas.

Este proceso consistió en dividir los campos tipo fecha en el formato dd/mm/yyyy en atributos separados, es decir, un atributo contiene el día, otro el mes y otro el año, generando la división de un atributo en 3 atributos respectivamente. La nomenclatura para etiquetar los atributos fue anteponiendo la palabra día, mes o anio al nombre del atributo, de esta forma por ejemplo el atributo CMOFECHAVENC, fue dividido en diaCMOFECHADEVENC, mesCMOFECHADEVENC y anioFECHADECVENC. Con esta transformación de los almacenes de datos propuestos para los algoritmos de selección de atributos, se generaron dos propuestas adicionales de Datasets con la cantidad de atributos presentados en la Tabla 3.4.

Tabla 3.4 *Dataset originales y generados para aplicar algoritmos de selección de atributos.*

DATASET	CANTIDAD DE ATRIBUTOS
1.- Dataset Original de la Propuesta 1	11 Atributos
2.-Dataset generado de la propuesta 1 con la transformación	28 atributos
3.-Dataset Original de la Propuesta 2	72 atributos
4.-Dataset generado de la propuesta 1 con la transformación	92 Atributos

3.3 Implementación de algoritmos de selección de atributos

Para realizar la implementación de los algoritmos de selección de atributos contenidos en el paquete FSelector, se tiene que instalar el paquete, utilizando la instrucción install.package ("FSelector"), para este trabajo se utiliza la versión 0.21 del paquete y se utiliza R en su versión 3.4.1.

Como se menciona anteriormente, los algoritmos aplicados en este trabajo son Random Forest (RF), CFS, Chi Square(X^2), Consistency e Information Gain (IG), cuyo funcionamiento se encuentra descrito en el capítulo 2.

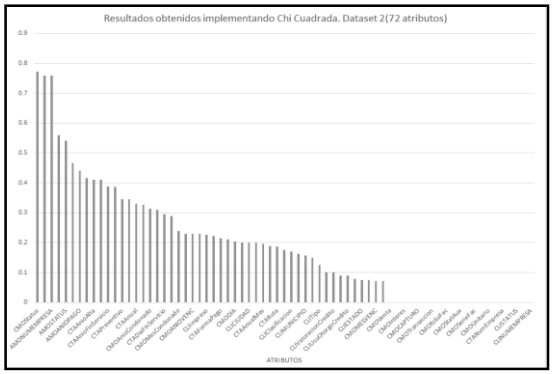
Con esto se realizan los experimentos, aplicando los 5 algoritmos de selección de atributos a cada uno de los Dataset, por la naturaleza de la información contenidas en los almacenes de datos, algunos de los algoritmos no se implementaron, generando errores al momento de su ejecución, mientras que otros algoritmos funcionaron correctamente. Se identifican de manera visual con los resultados arrojados por los algoritmos de selección, que los atributos BAFechaDia, diferenciasDias, CMODia, AMOTRANSACCION, CMOStatus, CTAAnioAlta y AMOClave como los atributos mejor rankeados y/o seleccionados para la clasificación del tipo de cliente con respecto a los pagos realizados. El resumen de la cantidad de algoritmos implementados para cada uno de los Dataset, se muestra en la Tabla 3.5.

Tabla 3.5 *Resultados de la Ejecución correcta de los algoritmos de selección de atributos para cada Dataset.*

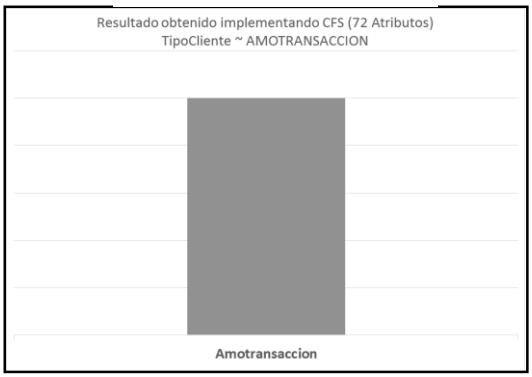
ALGORITMO	DATASET 1 CON 11 ATRIBUTOS	DATASET 2 CON 28 ATRIBUTOS	DATASET 3 CON 72 ATRIBUTOS	DATASET 4 CON 92 ATRIBUTOS
RF	✓	X	✓	X
CFS	✓	✓	✓	✓
X2	X	✓	✓	✓
Consistency	X	✓	✓	✓
IG	X	✓	✓	✓
TOTAL DE ALGORITMOS IMPLEMENTADOS	**2**	**4**	**5**	**4**

Es necesario tener en cuenta, que los algoritmos de selección de atributos, generan una salida que puede ser en dos formas distintas, la primera generando un Dataset (subset) con los atributos más relevantes, la segunda, generando una salida con un valor ponderado que representa la importancia que tiene cada atributo con respecto atributo de clasificación (ranking). Tomando en cuenta los resultados obtenidos de la aplicación de los algoritmos de selección de atributos, se concluye que el Dataset formado por 72 atributos, es el que genera resultados para todos los algoritmos de selección, descartando los Dataset 1, 2 y 4 debido a que no generan resultados en algunos de los algoritmos implementados (Tabla 3.5). En la Figura 3.7 se muestran los resultados obtenidos utilizando el Dataset 3 formado por 72 atributos. Este Dataset con 72 atributos para el presente trabajo, representa el almacén de datos extraído de la base de datos original, por lo que es considerado la base de datos origen y que será utilizado para realizar el clasificador en WEKA como propuesta del experto de la empresa.

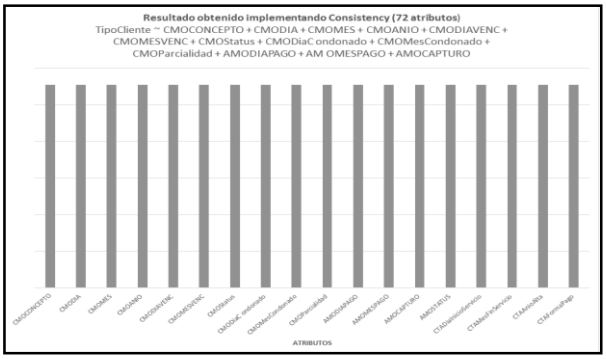
ALGORITMO CHI SQUARE



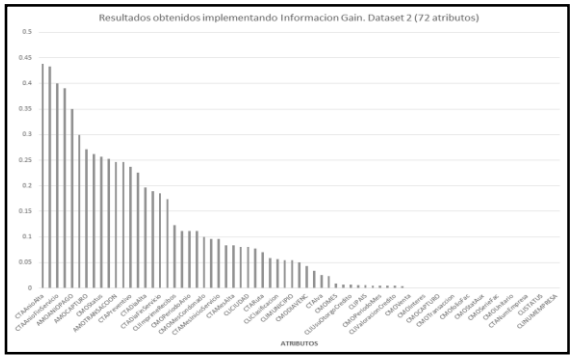
ALGORITMO CFS



ALGORITMO CONSISTENCY



ALGORITMO INFORMATION



ALGORITMO RANDOM FOREST

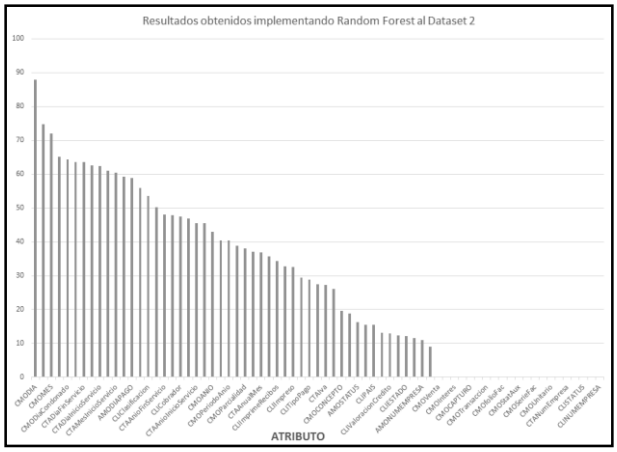


Figura 3.6 Conclusión de la implementación de Algoritmos de Selección de Atributos.

3.4 Generación del nuevo Dataset rankeado.

La cuarta etapa en la metodología propuesta en la investigación (ver Figura 1.1), se refiere a la integración de un nuevo Dataset, el cual debe contener los atributos mejor rankeados y/o seleccionados que resulten de la implementación de los algoritmos de selección de atributos. Para esta operación se realizó la selección de los primeros 18 atributos que resultaron de la implementación de los algoritmos de selección de atributos para el Dataset 3 (formado por 72 atributos); el corte se realizó basado el resultado obtenido por el algoritmo CONSISTENCY, el cual genero un subset con 18 atributos (ver Figura 3.15), por lo que fueron seleccionados los 18 atributos mejores rankeados de los algoritmo RF, X^2 y IG, mientras que del algoritmo CFS se seleccionó el único atributo que genero su implementación(ver Figura 3.6), como se mencionó anteriormente, todos para el Dataset 3 (72 atributos). La selección de los atributos seleccionados de cada algoritmo de selección se muestra en la Tabla 3.6, generando un conjunto de 73 atributos mejores rankeados.

Tabla 3.6 *Selección de 18 atributos mejores rankeados para el nuevo Dataset.*

RESULTADOS DE LOS ALGORITMOS DE SELECCIÓN DE ATRIBUTOS (DATASET CON 72 ATRIBUTOS)

No	CFS	CONSISTENCY	CHI CUADRADA		INFORMATION GAIN		RANDOM FOREST	
	RESULTADO	RESULTADO	RESULTADO	PONDERACIÓN	RESULTADO	PONDERACIÓN	RESULTADO	PONDERACIÓN
1	AMOTRANSACCION	CMOCONCEPTO	CMOStatus	0.77298074	CTAAnioAlta	0.438232362	CMODIA	0.87992152
2		CMODIA	AMOTRANSACCION	0.75924584	CTAAnioInicioServicio	0.433305766	CMOPeriodoMes	0.74767689
3		CMOMES	AMONUMEMPRESA	0.75924584	CTAAnioFinServicio	0.399424048	CMOMES	0.71953121
4		CMOANIO	AMOCAPTURO	0.55925449	AMODIAPAGO	0.390848371	CMOMesCondonado	0.65120306
5		CMODIAVENC	AMOSTATUS	0.54096126	AMOANIOPAGO	0.349845194	CMODiaCondonado	0.64305978
6		CMOMESVENC	AMODIAPAGO	0.4671396	AMOCONCEPTO	0.29938341	CTADiaAlta	0.63575542
7		CMOStatus	AMOANIOPAGO	0.43939471	AMOCAPTURO	0.271407555	CTADiaFinServicio	0.63537698
8		CMODiaCondonado	AMOCONCEPTO	0.41584584	AMOMESPAGO	0.262337661	CMOMESVENC	0.62504861
9		CMOMesCondonado	CTAAnioAlta	0.41106303	CMOStatus	0.256801097	CTADiaInicioServicio	0.62337037
10		CMOParcialidad	CTAAnioInicioServicio	0.41039924	AMOSTATUS	0.252421239	CTAMesFinServicio	0.61064146
11		AMODIAPAGO	CTAAnioFinServicio	0.38886267	AMOTRANSACCION	0.246823821	CTAMesInicioServicio	0.60434236
12		AMOMESPAGO	AMOMESPAGO	0.38586419	AMONUMEMPRESA	0.246823821	CTAMesAlta	0.59223178
13		AMOCAPTURO	CTAPreventivo	0.34499872	CTAPreventivo	0.237142969	AMODIAPAGO	0.58791476
14		AMOSTATUS	CMODiaCondonado	0.34456409	CMODiaCondonado	0.225133713	CTAFormaPago	0.56016124
15		CTADiaInicioServicio	CTAAnual	0.33112466	CTADiaAlta	0.196599268	CLIClasificacion	0.53495545
16		CTAMesFinServicio	CLIImprimeRecibos	0.32703373	CMOAnioCondonado	0.18964052	CTARuta	0.50220442
17		CTAAnioAlta	CMOAnioCondonado	0.31257274	CTADiaFinServicio	0.185318002	CTAAnioFinServicio	0.4812131
18		CTAFormaPago	CTADiaAlta	0.31028424	CTADiaInicioServicio	0.173893505	CLICIUDAD	0.47896182

La actividad siguiente que se realizó con este conjunto de 73 atributos, fue visualizar las coincidencias existentes de los atributos en los diferentes algoritmos, esto con el objetivo de cuantificar el número de algoritmos en los que se repite cada atributo y con ello determinar la importancia de este último con respecto la clasificación de clientes. El proceso para cuantificar la coincidencia de un atributo en cada algoritmo se muestra en la Tabla 3.7.

