RECONOCIMIENTO DE PATRONES PARA IDENTIFICACIÓN DE USUARIOS EN ACCESOS INFORMÁTICOS

CÉSAR BYRON GUEVARA MALDONADO

MÁSTER EN INVESTIGACIÓN EN