Tabla 3.7 *Resultados de aplicar algoritmos de selección de atributos para el Dataset con 72 atributos.*

RESULTADOS DE LOS ALGORITMOS DE SELECCIÓN DE ATRIBUTOS (DATASET CON 72 ATRIBUTOS)

No	CFS	CONSISTENCY	CHI CUADRADA		INFORMATION GAIN		RANDOM FOREST	
	RESULTADO	RESULTADO	RESULTADO	PONDERACIÓN	RESULTADO	PONDERACIÓN	RESULTADO	PONDERACIÓN
1	AMOTRANSACCION	CMOCONCEPTO	CMOStatus	0.77298074	CTAAnioAlta	0.438232362	CMODIA	0.87992152
2		CMODIA	AMOTRANSACCION	0.75924584	CTAAnioInicioServicio	0.433305766	CMOPeriodoMes	0.74767689
3		CMOMES	AMONUMEMPRESA	0.75924584	CTAAnioFinServicio	0.399424048	CMOMES	0.71953121
4		CMOANIO	AMOCAPTURO	0.55925449	AMODIAPAGO	0.390848371	CMOMesCondonado	0.65120306
5		CMODIAVENC	AMOSTATUS	0.54096126	AMOANIOPAGO	0.349845194	CMODiaCondonado	0.64305978
6		CMOMESVENC	AMODIAPAGO	0.4671396	AMOCONCEPTO	0.29938341	CTADiaAlta	0.63575542
7		CMOStatus	AMOANIOPAGO	0.43939471	AMOCAPTURO	0.271407555	CTADiaFinServicio	0.63537698
8		CMODiaCondonado	AMOCONCEPTO	0.41584584	AMOMESPAGO	0.262337661	CMOMESVENC	0.62504861
9		CMOMesCondonado	CTAAnioAlta	0.41106303	CMOStatus	0.256801097	CTADiaInicioServicio	0.62337037
10		CMOParcialidad	CTAAnioInicioServicio	0.41039924	AMOSTATUS	0.252421239	CTAMesFinServicio	0.61064146
11		AMODIAPAGO	CTAAnioFinServicio	0.38886267	AMOTRANSACCION	0.246823821	CTAMesInicioServicio	0.60434236
12		AMOMESPAGO	AMOMESPAGO	0.38586419	AMONUMEMPRESA	0.246823821	CTAMesAlta	0.59223178
13		AMOCAPTURO	CTAPreventivo	0.34499872	CTAPreventivo	0.237142969	AMODIAPAGO	0.58791476
14		AMOSTATUS	CMODiaCondonado	0.34456409	CMODiaCondonado	0.225133713	CTAFormaPago	0.56016124
15		CTADiaInicioServicio	CTAAnual	0.33112466	CTADiaAlta	0.196599268	CLIClasificacion	0.53495545
16		CTAMesFinServicio	CLIImprimeRecibos	0.32703373	CMOAnioCondonado	0.18964052	CTARuta	0.50220442
17		CTAAnioAlta	CMOAnioCondonado	0.31257274	CTADiaFinServicio	0.185318002	CTAAnioFinServicio	0.4812131
18		CTAFormaPago	CTADiaAlta	0.31028424	CTADiaInicioServicio	0.173893505	CLICIUDAD	0.47896182

Los colores identifican los atributos repetidos que coinciden en cada uno de los algoritmos, el resultado del número de coincidencia o repetición junto con el valor máximo en alguna de las coincidencias para en cada algoritmo, se encuentra resumida Tabla 3.8.

Tabla 3.8 *Numero de Coincidencias de los atributos y valor de ponderación máximo.*

ORDENADO POR COINCIDENCIA EN LOS ALGORITMOS					
ATRIBUTO	**COINCIDENCIAS**	**VALOR MAXIMO**	**ATRIBUTO**	**COINCIDENCIAS**	**VALOR MAXIMO**
CMODiaCondonado	4	0.64305978	AMOANIOPAGO	2	0.43939471
AMODIAPAGO	4	0.58791476	AMOCONCEPTO	2	0.41584584
AMOTRANSACCION	3	0.75924584	CTAAnioInicioServicio	2	0.433305766
CMOStatus	3	0.77298074	CTAPreventivo	2	0.34499872
AMOMESPAGO	3	0.38586419	CMOAnioCondonado	2	0.31257274
AMOCAPTURO	3	0.55925449	CTADiaFinServicio	2	0.63537698
AMOSTATUS	3	0.54096126	CMOCONCEPTO	1	Sin ponderacion
CTADiaInicioServicio	2	0.62337037	CMOANIO	1	Sin ponderacion
CTAAnioAlta	3	0.438232362	CMODIAVENC	1	Sin ponderacion
CTAAnioFinServicio	3	0.4812131	CMOParcialidad	1	Sin ponderacion
CTADiaAlta	2	0.63575542	CTAAnual	1	Sin ponderacion
CMODIA	2	0.87992152	CLIImprimeRecibos	1	Sin ponderacion
CMOMES	2	0.71953121	CMOPeriodoMes	1	0.74767689
CMOMESVENC	2	0.62504861	CTAMesInicioServicio	1	0.60434236
CMOMesCondonado	2	0.65120306	CTAMesAlta	1	0.59223178
CTAMesFinServicio	2	0.61064146	CLIClasificacion	1	0.53495545
CTAFormaPago	1	0.56016124	CTARuta	1	0.50220442
AMONUMEMPRESA	2	0.75924584	CLICIUDAD	0	0.47896182

En la Tabla 3.8, se observa lo siguiente: no existe ningún atributo que coincida en los 5 algoritmos, existen 2 atributos que aparecen en 4 algoritmos, 9 atributos aparecen en 3 algoritmos, son 13 atributos que aparecen en al menos 2 algoritmos, y por último apareciendo de manera única en alguno de los algoritmos tenemos 12 atributos, de los cuales 6 de ellos no tiene valor de ponderación, debido a que fueron resultado de un algoritmo que no genera ranking sino subset. Con estos resultados, se identifican claramente la cantidad de coincidencias que existen en los atributos generados por los algoritmos de selección de atributos y los valores máximos de ponderación, que permiten definir un ranking de relevancia de los atributos basándose en el comportamiento que tienen los clientes con relación a sus pagos.

Como conclusión, se determina, generar un nuevo Dataset con 30 atributos, tomando como base el número de coincidencia o apariciones en los algoritmos (4, 3, 2 y 1 veces), excluyendo a los 6 atributos que solo aparecen en un solo algoritmo y que no tienen valor de ponderación, debido a que no se pueden considerar dentro del nuevo Dataset por desconocer la importancia cuantificable que tienen para la clasificación. Por lo que este nuevo Dataset formado por 30 atributos es considerado como el Dataset formado con los atributos mejores rankeados resultado de la implementación de los algoritmos de selección de atributos.

Para la realización de la etapa 5 de la metodología “Análisis de Resultados” (ver Figura 1.1), es necesario aplicar en el nuevo Dataset generado los algoritmos de aprendizaje profundo, los cuales se describen en el siguiente capítulo.

CAPÍTULO 4.
EXPERIMENTOS DE CLASIFICACIÓN Y RESULTADOS

4.1 Introducción

Para la generación de los resultados mostrados en esta sección, se realiza la clasificación del nuevo Dataset generado utilizando algoritmos ensamblados y algoritmos de aprendizaje profundo, los cuales permiten realizar un análisis comparativo del rendimiento con respecto a lo señalado en el trabajo de Toledo, S. (2018), utilizando meta-clasificadores. La estrategia que se sigue para el desarrollo de experimentos en la siguiente:

Para la clasificación se utilizan tres configuraciones, tomando en cuenta que la cantidad de registros en cada uno de los Dataset es de 22245, se implementa la clasificación utilizando el algoritmo Dl4jMlpClassifier de la siguiente manera:

- La primera configuración utilizada es validación cruzada (CROSS VALIDATION) para el que se utilizó el valor por defecto que marca WEKA en 10.
- Para la segunda fase de experimentos, se utilizó la estrategia de asignar $2/3$ de los registros para la fase de entrenamiento (training) y el otro $1/3$ para la etapa de prueba (test), por lo que un total de 14830 registros son destinados a training y 7415 registros para test.
- La última configuración utilizada en WEKA fué la opción de “Percentage Split”, la cual permite definir el porcentaje del almacén de datos que se utiliza para la fase de entrenamiento. Para los experimentos se definió en este apartado el porcentaje 85%(training) y 15%(test). Haciendo en calculo con una muestra representativa del 95% y un margen de error de 5%, estadísticamente da como cantidad el valor de 378 registros para una población que consta de 22245 registros que es lo que contiene el Dataset, calculando el porcentaje que representaría esos 378 registros da como resultado 1.7%, por lo que el porcentaje que debe ir en la sección percentage split de WEKA es 98.3%, con estos valores queda definido un training de 21867 y un set de 378 registros con un nivel de confianza de 95% y 5%.

Para cada sección se obtienen los clasificados correctamente e incorrectamente con sus porcentajes respectivos aplicados al Dataset con 72 atributos extraídos de la base de datos original (ver figura 3.1) como propuesta del experto de la empresa y al Dataset con 31 atributos resultado de la aplicación de los algoritmos de selección de atributos. A continuación, se muestran los resultados obtenidos de la aplicación del algoritmo Dl4MlpClassifier de manera individual y en combinación con los algoritmos meta-clasificadores en cada uno de los Dataset.

4.2 Dataset 31 Atributos

4.2.1 Algoritmo Dl4jMlpClassifier individual

Aplicando el algoritmo Dl4jMlpClassifier al Dataset de 31 Atributos (Anexo 11), se obtienen los porcentajes mostrados en la Tabla 4.1.

Tabla 4.1 *Resultados aplicando Dl4jMlpClassifier al Dataset 31 atributos*.

ESTRATEGIA	CORRECTAMENTE CLASIFICADO	INCORRECTAMENTE CLASIFICADOS
Cross validation	64.2616%	35.7384%
Use training set	64.4957%	35.5043%
Percentage split	65.6085%	34.3915%

La figura 4.1, muestra visualmente los mejores resultados obtenidos de aplicar el algoritmo Dl4jMlpClassifier en el Dataset con 31 atributos (ver Anexo 12).

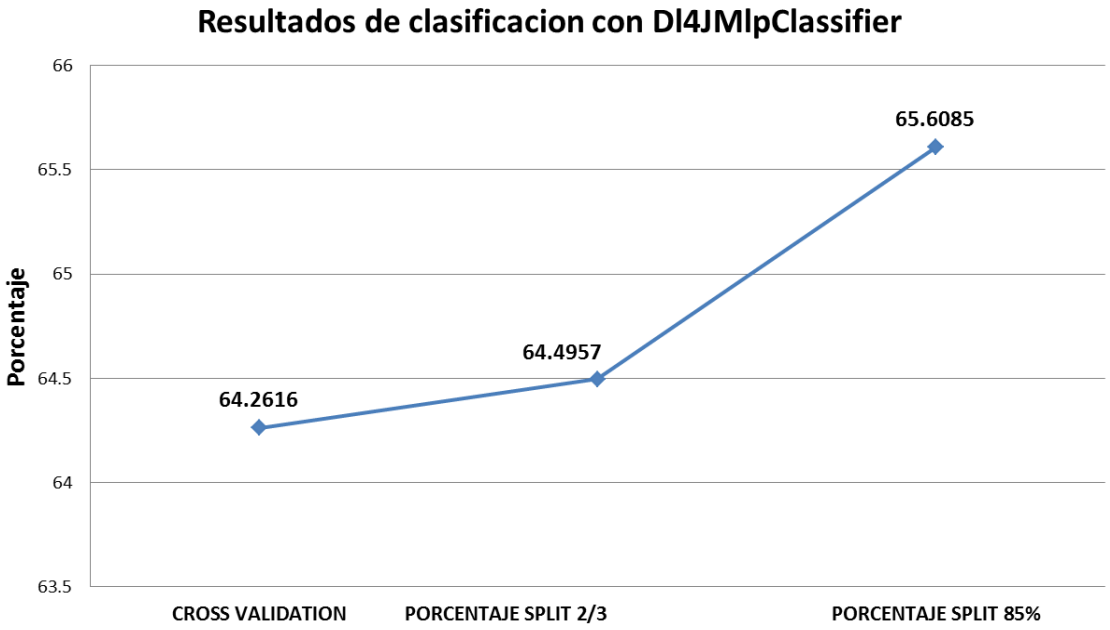


Figura 4.1 Gráfica aplicando Dl4jMlpClassifier al Dataset 31 atributos.

Como se observa en la tabla 4,1, la configuración “Percentage Split” representa el mejor resultado con un 65.6085% de clasificados correctamente. Se muestran a continuación, los resultados obtenidos en combinación con los meta-clasificadores con las configuraciones antes mencionadas.

4.2.2 Resultados utilizando configuración combinación Deep Learning y Meta-clasificadores

En los tres grupos de configuración del conjunto de datos, la combinación que con la que se obtuvieron los mejores resultados fue la combinación del algoritmo Dl4MlpClassifier y el metaclasificador FilteredClassifier (ver Anexo 12). En la Tabla 4.2, se muestran los porcentajes mejor clasificados para cada grupo de configuración.

Tabla 4.2 *Resultados aplicando Dl4jMlpClassifier combinados con algoritmos META al Dataset 31 atributos.*

Combinación algoritmo Dl4jmlpClassifier y meta-clasificador FilteredClassifier			
	Cross Validation	**Percentage Split 2/3**	**Percentage Split 85%**
Clasificados Correctamente	77.42%	74.96%	81.75%
Incorrectamente Clasificados	22.58%	25.04%	18.25%

Gráficamente, los mejores resultados obtenidos de la combinación Dl4MlpClassifier y el metaclasificador FilteredClassifier, se muestran en la figura 4.2 (ver Anexo 12).

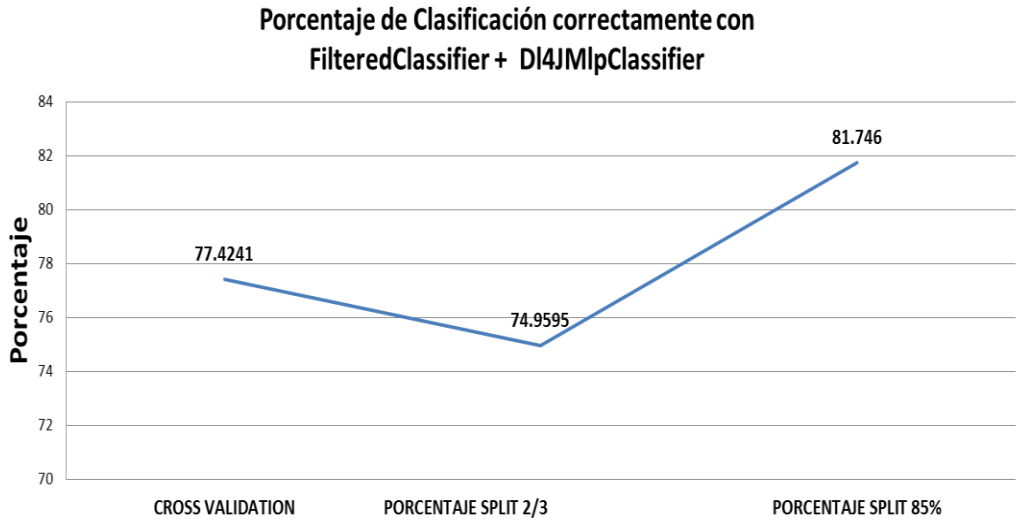


Figura 4.2 Gráfica aplicando Dl4jMlpClassifier y FilteredClassifier al Dataset 31 atributos.

Se observa en la tabla 4.2, que la configuración “Percentage Split 85%” representa el mayor porcentaje obtenido con un 81.746% de clasificados correctamente.

4.3 Dataset 72 Atributos

4.3.1 Algoritmo Dl4jMlpClassifier individual

Aplicando el algoritmo Dl4jMlpClassifier al Dataset de 72 Atributos (Anexo 13), se obtienen los porcentajes mostrados en la Tabla 4.3.

Tabla 4.3 *Resultados aplicando Dl4jMlpClassifier al Dataset 72 atributos.*

ESTRATEGIA	CORRECTAMENTE CLASIFICADO	INCORRECTAMENTE CLASIFICADOS
Cross validation	67.76%	32.24%
Use training set	67.28%	32.72%
Percentage split	68.25%	31.75%

La Figura 4.3, muestra visualmente los mejores resultados obtenidos de aplicar el algoritmo Dl4jMlpClassifier en el Dataset con 72 atributos (ver Anexo 14).

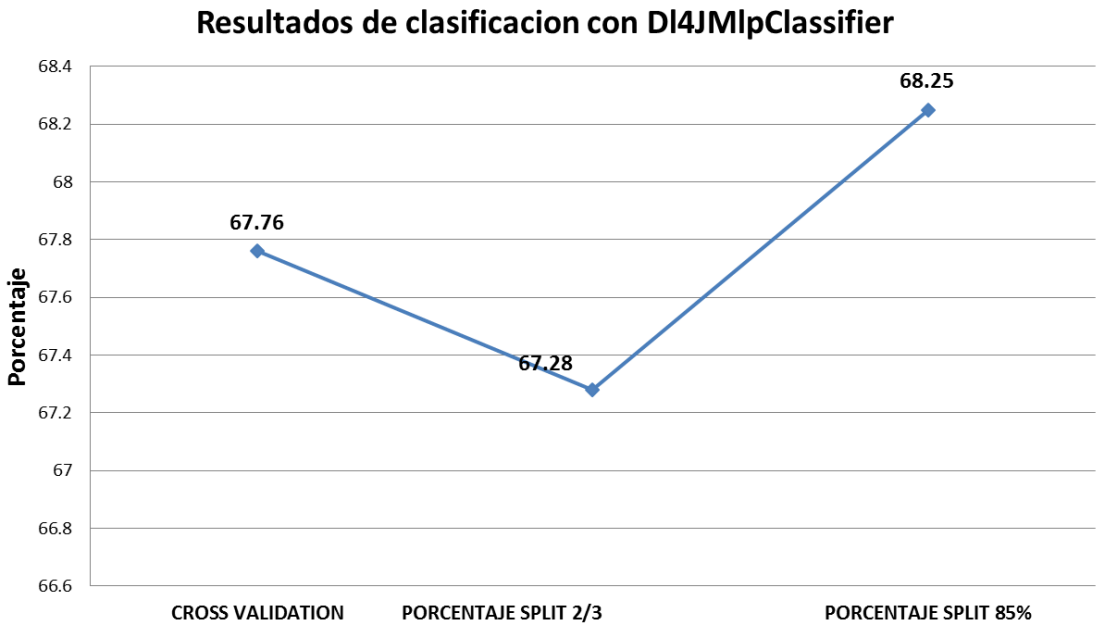


Figura 4.3 Gráfica aplicando Dl4jMlpClassifier al Dataset 72 atributos.

Como se observa en la Tabla 4.3, la configuración “Percentage Split” representa el mayor con un 68.25% de clasificados correctamente. Los resultados obtenidos del algoritmo Deep Learning en combinación con los meta-clasificadores con las configuraciones antes mencionadas, se muestran a continuación.

4.3.2 Resultados utilizando configuración combinación Deep Learning y Meta-clasificadores

En los tres grupos de configuración del conjunto de datos, la combinación con la que se obtuvieron los mejores resultados al igual que en el dataset de 31 atributos, fué la combinación del al algoritmo Dl4MlpClassifier y el metaclasificador FilteredClassifier (ver Anexo 14). En la Tabla 4.4, se muestran los porcentajes de registros mejor clasificados para cada grupo de configuración.

Tabla 4.4 *Resultados aplicando Dl4jMlpClassifier combinados con algoritmos META al Dataset 72 atributos.*

Combinación algoritmo Dl4jmlpClassifier y meta-clasificador FilteredClassifier			
	Cross Validation	**Percentage Split 2/3**	**Percentage Split 85%**
Clasificados Correctamente	79.46%	79.35%	81.75%
Incorrectamente Clasificados	20.54%	20.65%	18.25%

Gráficamente, los mejores resultados obtenidos de la combinación Dl4MlpClassifier y el metaclasificador FilteredClassifier, se muestran en la Figura 4.4 (ver Anexo 14).

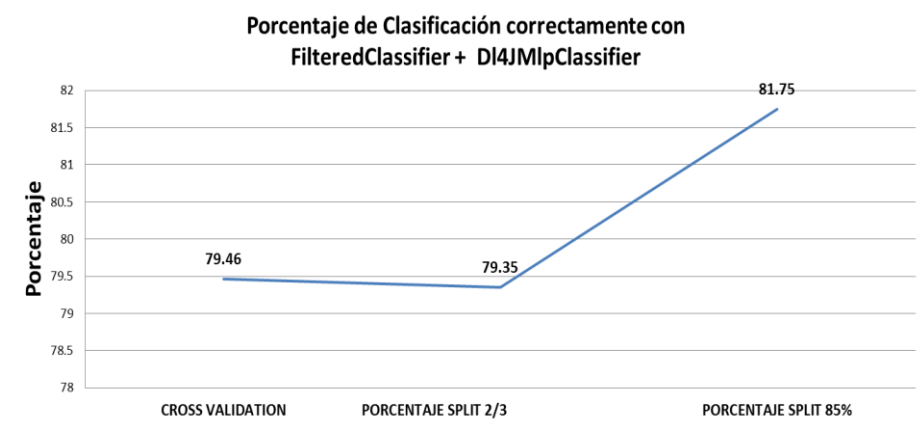


Figura 4.4 Gráfica aplicando Dl4jMlpClassifier y FilteredClassifier al Dataset 72 atributos.

Como se observa en la Tabla 4.4, la configuración “Percentage Split 85%” representa el mayor con un 81.75% de clasificados correctamente.

4.4 Matriz de Confusión y métricas de rendimiento

En base a los resultados observados de los experimentos, tanto para el Dataset de 31 atributos(Anexo 15) como para el Dataset de 72 atributos(Anexo 16), el clasificador META FilteredClassifier presentó los mejores resultados en combinación con el algoritmo Dl4JMlpClassifier registrando un porcentaje idéntico de 81.75% para ambos Dataset, utilizando la configuración “Percentage Split 85%”, esto representa un rendimiento de 13.5% mayor que la aplicación del algoritmo Dl4jMlpClassifier de manera individual, como se muestra en la Tabla 4.5.

Tabla 4.5 *Conclusión de Clasificación para Dataset 31 y 72 Atributos.*

Dl4JMlpClassifier combinado con FilteredClassifier		
Categoría	Dataset 31 Atributos	Dataset 72 Atributos
Correctamente Clasificados	81.75%	81.75%
Incorrectamente Clasificados	18.25%	18.25%

Con estos resultados y tomando en cuenta que generan un porcentaje de clasificación idéntico, se muestra en la Tabla 4.6 la matriz de confusión del Dataset con 72 atributos (ver Anexo 16).

Tabla 4.6 *Matriz de Confusión en WEKA para el Dataset con 72 Atributos.*

a	b	c	Classified as
187	24	3	a = MOROSO
23	101	9	b = NORMAL
3	7	21	c = ANTICIPADO

Se puede observar en la Tabla 4.6, que para la clasificación MOROSO, se clasificaron correctamente 187 registros, confundiendo con la clase NORMAL 24 registros y con la clase ANTICIPADO 3 registros; para la clase NORMAL, se identificaron correctamente 101 registros confundiendo 23 registros con la clase MOROSO y 9 con la clase ANTICIPADO; por último se obtuvieron 21 registros clasificados correctamente en la clase ANTICIPADO con 3 registros para la clase MOROSO y 7 para la clase NORMAL que no se determinaron correctamente, para un total de 378 registros que se utilizaron como test en esta configuración.

La matriz de confusión obtenida con WEKA, presentada en la Tabla 4.6, muestra en la diagonal principal el número de instancias correctamente clasificadas (187+101+21=309 instancias), mientras que los elementos que están por arriba y por debajo de la diagonal principal, representan el total de instancias incorrectamente clasificadas (24+3+23+9+3+7=69 instancias).

Las métricas obtenidas con esta clasificación, se muestra en la Tabla 4.7, estas métricas permitieron validar los resultados obtenidos en la clasificación (Anexo 17).

Tabla 4.7 *Valores de las métricas de Clasificación.*

CLASE	TASA TP	TASA FP	PRECISION	CURVA ROC
MOROSO	0.874	0.159	0.878	0.919
NORMAL	0.759	0.127	0.765	0.902
ANTICIPADO	0.677	0.035	0.636	0.939
Total de Instancias: 378				
Instancias correctamente clasificadas: 309 (81.74%)				
Instancias incorrectamente clasificadas: 69 (18.254%)				

La Tabla 4.7. muestra que del total de instancias utilizadas para el test (378 instancias), de este conjunto de registros el 81.74% (309 instancias) fueron correctamente clasificadas, mientras que el 18.254% (69 instancias) fue clasificada de manera incorrecta.

Para el análisis de las métricas de evaluación se utilizó sensibilidad, especificidad, precisión y curva ROC, los cuales generan los porcentajes siguientes: para la evaluación de la sensibilidad y especificidad, se toman en cuenta los valores obtenidos de tasa de verdaderos positivos (TP) y la tasa de falsos positivos (FP) respectivamente. La Figura 4.5. muestra los valores obtenidos para estos indicadores.

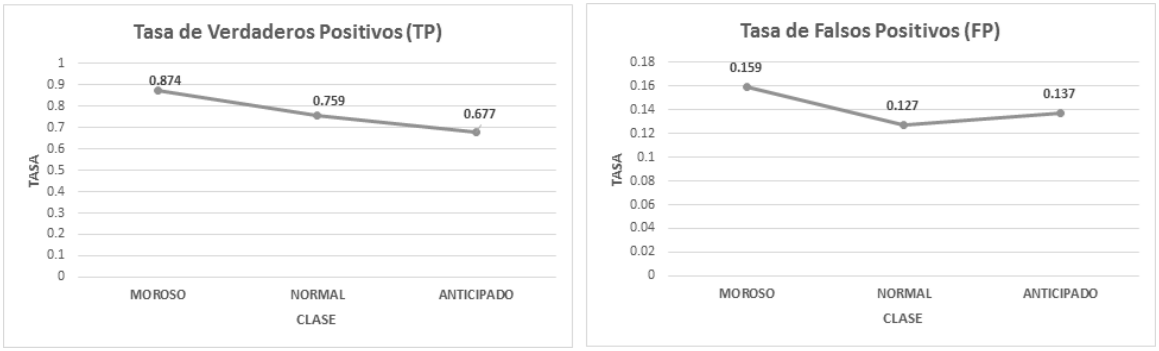


Figura 4.5 Indicadores de Verdaderos Positivos y Falsos Positivos.

Los valores mostrados en la Figura 4.5, son necesarios para definir las métricas de Precisión, Especificidad y Sensibilidad. Se debe recordar que la tasa de "**verdaderos positivos"** (TP) representan puntos ubicados por el algoritmo de clasificación en la misma clase definida en el Dataset, mientras que los **"falsos positivos" (FP)**, indican puntos ubicados por el algoritmo en una clase distinta a la que es definida en el Dataset. Con los resultados obtenidos y mostrados en la Figura 4.5, se obtiene los siguientes resultados para la métrica de Precisión mostrados en la Figura 4.6, tomando en cuenta que la precisión mide el porcentaje de muestras correctamente clasificadas.

La Figura 4.6, muestra que se obtuvo una Precisión mayor para la clase MOROSO, obteniendo un 11.3% de mayor precisión con respecto a la clase NORMAL y un 6% mayor que la clase ANTICIPADO.

Para la métrica de Especificidad se utiliza además del indicador FP el indicador de “verdaderos negativo” (VN), este indicador ubica a los puntos que se encuentran fuera del clúster y que efectivamente no correspondían a la clase especificada en el Dataset; por lo que, la métrica de Especificidad mide la proporción de muestras negativas correctamente clasificadas, es decir, la detección de puntos que no fueron puestos en una clase y que verdaderamente no corresponden a esa clase.

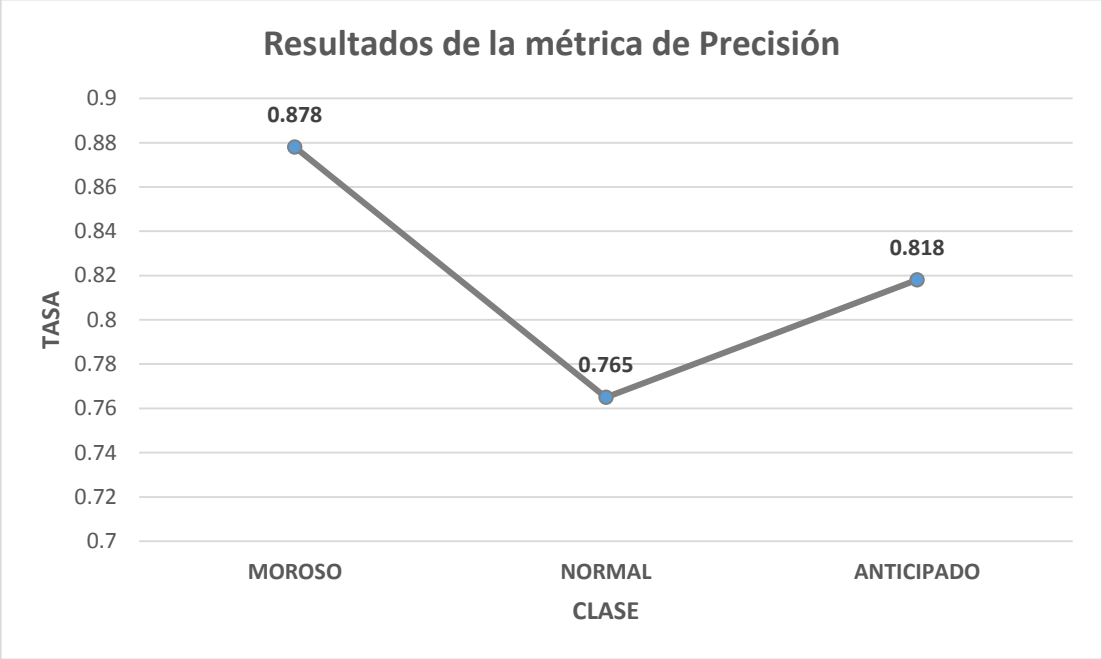


Figura 4.6 Valores obtenidos para la métrica de Precisión.

Para la métrica de Sensibilidad, se utiliza el indicador de VP en relación con el indicador de **“falsos negativos”** (FN), este indicador FN, ubica a los puntos que fueron detectados en una clasificación diferentes a la que indicaba su etiqueta en el Dataset.

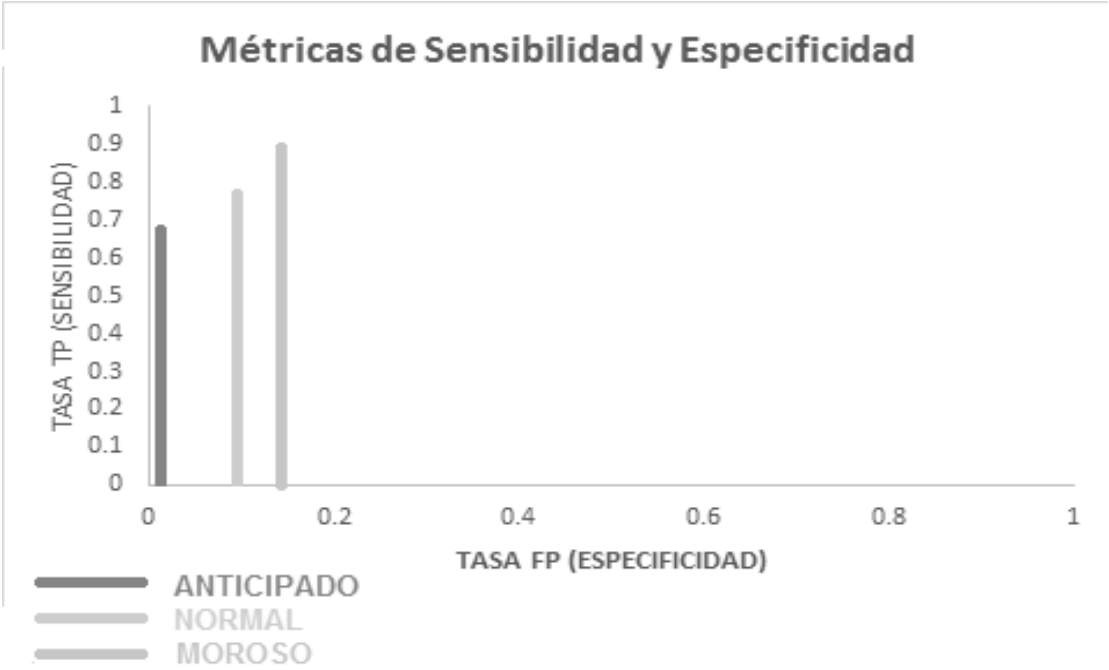


Figura 4.7 Métrica de Sensibilidad y Especificidad.

Con esos indicadores VP, FP, VN y FN, se obtiene la gráfica mostrada en la Figura 4.7, la cual muestra la relación existente con las métricas de Sensibilidad y Especificidad, estos resultados se pueden validar, visualizando las curvas ROC, estas son gráficos que se observan todos los pares sensibilidad/especificidad resultante de la variación continua de los puntos de corte en todo el rango de resultados observados Sackett, DL. (1989). En el eje y de coordenadas se sitúa la sensibilidad o fracción de verdaderos positivos, en el eje x se sitúa la fracción de falsos positivos o 1-especificidad, definida como FP/VN + FP y calculada en el subgrupo no afectado. Algunos autores sitúan en el eje x la especificidad, pero es lo menos frecuente Gerhardt W & Keller H. (1986). A continuación, se muestran las curvas ROC generadas en WEKA para cada una de las clases: MOROSO (Figura 4.8), NORMAL (Figura 4.9) y ANTICIPADO (Figura 4.10).

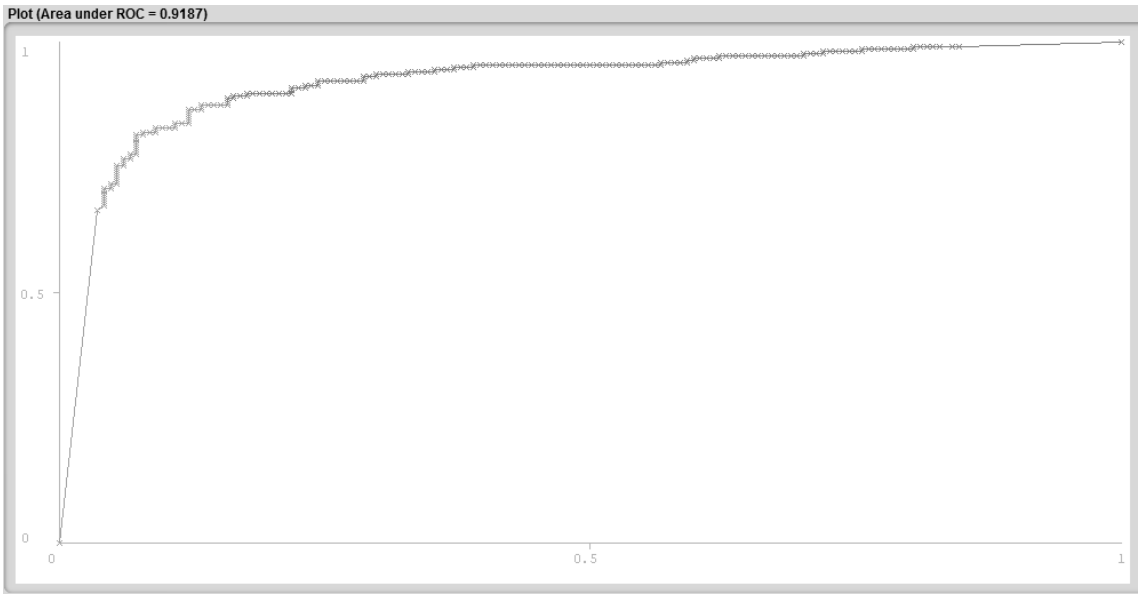


Figura 4.8 Curva ROC en WEKA para la clase MOROSO.

De acuerdo a la Figura 4.8 se aprecia que la sensibilidad de los resultados para la clase moroso es buena, tenemos que el área de la curva queda arriba de la diagonal imaginaria de 45º o del valor 0.5.

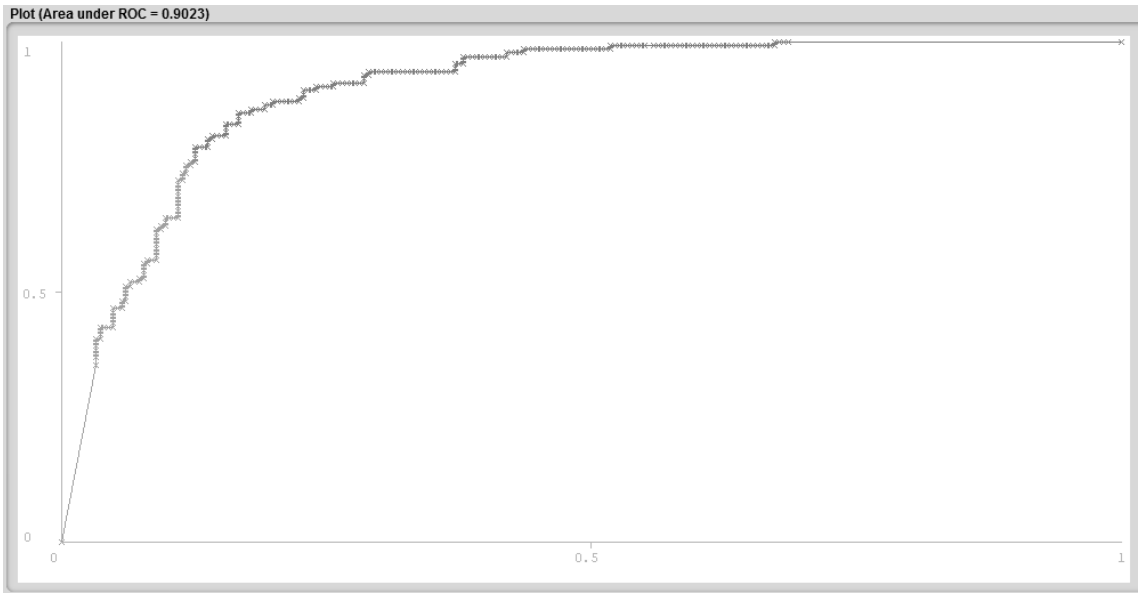


Figura 4.9 Curva ROC en WEKA para la clase NORMAL.

En la Figura 4.9 se muestra que la sensibilidad de los resultados para la clase normal es buena, tenemos que el área de la curva queda arriba de la diagonal imaginaria de 45º o del valor 0.5, pero comparada con la curva ROC de la clase moroso se encuentra con resultados menores.

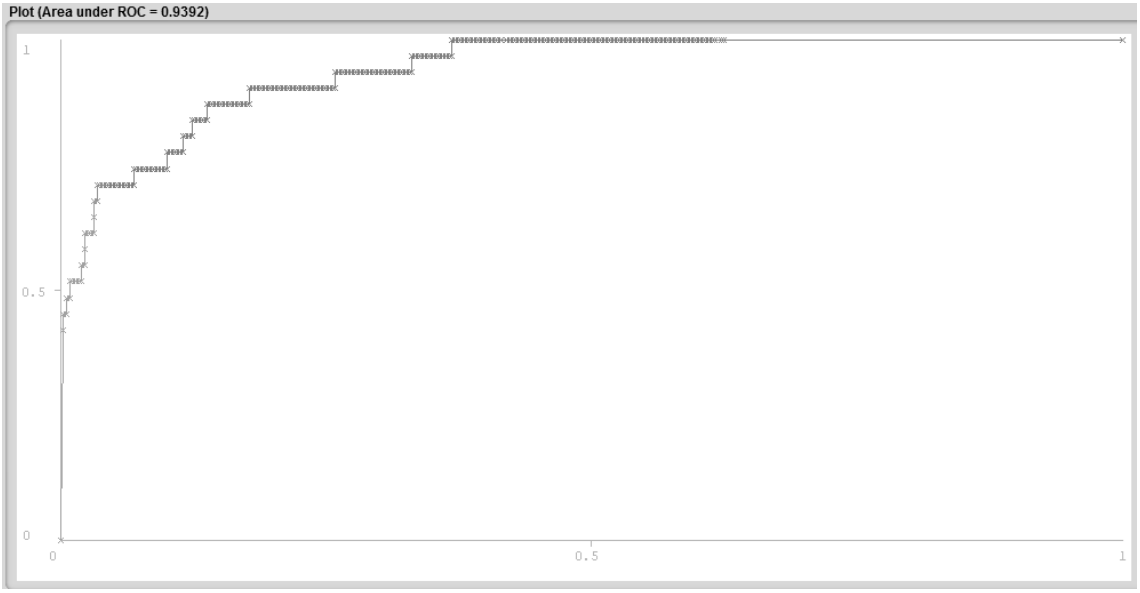


Figura 4.10 Curva ROC en WEKA para la clase ANTICIPADO.

La Figura 4.10 se aprecia que la sensibilidad de los resultados para la clase anticipado es buena, pero comparada con las dos clases anteriores se tiene que la curva ROC muestra resultados de clasificación menor. Por ello, no es confiable la clasificación.

4.5 Comparación de resultados con el estado del arte

Para la comparación de resultados, se debe recordar que se comparan los resultados obtenidos en la presente investigación con el trabajo realizado por Toledo, S. (2018), donde se aplican algoritmos Meta-clasificadores en combinación con algoritmos de selección de atributos para clasificar Dataset con información histórica de clientes. Tomando como base los resultados obtenidos en Toledo, S. (2018), donde se demuestra que se obtuvo un porcentaje de clasificación mayor con el Dataset de 31 atributos (utilizando el análisis exploratorio del almacén de datos), se comparan los resultados obtenidos en el presente trabajo en la Tabla 4.8.

Tabla 4.8 *Comparación de resultados reportados en Toledo, S. (2018).*

DATASET	Meta-clasificadores Reportado en Toledo, S. (2018)	Algoritmos Deep Learning (Trabajo presente)	DIFERENCIA
31 Atributos	98.93%	81.75%	- 17.18
72 Atributos	98.84%	81.75%	-17.09

Como se observa en la Tabla 4.8, el resultado obtenido al realizar la clasificación de datos históricos de clientes utilizando algoritmos meta-clasificadores y algoritmos de selección de atributos (98.93%), presenta una diferencia negativa de 17.18% con relación a la aplicación de algoritmos meta-clasificadores en combinación con algoritmos de redes neuronales (81.75% en ambos Datasets), por lo que no se mejoraron los porcentajes de clasificación reportados en Toledo, S. (2018).

Por lo que para la hipótesis propuesta en el presente trabajo se rechaza debido a que la combinación de un clasificador ensamblado con un algoritmo de red neuronal artificial profunda no genera mejores porcentajes de clasificación de los tipos de clientes registrados en un Dataset histórico de pagos.

Con esto, terminamos el apartado de análisis de resultados y se realiza en la siguiente sección las conclusiones y trabajos futuros.

CAPÍTULO 5
CONCLUSIÓN Y TRABAJOS FUTUROS

En el presente trabajo se obtienen dos conclusiones generales en relación al pre-procesamiento y la clasificación:

i. Se puede obtener otra u otras variantes, a partir de los dos datasets utilizados para las pruebas, es decir, aplicar un refinamiento de datos teniendo como objetivo una discriminación diferente a la de clases moroso, normal o anticipado. Por otro lado, se puede explorar el uso de otros algoritmos de selección de atributos recientes en la literatura, ya que siendo el área de sistemas computacionales un ámbito altamente cambiante debido a su naturaleza tecnológica, frecuentemente se extienden o proponen nuevos algoritmos para mejorar (ampliar o disminuir las dimensiones) de las bases datos.

ii. Se ha mostrado que utilizando algoritmos ensamblados combinados con un algoritmo de Deep Learning, se obtienen porcentajes aceptables respecto a los reportados en la literatura especializada. Dichos porcentajes pueden ser mejorados, como en el punto anterior, con base en nuevos métodos de clasificación que sean afines para su implementación en base de datos históricas de clientes.

Por lo tanto, es posible una siguiente fase del presente estudio, considerando los siguientes trabajos futuros:

- Utilizar algoritmos recientes en el Estado del arte, de selección de atributos.
- Implementar el método de *Selección de algoritmos e hiperparametrización*, para aumentar, si es posible, el porcentaje de correcta clasificación de instancias de los dos Datasets usados.
- Encontrar y explorar con un nuevo Dataset con especificaciones semejantes a la del presente trabajo, para una comparación homogénea y más justa.
- Desarrollar un sistema experto que permita a la empresa detectar con alto grado de confianza, si un cliente potencial será bueno o no, en términos de sus pagos por servicio.
- Se recomienda, realizar trabajos que permitan medir con diferentes métricas, la eficiencia y costo de procesamiento de los diferentes Dataset generados.

REFERENCIAS

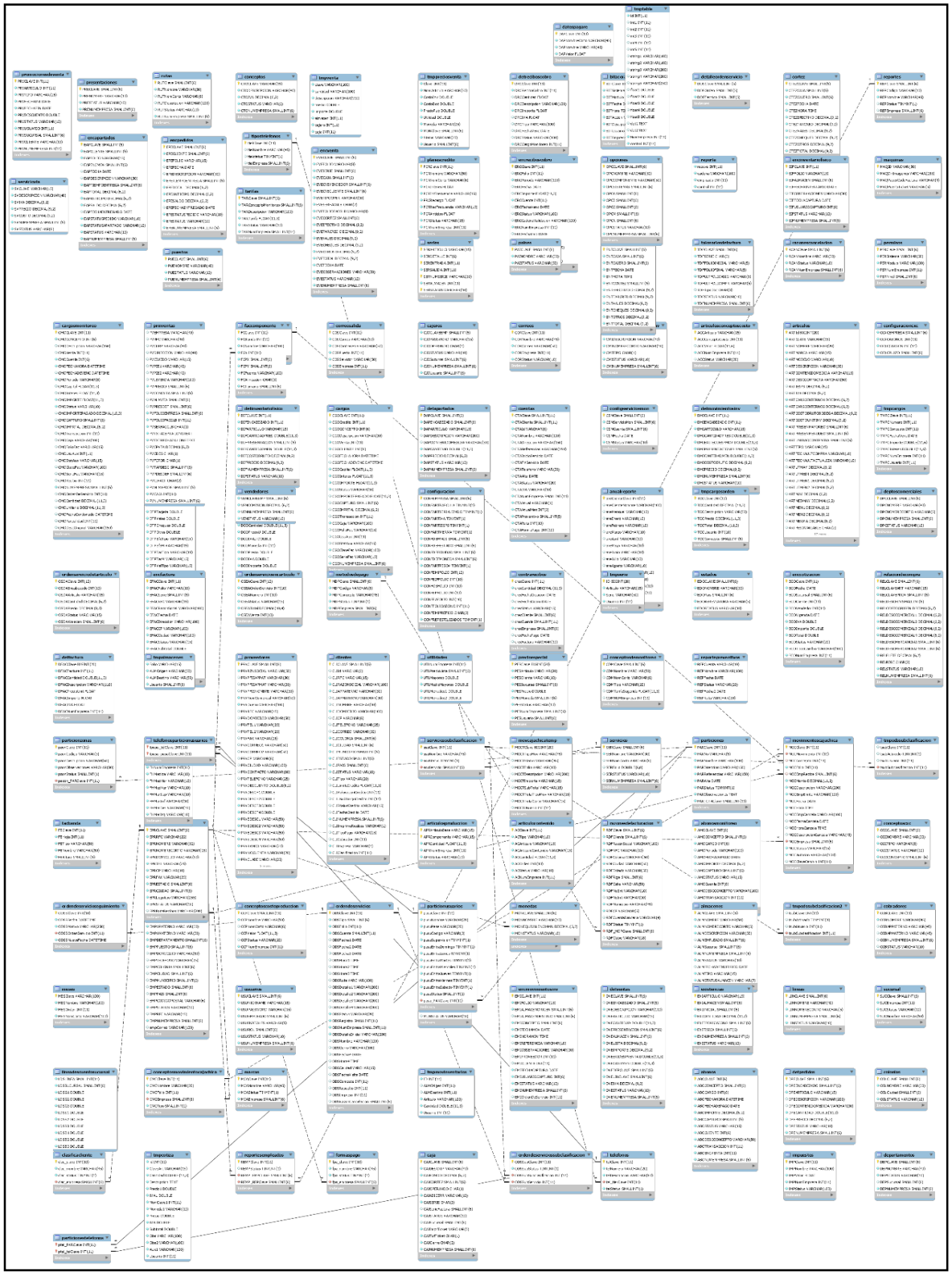
1. Abadi, Martın, et al. (2016). "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." arXiv preprint arXiv:1603.04467.
2. Aboobyda J. & Taring M. (2016). Developing prediction model of loan risk in Banks using data mining (Vol. 3). Khartoum, Sudan.
3. Aguilar, J. & Diaz, N. (2005). Selección de atributos relevantes basada en bootstrapping. Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.21-30 ISBN: 84-9732-449-8
4. Arel, Itamar, Derek C. Rose, and Thomas P. Karnowski (2010). "Deep machine learning-a new frontier in artificial intelligence research [research frontier]."IEEE Computational Intelligence Magazine Vol. 5. No. 4, 13-18.
5. Cárdenas Cardona, Alejandro (2012). "Inteligencia artificial, métodos bio-inspirados: un enfoque funcional para las ciencias de la computación.".
6. Cristóbal Romero, José Raúl Romero, and Sebastián Ventura (2014). A survey on pre-processing educational data. In Alejandro Peña Ayala, editor, Educational Data Mining, volume 524 of Studies in Computational Intelligence, pages 29–64. Springer International Publishing, ISBN 978-3-319-02737-1. d o i : 10.1007/978-3-319-0273
7. Cruz, R., Lavernia, A., Franco, A., Simón, I. (2017). Estudio del comportamiento de algoritmos de clasificación según la naturaleza de los datos. Revista de Tecnología Informática. Vol. 1 No. 2, 9-18.
8. Duval, M., Vega, S., Ruiz S. (2012). Combinación de Clasificadores Supervisados: estado del arte. Centro de aplicaciones de tecnología avanzada. Reporte técnico. RNPS No. 2142, ISSN 2072-6287
9. Gerhardt W & Keller H. (1986). Evaluation of test data from clinical studies. II. Critical review of concepts of efficiency, Receiver Operated Characteristic (ROC) and likelihood ratios. Scand J Clin Lab lnvest; 46 Supl 181: 47- 74.
10. Jiawei Han and Micheline (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann, ISBN 1-55860-489-8.
11. Kuncheva, L. (2004). Combining pattern Classifier. Methods and Algorithms. United State: Editorial Wiley-InterScience.

12. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning." Nature Vol. 521 No.7553, 436-444.
13. Liu H, Setiono R. Chi2 (2005). Feature Selection and Discretization of Numeric Attributes. Department of Information System and Computer Science.
14. Martín, R. (2017). Predicción y selección de características, mediante análisis local de la fiabilidad, para el mercado de valores y su extensión a problemas de clasificación (Tesis Doctoral). Departamento de Informática, Universidad Carlos III de Madrid. España.
15. McCulloch, Warren S., and Walter Pitts (1943). "A logical calculus of the ideas immanent in nervous activity." The bulletin of mathematical biophysics vol.5. No. 4, 115-133.
16. Ortiz, A., (2014). Algoritmo multiclasificador con aprendizaje incremental que manipula cambios de conceptos. (Tesis Doctoral). Universidad de Granada. España.
17. Rivero, J., et al. (2016). Proposal of data processing platform for direct marketing data. Revista Universidad y Sociedad [seriada en línea], 8 (2). pp. 65-71
18. Ruiz, R. et al. (2005). Evaluación de Rankings de atributos para clasificación. Departamento de Lenguajes y sistemas informáticos. Universidad de Sevilla, España.
19. Ruiz, R. et al. (2005). Búsqueda secuencial de subconjuntos de atributos sobre un ranking. Departamento de Lenguajes y sistemas informáticos. Universidad de Sevilla, España. Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.251-260 ISBN: 84-9732-449-8.
20. Sackett, DL. et al. (1989). Epidemiología clínica. Una ciencia básica para la medicina clínica. Madrid: Díaz Santos S.A.
21. Sánchez, E. (2012). Detección de anormalidades en mastografías digitales usando el proceso KDD (Tesis Doctoral). División académica de informática y sistemas, Universidad Juarez Autónoma de Tabasco. México.

22. Toledo, S. (2018). Preprocesamiento y clasificación de datos históricos de cliente. Caso SISCOM S.A.
23. Yu, L., Liu, H. (2003). Feature Selection for High-Dimensional Data). Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC

ANEXOS

Anexo 1. Diagrama Entidad-Relación de la base de datos Issai de la empresa Siscom S.A.



Anexo 2. Propuesta 1, integrada con los atributos recomendados por el experto (25 atributos y 19443 registros).

CMO=cargosmonitoreo AMO=abonosmonitore CTA=cuentas CLI=clientes				
No.	**ATRIBUTOS**			
1	CLIClave	CTAId	CMOCONCEPTO	CMOStatus
2	CLIRAZONSOCIAL	CTANombre	CMODescripcion	AMOClave
3	CLIStatus	CTAStatus	CMOCuenta	AMOConcepto
4	CLIClasificacion	CTARuta	CMOFECHAHORA	AMOCargo
5	CTAClave	CTAFormaPago	CMOFECHADEVENC	AMOFechadePago
6	CTACliente	CMOClave	CMOIMPORTE	AMOImporte
7				AMODescConcepto

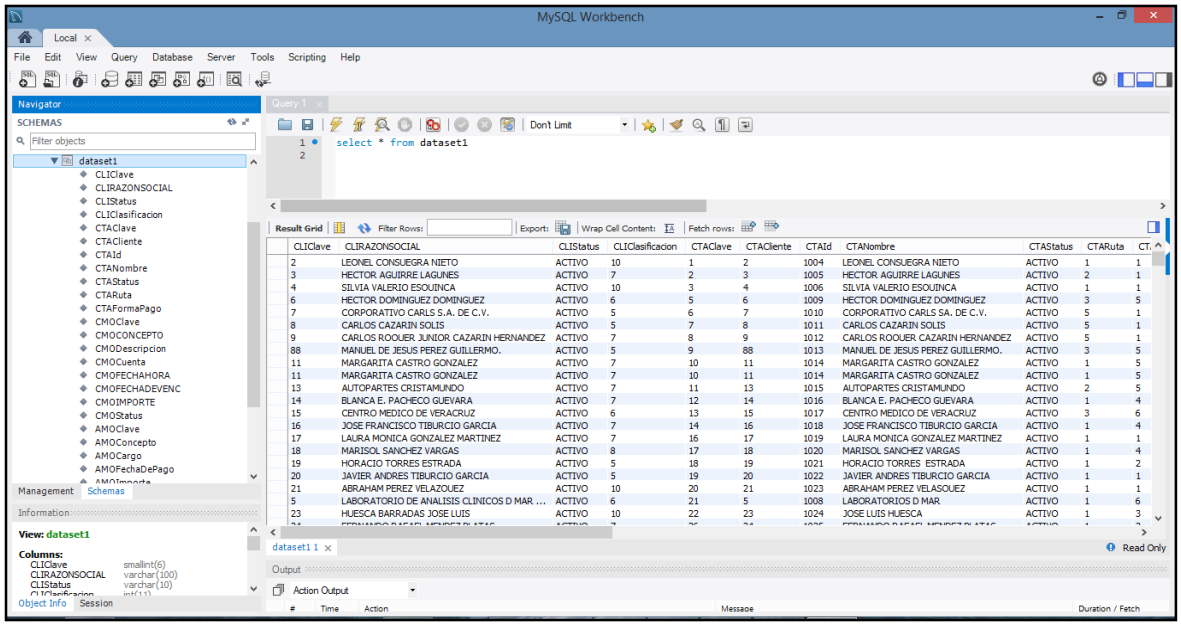
Anexo 3. Propuesta 2, integrada con los atributos que resultan del análisis de información (88 ATRIBUTOS Y 22245 REGISTROS).

CMO=cargosmonitoreo AMO=abonosmonitore CTA=cuentas CLI=clientes				
No	**ATRIBUTOS**			
1	CMOCLAVE	CMOCONCEPTO	CMODescripcion	CMOVenta
2	CMOFECHAHORA	CMOFECHADEVENC	CMOPeriodo	CMOCapital
3	CMOInteres	CMOIMPORTE	CMOStatus	CMOIMPORTEPAGADO
4	CMOCAPTURO	CMOIMPRTAL	CMOTransaccion	CMOCaja
5	CMOfolioFac	CMOUsuAux	CMOStatAux	CMOClaveFac
6	CMOSerieFac	CMORecibo	CMONUMEMPRESA	CMOOrdenDeServicio
7	CMOCantidad	CMOUnitario	CMOFechaCondonado	CMOParcialidad
8	CMODescOriginal	AMOClave`,	AMOCONCEPTO	AMOCARGO
9	AMOPeriodo	AMOFECHADEPAGO	AMOIMPORTE	AMOCAPTURO
10	AMOSTATUS	AMOCuenta	AMODESCCONCEPTO	AMOTRANSACCION
11	AMONUMEMPRESA	CTAClave	CTACliente	CTAId
12	CTANombre	CTADireccion	CTAReferencias	CTAInicioServicio
13	CTAFinServicio	CTATelefono	CTAAlta	CTAStatus
14	CTAIva	CTANumEmpresa	CTAAnual	CTAAnualMes
15	CTAPreventivo	CTARuta	CTAFormaPago	CLICLAVE
16	CLIId	CLIRFC	CLIRAZONSOCIAL	CLIAPPATERNO
17	CLIAPMATERNO	CLINOMBRE	CLIDOMICILIO	CLICP
18	CLITELEFONO	CLICORREO	CLICOLONIA	CLICIUDAD
19	CLIMUNICIPIO	CLIESTADO	CLIPAIS	CLISTATUS
20	CLITipo	CLILimiteCredito	CLIValoracionCredito	CLIUsuOtorgoCredito
21	CLIStatusCredito	CLIFechaCredito	CLINUMEMPRESA	CLIImprimeRecibos
22	CLITipoPago	CLICobrador	CLIImpreso	CLIClasificacion

Anexo 4. Propuesta 1. Relación Abono-Cargo (abonomonitoreo - cargosmonitoreo).

```
CREATE
ALGORITHM = UNDEFINED
SQL SECURITY DEFINER
VIEW siscom.dataset1 AS
SELECT
`cl`.`CLICLAVE` AS `CLIClave`, `cl`.`CLIRAZONSOCIAL` AS `CLIRAZONSOCIAL`, `cl`.`CLISTATUS` AS `CLIStatus`,
`cl`.`CLIClasificacion` AS `CLIClasificacion`,
`ct`.`CTAClave` AS `CTAClave`, `ct`.`CTACliente` AS `CTACliente`, `ct`.`CTAId` AS `CTAId`, `ct`.`CTANombre` AS `CTANombre`,
`ct`.`CTAStatus` AS `CTAStatus`, `ct`.`CTARuta` AS `CTARuta`, `ct`.`CTAFormaPago` AS `CTAFormaPago`,
`cm`.`CMOCLAVE` AS `CMOClave`, `cm`.`CMOCONCEPTO` AS `CMOCONCEPTO`, `cm`.`CMODescripcion` AS `CMODescripcion`,
`cm`.`CMOCuenta` AS `CMOCuenta`, `cm`.`CMOFECHAHORA` AS `CMOFECHAHORA`, `cm`.`CMOFECHADEVENC` AS
`CMOFECHADEVENC`, `cm`.`CMOIMPORTE` AS `CMOIMPORTE`, `cm`.`CMOStatus` AS `CMOStatus`,
`am`.`AMOCLAVE` AS `AMOClave`, `am`.`AMOCONCEPTO` AS `AMOConcepto`, `am`.`AMOCARGO` AS `AMOCargo`,
`am`.`AMOFECHADEPAGO` AS `AMOFechaDePago`, `am`.`AMOIMPORTE` AS `AMOImporte`, `am`.`AMODESCCONCEPTO` AS
`AMODescConcepto`
FROM
(((`lssai`.`abonosmonitoreo` `am`
LEFT JOIN `lssai`.`cargosmonitoreo` `cm` ON ((`cm`.`CMOCLAVE` = `am`.`AMOCARGO`)))
JOIN `lssai`.`cuentas` `ct` ON ((`cm`.`CMOCuenta` = `ct`.`CTAClave`)))
JOIN `lssai`.`clientes` `cl` ON ((`ct`.`CTACliente` = `cl`.`CLICLAVE`)))
ORDER BY `cm`.`CMOFECHAHORA` , `cm`.`CMOCuenta`
```

Consulta SQL para integrar propuesta 1 de almacén de datos.



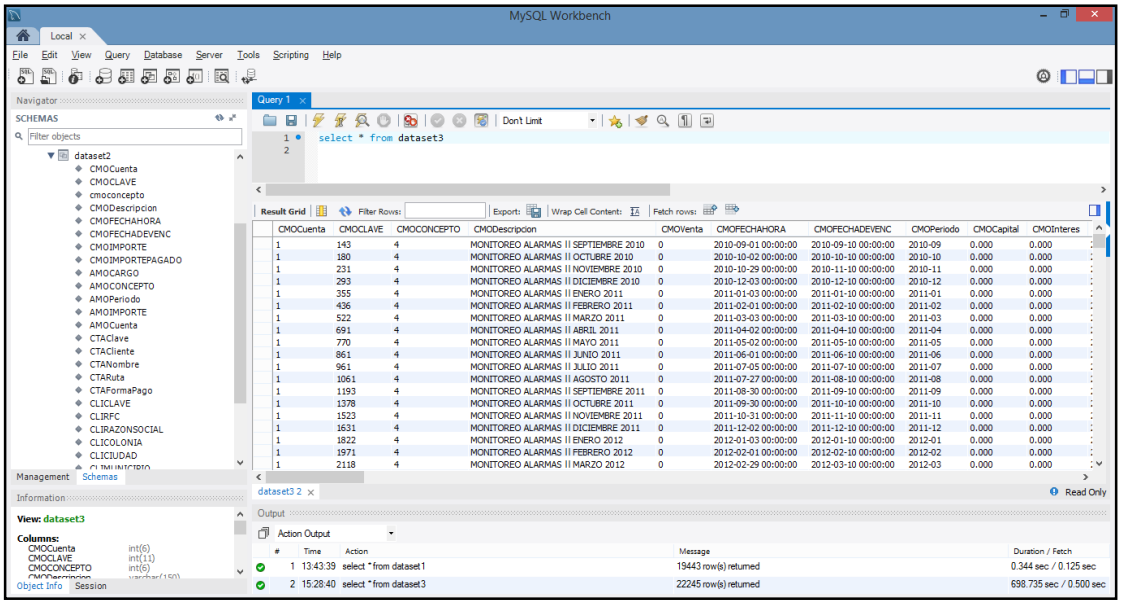
Vista de la propuesta 1 en Mysql Workbench.

Anexo 5. Propuesta 2. Relación Cargo-Abono (cargosmonitoreo - abonosmonitoreo).

```
CREATE
ALGORITHM = UNDEFINED
SQL SECURITY DEFINER
VIEW siscom.dataset3 AS

SELECT
CMOCuenta, CMOCLAVE, CMOCONCEPTO, CMODescripcion, CMOVenta, CMOFECHAHORA, CMOFECHADEVENC,
CMOPeriodo, CMOCapital, CMOInteres, CMOIMPORTE, CMOStatus, CMOIMPORTEPAGADO, CMOCAPTURO, CMOIMPRTAL,
CMOTransaccion, CMOCaja, CMOfolioFac, CMOUsuAux, CMOStatAux, CMOClaveFac, CMOSerieFac, CMORecibo,
CMONUMEMPRESA, CMOOrdenDeServicio, CMOCantidad, CMOUnitario, CMOFechaCondonado, CMOParcialidad,
CMODescOriginal,
AMOClave, AMOCONCEPTO, AMOCARGO, AMOPeriodo, AMOFECHADEPAGO, AMOIMPORTE, AMOCAPTURO, AMOSTATUS,
AMOCuenta, AMODESCCONCEPTO, AMOTRANSACCION, AMONUMEMPRESA,
CTAClave, CTACliente, CTAId, CTANombre, CTADireccion, CTAReferencias, CTAInicioServicio, CTAFinServicio, CTATelefono,
CTAAlta, CTAStatus, CTAIva, CTANumEmpresa, CTAAnual, CTAAnualMes, CTAPreventivo, CTARuta, CTAFormaPago,
CLICLAVE, CLIId, CLIRFC, CLIRAZONSOCIAL, CLIAPPATERNO, CLIAPMATERNO, CLINOMBRE, CLIDOMICILIO,
CLICP, CLITELEFONO, CLICORREO, CLICOLONIA, CLICIUDAD, CLIMUNICIPIO, CLIESTADO, CLIPAIS, CLISTATUS,
CLITipo,CLILimiteCredito, CLIValoracionCredito, CLIUsuOtorgoCredito, CLIStatusCredito, CLIFechaCredito,
CLINUMEMPRESA, CLIImprimeRecibos, CLITipoPago, CLICobrador, CLIImpreso, CLIClasificacion
from (((cargosmonitoreo left join abonosmonitoreo on (CMOCLAVE=AMOCargo))
join cuentas on (CMOCuenta=CTAClave) join clientes on (CTACliente=CLIClave)))
order by CMOCuenta
```

Consulta SQL para integrar propuesta 2 de almacén de datos.



Vista de la propuesta 2 en Mysql Workbench.

Anexo 6: Paso 1 de la etapa de pre-procesamiento. Completar o eliminar valores incompletos o faltantes en el Dataset 1.

ATRIBUTOS	ESPACIOS EN BLANCO	VALORES NULOS
CLIClave	0	0
CLIRAZONSOCIAL	0	0
CLIStatus	0	0
CLIClasificacion	0	0
CTAClave	0	0
CTACliente	0	0
CTAId	0	0
CTANombre	0	0
CTAStatus	0	0
CTARuta	0	0
CTAFormaPago	0	0
CMOClave	0	0
CMOCONCEPTO	0	0
CMODescripcion	0	0
CMOCuenta	0	0
CMOFECHAHORA	0	0
CMOFECHADEVENC	0	0
CMOIMPORTE	0	0
CMOStatus	0	0
AMOClave	0	0
AMOConcepto	0	0
AMOCargo	0	0
AMOFechaDePago	0	0
AMOImporte	0	0
AMODescConcepto	0	0
TOTAL	**0**	**0**

Anexo 7: Paso 1 de la etapa de pre-procesamiento. Completar o eliminar valores incompletos o faltantes en el Dataset 2.

DATASET 2		
ATRIBUTO	ESPACIOS EN BLANCO	VALORES NULOS
CMOCuenta	0	0
CMOCLAVE	0	0
CMOCONCEPTO	0	0
CMODescripcion	0	0
CMOVenta	0	0
CMOFECHAHORA	2	0
CMOFECHADEVENC	0	0
CMOPeriodo	1838	0
CMOCapital	0	0
CMOInteres	0	0
CMOIMPORTE	0	2
CMOStatus	0	0
CMOIMPORTEPAGADO	0	22243
CMOCAPTURO	0	0
CMOIMPRTAL	2	0
CMOTransaccion	2	0
CMOCaja	22243	0
CMOfolioFac	22245	0
CMOUsuAux	2	0
CMOStatAux	22243	2
CMOClaveFac	22243	0
CMOSerieFac	0	22243
CMORecibo	0	2
CMONUMEMPRESA	0	0
CMOOrdenDeServicio	0	22189
CMOCantidad	0	20590
CMOUnitario	0	19820
CMOFechaCondonado	0	16466
CMOParcialidad	0	0
CMODescOriginal	0	21676

DATASET 2		
ATRIBUTO	ESPACIOS EN BLANCO	VALORES NULOS
AMOClave	0	2802
AMOCONCEPTO	2	2802
AMOCARGO	0	2802
AMOPeriodo	1632	2802
AMOFECHADEPAGO	0	2802
AMOIMPORTE	0	2802
AMOCAPTURO	0	2802
AMOSTATUS	0	2802
AMOCuenta	0	2802
AMODESCCONCEPTO	0	2802
AMOTRANSACCION	0	2802
AMONUMEMPRESA	0	2802
CTAClave	0	0
CTACliente	0	0
CTAId	0	0
CTANombre	0	0
CTADireccion	0	0
CTAReferencias	9999	0
CTAInicioServicio	0	0
CTAFinServicio	0	0
CTATelefono	4518	0
CTAAlta	0	0
CTAStatus	0	0
CTAIva	2	0
CTANumEmpresa	0	0
CTAAnual	0	0
CTAAnualMes	0	0
CTAPreventivo	0	0
CTARuta	0	0
CTAFormaPago	0	0

DATASET 2		
ATRIBUTO	ESPACIOS EN BLANCO	VALORES NULOS
CLICLAVE	0	0
CLIId	22243	0
CLIRFC	1273	0
CLIRAZONSOCIAL	0	0
CLIAPPATERNO	460	0
CLIAPMATERNO	1816	0
CLINOMBRE	0	0
CLIDOMICILIO	0	0
CLICP	2281	0
CLITELEFONO	3285	0
CLICORREO	3282	0
CLICOLONIA	0	0
CLICIUDAD	0	0
CLIMUNICIPIO	2	0
CLIESTADO	0	0
CLIPAIS	0	0
CLISTATUS	0	0
CLITipo	0	0
CLILimiteCredito	0	0
CLIValoracionCredito	0	0
CLIUsuOtorgoCredito	0	0
CLIStatusCredito	22127	0
CLIFechaCredito	2	0
CLINUMEMPRESA	2	0
CLIImprimeRecibos	2	15387
CLITipoPago	173	0
CLICobrador	2	0
CLIImpreso	2	0
CLIClasificacion	2	0
TOTAL	163927	194244

Anexo 8. Detalle del paso 1 de la etapa de pre-procesamiento. Completar o eliminar valores incompletos o faltantes en el Dataset 2.

ACCIONES: 1.-Ignorar la tupla, 2.- Completar el dato faltante a mano, 3.- Usar una constante global, 4.- Usar el valor medio del atributo moda, 5.- Usar el valor más probable basado en inferencia		
ATRIBUTO	**PROCEDIMIENTO**	**DESCRIPCIÓN**
CMOVenta	2 y 3	Se completó con valores 0, solo 2 registros contienen información de la fecha de la venta.
CMOFECHAHORA	2	Tiene desfase en la información de los registros 5207 y 5211, los cuales tienen un número de cargo 12368.
CMOCapital,	3	Tienen en su totalidad el valor 0.
CMOInteres,	3	
CMOCAPTURO,	3	
CMOIMPRTAL,	3	
CMOTransaccion,	3	
CMOUsuAux	3	
CMOIMPORTEPAGADO	3	Tiene todos los datos con el valor NULL, por lo que se sustituyó con el valor de 0
CMOCAJA y	3	
CMOfoliofac,	3	
CMOStatAux,	3	
CMOClaveFac,	3	
CMOSerieFac	3	
CMOOrdenDeServicio	3	
CMOCantidad	4	Se sustituyó por el valor de 1, el cual es moda ya que se refiere a la cantidad de servicios realizados de monitoreo.
CMOUnitario	4	Se sustituyó por el valor 0, tomando como base la moda de los demás registros
CMOFechaCondonado	2	Se sustituyeron los atributos con valores NULL y lo que contenían el valor 0000-00-00 00:00:00, con el valor contenido en el campo AMOFECHADEPAGO, ya que este valor representa la fecha en la que se realizó el pago o la condonación de este
CMODescOriginal	2	Se sustituyeron los campos faltantes con el contenido del atributo CMODescripcion.
CTADireccion	3	Se sustituyeron las celdas vacías con el valor las siglas S/D (Sin Definir), ya que estos campos tenia espacios en blanco, o diferentes textos para identificar que no estaba definido
CTAReferencias	3	
CTATelefono	3	
CTAAnual	3	
CLIId	3	
CLIRFC	3	
CTAAnual	3	
CLIId	3	

CLIRFC	3	
CLIPPATERNO	3	
CLIAPMATERNO	3	
CLINOMBRE	3	
CLIDOMICILIO	3	
CLICP	3	
CLITELEFONO	3	
CLICORREO	3	
CLIPAIS	4	Se asignó el Número 2 por moda en este atributo.
CLISTATUS	2 y3	Se sustituyó con el campo CTAStatus que contiene el tipo que status que guarda el cliente, al tratarse de clientes vigentes, se modificaron estos registros con la palabra ACTIVO.
CLITipoPago	2 y3	se sustituyeron los registros vacíos por el tipo de datos OTROS tomando como base que no se especifica si es FACTURA o RECIBO

Anexo 9. Detalle del Paso 3 de la etapa de pre-procesamiento. Corregir Inconsistencias en el Dataset 1.

DATO INCONSISTENTE	SE SUSTITUYO POR:	REEEMPLAZOS	ATRIBUTO
CARLOS ISAÃ□AS HERNÃ□NDEZ LASTRA	CARLOS ISAIAS HERNANDEZ LASTRA	7	CLIRAZONSOCIAL
CARLOS SCHMIDT NUÃ‘EZ -CLAUDIA ALVARADO	CARLOS SCHMIDT NUNEZ - CLAUDIA ALVARADO	10	CLIRAZONSOCIAL
CASA DISEÃ‘OS	CASA DISENOS	32	CLIRAZONSOCIAL
CIRENIA SANTOS CAÃ‘ETE	CIRENIA SANTOS CANETE	40	CLIRAZONSOCIAL
GABRIELA PIÃ‘EIRO	GABRIELA PINEIRO	62	CLIRAZONSOCIAL
JULIETA ELENA NUÃ‘EZ SAAVEDRA	JULIETA ELENA NUNEZ SAAVEDRA	2	CLIRAZONSOCIAL
LAURA SAN JUAN BAÃ‘OS	LAURA SAN JUAN BANOS	86	CLIRAZONSOCIAL
MERCEDES GARCIA PEREZ - CARLOS ALBERTO MUÃ‘OZ GARCI	MERCEDES GARCIA PEREZ - CARLOS ALBERTO MUNOZ GARCIA	79	CLIRAZONSOCIAL
OLGA LUCY GUTIERREZ TRIVIÃ‘O	OLGA LUCY GUTIERREZ TRIVINO	80	CLIRAZONSOCIAL
OSCAR PEÃ‘A HERNANDEZ	OSCAR PENA HERNANDEZ	10	CLIRAZONSOCIAL
RAUL PEÃ‘A TOVAR	RAUL PENA TOVAR	50	CLIRAZONSOCIAL
REYNA ELIZABETH DOMÃ□NGUEZ VIVEROS	REYNA ELIZABETH DOMINGUEZ VIVEROS	74	CLIRAZONSOCIAL
RICARDO COUTIÃ‘O LARA	RICARDO COUTINO LARA	142	CLIRAZONSOCIAL
SAUL MUÃ‘OZ RODRIGUEZ	SAUL MUNOZ RODRIGUEZ	44	CLIRAZONSOCIAL
SUSANA PATRICIA AKE CHIÃ‘AS	SUSANA PATRICIA AKE CHINAS	152	CLIRAZONSOCIAL
SAMUEL YAÃ‘EZ PICAZO	SAMUEL YANEZ PICAZO	175	CLIRAZONSOCIAL
GRUAS Y MANIOBRAS DE MÃ‰XICO, S. A. DE C. V.	GRUAS Y MANIOBRAS DE MEXICO, S. A. DE C. V.	237	CLIRAZONSOCIAL
ING FRANCISCO J DIAZ MUÃ‘OZ	ING FRANCISCO J DIAZ MUNOZ	38	CLIRAZONSOCIAL

N	NETSOLUTION SHOP S. A. DE C. V.	6	CLIRAZONSOCIAL
JORGE LUIS VAZQUEZ ORDOÃ‘EZ	JORGE LUIS VAZQUEZ ORDONEZ	19	CTANombre
DISEÑOS EN AIRE ACONDICIONADO CONSTR. Y MANTTO / JOSE LUIS GARCIA CRUZ	DISENOS EN AIRE ACONDICIONADO CONSTR. Y MANTTO / JOSE LUIS GARCIA CRUZ	14	CTANombre
SEÃ‘AL	SENAL	14	CMODescripcion
TELEFÃ“NICA	TELEFONICA	3	CMODescripcion
INSTALACIÃ“N	INSTALACION	88	CMODescripcion
REUBICACIÃ“N	REUBICACION	10	CMODescripcion
REVISIÃ“N	REVISION	10	CMODescripcion
COTIZACIÃ“N	COTIZACION	2	CMODescripcion
COLOCACIÃ“N	COLOCACION	2	CMODescripcion
ADQUISICIÃ“N	ADQUISICION	2	CMODescripcion
MODIFICACIÃ“N	MODIFICACION	4	CMODescripcion
PROGRAMACIÃ“N	PROGRAMACION	4	CMODescripcion
ALIMENTACIÃ“N	ALIMENTACION	5	CMODescripcion
CONTRATATCIÃ“N	CONTRATACION	2	CMODescripcion
DAÃ‘OS	DATOS	2	CMODescripcion
RECONEXIÃ“N	RECONEXION	14	CMODescripcion
LÃNEA	LINEA	8	CMODescripcion
CONEXIÃ“N	CONEXION	6	CMODescripcion
PORTÃ“N	PORTON	2	CMODescripcion
EXTENSIÃ“N	EXTENSION	2	CMODescripcion
INSTALACIÃ“N	INSTALACION	10	AMODescConcepto
AMPLIACIÃ“N	AMPLIACION	2	AMODescConcepto
EXPANSIÃ“N	EXPANSION	2	AMODescConcepto

LIQUIDACIÃ“N	LIQUIDACION	4	AMODescConcepto
REVISIÃ“N	REVISION	2	AMODescConcepto
COMUNICACIÃ“N	COMUNICACION	2	AMODescConcepto
CENTRALIZACIÃ“N	CENTRALIZACION	2	AMODescConcepto

Anexo 10. Detalle del Paso 3 de la etapa de pre-procesamiento. Corregir Inconsistencias en el Dataset 2.

DATO INCONSISTENTE	SE SUSTITUYO POR:	REEEMPLAZOS	ATRIBUTO
SEÃ‘AL	SENAL	16	CMODescripcion
TELEFÃ“NICA	TELEFONICA	3	CMODescripcion
INSTALACIÃ“N	INSTALACION	103	CMODescripcion
REUBICACIÃ“N	REUBICACION	10	CMODescripcion
REVISIÃ“N	REVISION	10	CMODescripcion
COTIZACIÃ“N	COTIZACION	2	CMODescripcion
2Âº	2DO	2	CMODescripcion
COLOCACIÃ“N	COLOCACION	2	CMODescripcion
ADQUISICIÃ“N	ADQUISICION	2	CMODescripcion
MODIFICACIÃ“N	MODIFICACION	6	CMODescripcion
PROGRAMACIÃ“N	PROGRAMACION	4	CMODescripcion
ALIMENTACIÃ“N	ALIMENTACION	7	CMODescripcion
CONTRATATCIÃ“N	CONTRATACION	2	CMODescripcion
DAÃ‘OS	DATOS	2	CMODescripcion
RECONEXIÃ“N	RECONEXION	14	CMODescripcion
LÃ□NEA	LINEA	8	CMODescripcion
CONEXIÃ“N	CONEXION	6	CMODescripcion
PORTÃ“N	PORTON	2	CMODescripcion
EXTENSIÃ“N	EXTENSION	2	CMODescripcion
AMPLIACIÃ“N	AMPLIACION	2	CMODescripcion
EXPANSIÃ“N	EXPANSION	2	CMODescripcion
LIQUIDACIÃ“N	LIQUIDACION	4	CMODescripcion
COMUNICACIÃ“N	COMUNICACION	4	CMODescripcion
CENTRALIZACIÃ“N	CENTRALIZACION	7	CMODescripcion
ELÃ‰CTRICO	ELECTRICO	1	CMODescripcion

RE-INSTALACION	REINSTALACION	8	CMODESCORIGINAL
ABLES	CABLES	2	CMODESCORIGINAL
MAGNÃ‰TICO	MAGNETICO	1	CMODESCORIGINAL
REMODELACIÃ“N	REMODELACION	5	AMODESCCONCEPTO
REPARACIÃ“N	REPARACION	2	AMODESCCONCEPTO
CARLOS ISAÃ□AS HERNÃ□NDEZ LASTRA	CARLOS ISAIAS HERNANDEZ LASTRA	29	CTANOMBRE
CARLOS SCHMIDT NUÃ‘EZ -CLAUDIA ALVARADO	CARLOS SCHMIDT NUNEZ - CLAUDIA ALVARADO	12	CTANOMBRE
CASA DISEÃ‘OS	CASA DISENOS	40	CTANOMBRE
CIRENIA SANTOS CAÃ‘ETE	CIRENIA SANTOS CANETE	56	CTANOMBRE
GABRIELA PIÃ‘EIRO	GABRIELA PINEIRO	64	CTANOMBRE
JULIETA ELENA NUÃ‘EZ SAAVEDRA	JULIETA ELENA NUNEZ SAAVEDRA	21	CTANOMBRE
LAURA SAN JUAN BAÃ‘OS	LAURA SAN JUAN BANOS	88	CTANOMBRE
MERCEDES GARCIA PEREZ - CARLOS ALBERTO MUÃ‘OZ GARCI	MERCEDES GARCIA PEREZ - CARLOS ALBERTO MUNOZ GARCIA	84	CTANOMBRE
OLGA LUCY GUTIERREZ TRIVIÃ‘O	OLGA LUCY GUTIERREZ TRIVINO	90	CTANOMBRE
OSCAR PEÃ‘A HERNANDEZ	OSCAR PENA HERNANDEZ	10	CTANOMBRE
RAUL PEÃ‘A TOVAR	RAUL PENA TOVAR	58	CTANOMBRE
REYNA ELIZABETH DOMÃ□NGUEZ VIVEROS	REYNA ELIZABETH DOMINGUEZ VIVEROS	81	CTANOMBRE
RICARDO COUTIÃ‘O LARA	RICARDO COUTINO LARA	158	CTANOMBRE
SAUL MUÃ‘OZ RODRIGUEZ	SAUL MUNOZ RODRIGUEZ	51	CTANOMBRE
SUSANA PATRICIA AKE CHIÃ‘AS	SUSANA PATRICIA AKE CHINAS	160	CTANOMBRE
SAMUEL YAÃ‘EZ PICAZO	SAMUEL YANEZ PICAZO	179	CTANOMBRE
DISEÃ‘OS EN AIRE ACONDICIONADO CONSTR. Y MANTTO / JOSE LUIS GARCIA CRUZ	DISENOS EN AIRE ACONDICIONADO CONSTR. Y	16	CTANOMBRE

	MANTTO / JOSE LUIS GARCIA CRUZ		
GRUAS Y MANIOBRAS DE MÃ‰XICO, S. A. DE C. V.	GRUAS Y MANIOBRAS DE MEXICO, S. A. DE C. V.	270	CLIRAZONSOCIAL
ING FRANCISCO J DIAZ MUÃ‘OZ	ING FRANCISCO J DIAZ MUNOZ	49	CLIRAZONSOCIAL
N	NETSOLUTION SHOP S. A. DE C. V.	8	CLIRAZONSOCIAL
JORGE LUIS VAZQUEZ ORDOÃ‘EZ	JORGE LUIS VAZQUEZ ORDONEZ	26	CLIRAZONSOCIAL
DISEÃ‘OS EN AIRE ACONDICIONADO CONSTR. Y MANTTO / JOSE LUIS GARCIA CRUZ	DISENOS EN AIRE ACONDICIONADO CONSTR. Y MANTTO / JOSE LUIS GARCIA CRUZ	14	CLIRAZONSOCIAL
DOMÃ□NGUEZ	DOMINGUEZ	81	CLIAPATERNO
HERNÃ□NDEZ	HERNANDEZ	29	CLIAPATERNO
MUÃ‘OZ	MUNOZ	200	CLIAPATERNO
YAÃ‘EZ	YANEZ	100	CLIAPATERNO
COUTIÃ‘O	COUTINO	79	CLIAPATERNO
NUÃ‘EZ	NUNEZ	234	CLIAPATERNO
PIÃ‘EIRO	PIÃ‘EIRO	34	CLIAPATERNO
PEÃ‘A	PENA	77	CLIAPATERNO
CAÃ‘ETE	CANETE	28	CLIAMATERNO
TRIVIÃ‘O	TRIVINO	45	CLIAMATERNO
BAÃ‘OS	BANOS	132	CLIAMATERNO
CHIÃ‘AS	CHINAS	80	CLIAMATERNO
ZUÃ‘IGA	ZUNIGA	67	CLIAMATERNO
CARLOS ISAÃ□AS	CARLOS ISAIAS	29	CLINOMBRE
GONZALEZ PAGÃ‰S 256-C ENTRE CAÃ‘ONERO TAMPICO Y ABAS	GONZALEZ PAGES 256-C ENTRE CANONERO TAMPICO Y ABAS	945	CLIDOMICILIO
WASHINGTON 510 ESQ. PINZÃ“N	WASHINGTON 510 ESQ. PINZÃ“N	248	CLIDOMICILIO

ESPAÃ‘A NO.194-A ENTRE PASEO LAS FLORES Y JUAN PABL	ESPANA NO.194-A ENTRE PASEO LAS FLORES Y JUAN PABLO II	118	CLIDOMICILIO
MAGNOLIAS NO. 225 ENTRE ESPAÃ‘A Y 18 DE MARZO	MAGNOLIAS NO. 225 ENTRE ESPAÃ‘A Y 18 DE MARZO	212	CLIDOMICILIO

Anexo 11. Estructura del Dataset con 31 atributos.

DATASET 31 ATRIBUTOS		
CMODIA	AMOMESPAGO	CTAMesFinServicio
CMOMES	AMOANIOPAGO	CTAAnioFinServicio
CMOMESVENC	AMOCAPTURO	CTADiaAlta
CMOPeriodoMes	AMOSTATUS	CTAMesAlta
CMOStatus	AMOTRANSACCION	CTAAnioAlta
CMODiaCondonado	AMONUMEMPRESA	CTAPreventivo
CMOMesCondonado	CTADiaInicioServicio	CTARuta
CMOAnioCondonado	CTAMesInicioServicio	CTAFormaPago
AMOCONCEPTO	CTAAnioInicioServicio	CLICIUDAD
AMODIAPAGO	CTADiaFinServicio	CLIClasificacion
		tipoCliente

Anexo 12. Experimentos realizados combinando algoritmo Dl4jmlpClassifier y Meta-clasificadores META en WEKA con Dataset de 31 Atributos.

			CROSS VALIDATION (10)	**PORCENTAJE DE DIVISION (66.665%)**	**PORCENTAJE DE DIVISION (85%)**
	Clasificador ensamblado	**DeepLearning**	**Porcentaje**	**Porcentaje**	**Porcentaje**
1		Dl4JMlpClassifier	64.2616 b	64.4957	65.6085
2	AdBoostM1	Dl4JMlpClassifier	67.3724	68.568	68.5185
3	AttributeSelectedClassifier	Dl4JMlpClassifier	65.7901	66.7745	73.809
4	Bagging	Dl4JMlpClassifier	70.0742	66.7745	70.1058
5	ClassificationViaRegretion	Dl4JMlpClassifier	57.1544	57.2141	56.6138
6	CVParameterSelection	Dl4JMlpClassifier	64.2616	64.4957	65.6085
7	FilteredClasifier	Dl4JMlpClassifier	**77.4241**	**74.9595**	**81.746**
8	LogitBoost	Dl4JMlpClassifier	57.1544	57.2141	56.6138
9	MultiClassClassifier	Dl4JMlpClassifier	64.3336	64.8598	64.5503
10	MultiSearch	Dl4JMlpClassifier	64.2616	64.4957	65.6085
11	OrdinalClassClassifier	Dl4JMlpClassifier	64.1852	65.5879	66.6667
12	RandomCommittee	Dl4JMlpClassifier	69.9573	71.2109	70.1058
13	RandomizableFilteredClassifier	Dl4JMlpClassifier	65.6777	69.4175	68.5185
14	RandomSubSpace	Dl4JMlpClassifier	69.0088	70.712	66.9312
15	Stacking	Dl4JMlpClassifier	52.9018	57.2141	56.6138
16	WeightInstancesHandlerWrapper	Dl4JMlpClassifier	64.2616	64.4957	65.6085

Anexo 13. Estructura del Dataset con 72 atributos.

DATASET DE 72 ATRIBUTOS			
CMOCONCEPTO	CMOStatAux	CTADiaFinServicio	CLITipo
CMOVenta	CMOClaveFac	CTAMesFinServicio	CLILimiteCredito
CMODIA	CMOSerieFac	CTAAnioFinServicio	CLIValoracionCredito
CMOMES	CMONUMEMPRESA	CTADiaAlta	CLIUsuOtorgoCredito
CMOANIO	CMOUnitario	CTAMesAlta	CLIStatusCredito
CMODIAVENC	CMODiaCondonado	CTAAnioAlta	CLINUMEMPRESA
CMOMESVENC	CMOMesCondonado	CTAStatus	CLIImprimeRecibos
CMOANIOVENC	CMOAnioCondonado	CTAIva	CLITipoPago
CMOPeriodoAnio	CMOParcialidad	CTANumEmpresa	CLICobrador
CMOPeriodoMes	AMOCONCEPTO	CTAAnual	CLIImpreso
CMOCapital	AMODIAPAGO	CTAAnualMes	CLIClasificacion
CMOInteres	AMOMESPAGO	CTAPreventivo	tipoCliente
CMOStatus	AMOANIOPAGO	CTARuta	
CMOIMPORTEPAGADO	AMOCAPTURO	CTAFormaPago	
CMOCAPTURO	AMOSTATUS	CLIId	
CMOIMPRTAL	AMOTRANSACCION	CLICIUDAD	
CMOTransaccion	AMONUMEMPRESA	CLIMUNICIPIO	
CMOCaja	CTADiaInicioServicio	CLIESTADO	
CMOfolioFac	CTAMesInicioServicio	CLIPAIS	
CMOUsuAux	CTAAnioInicioServicio	CLISTATUS	

Anexo 14. Experimentos realizados combinando algoritmo Dl4jmlpClassifier y Meta-clasificadores META en WEKA con Dataset de 72 atributos.

			CROSS VALIDATION (10)	PORCENTAJE DE DIVISION (66.665%) 14830 Trainig - 7415 Test	PORCENTAJE DE DIVISION (85%) 18908 Training - 3337 Test
PRUEBA	**META**	**FUNCION**	**Porcentaje**	**Porcentaje**	**Porcentaje**
1		Dl4jMlpClassifier	67.76%	67.28%	68.25%
2	AdaBoostM1	Dl4jMlpClassifier	70.85%	71.56%	71.69%
3	AttributeSelectedClassifier	Dl4jMlpClassifier	66.68%	71.65%	74.34%
4	AutoWEKAClassifier		85.09%		
5	Bagging	Dl4jMlpClassifier	73.92%	71.33%	72.75%
6	ClassificationViaRegression	Dl4jMlpClassifier	57.15%	57.22%	56.61%
7	CostSensitiveClassifier	Dl4jMlpClassifier	On demand cost file doesn't exist	On demand cost file doesn't exist	On demand cost file doesn't exist
8	CVParameterSelection	Dl4jMlpClassifier	67.87%	67.28%	66.93%
9	**FilteredClassifier**	**Dl4jMlpClassifier**	79.46%	79.35%	**81.75%**
10	GridSearch	Dl4jMlpClassifier	Thread based execution of evaluation tasks failed	Thread based execution of evaluation tasks failed	Thread based execution of evaluation tasks failed
11	IterativeClassifierOptimizer		No permitio la combinacion con Dl4MlpClassifier	No permitio la combinacion con Dl4MlpClassifier	No permitio la combinacion con Dl4MlpClassifier
12	LogitBoost	Dl4jMlpClassifier	Tiempo Excedido	57.22%	56.61%
13	MetaCost	Dl4jMlpClassifier	On demand cost file doesn't exist	On demand cost file doesn't exist	On demand cost file doesn't exist
14	MultiClassClassifier	Dl4jMlpClassifier	67.39%	66.96%	69.05%
15	MultiClassClassifierUpdateable	Dl4jMlpClassifier	Base classifier must be updateable	Base classifier must be updateable	Base classifier must be updateable
16	MultiScheme	Dl4jMlpClassifier	66.75%	67.04%	64.29%
17	MultiSearch	Dl4jMlpClassifier	Cannot handle a multi-valued nominal class	67.28%	66.93%
18	OneClassClassifier	Dl4jMlpClassifier	Target Class value doesnt exist	Target Class value doesnt exist	Target Class value doesnt exist
19	OrdinalClassClasifier	Dl4jMlpClassifier	67.13%	66.19%	69.05%
20	RandomCommittee	Dl4jMlpClassifier	72.83%	73.36%	72.49%
21	RandomizableFilteredClassifier	Dl4jMlpClassifier	68.27%	67.55%	70.11%
22	RandomSubSpace	Dl4jMlpClassifier	71.68%	71.27%	70.63%
23	RegressionByDiscretization	Dl4jMlpClassifier	Opcion Inhabilitada	Opcion Inhabilitada	Opcion Inhabilitada
24	Stacking	Dl4jMlpClassifier	57.15%	57.22%	56.61%
25	Vote	Dl4jMlpClassifier	67.87%	67.28%	66.93%
26	WeightedInstancesHandlerWrapper	Dl4jMlpClassifier	67.87%	67.28%	66.93%

Anexo 15. Resultados de la clasificación de WEKA utilizando el Dataset con 31 atributos.

Time taken to build model: 195.89 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.08 seconds

=== Summary ===

Correctly Classified Instances	305	80.6878 %
ncorrectly Classified Instances	73	19.3122 %
Kappa statistic	0.6514	
Mean absolute error	0.1294	
Root mean squared error	0.3533	
Relative absolute error	35.6515 %	
Root relative squared error	82.5746 %	
Total Number of Instances	378	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.855	0.134	0.893	0.855	0.874	0.717	0.918	0.932	MOROSO
	0.774	0.159	0.725	0.774	0.749	0.607	0.892	0.762	NORMAL
	0.613	0.035	0.613	0.613	0.613	0.578	0.884	0.680	ANTICIPADO
Weighted Avg.	0.807	0.135	0.811	0.807	0.808	0.667	0.906	0.852	

=== Confusion Matrix ===

a	b	c	<-- classified as
183	29	2	a = MOROSO
20	103	10	b = NORMAL
2	10	19	c = ANTICIPADO

Anexo 16. Resultados de la clasificación de WEKA utilizando el Dataset con 72 atributos.

Time taken to build model: 164.22 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.04 seconds

=== Summary ===

Correctly Classified Instances	309	81.746 %
Incorrectly Classified Instances	69	18.254 %
Kappa statistic	0.6687	
Mean absolute error	0.1204	
Root mean squared error	0.3418	
Relative absolute error	33.1714 %	
Root relative squared error	79.8857 %	
Total Number of Instances	378	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.874	0.159	0.878	0.874	0.876	0.715	0.919	0.929	MOROSO
	0.759	0.127	0.765	0.759	0.762	0.634	0.902	0.791	NORMAL
	0.677	0.035	0.636	0.677	0.656	0.625	0.939	0.702	ANTICIPADO
Weighted Avg.	0.817	0.137	0.818	0.817	0.818	0.679	0.915	0.862	

=== Confusion Matrix ===

a	b	c	<-- classified as
187	24	3	a = MOROSO
23	101	9	b = NORMAL
3	7	21	c = ANTICIPADO

Anexo 17. Métricas de validación de la clasificación en WEKA.

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.874	0.159	0.878	0.874	0.876	0.715	0.919	0.929	MOROSO
	0.759	0.127	0.765	0.759	0.762	0.634	0.902	0.791	NORMAL
	0.677	0.035	0.636	0.677	0.656	0.625	0.939	0.702	ANTICIPADO
Weighted Avg.	0.817	0.137	0.818	0.817	0.818	0.679	0.915	0.862	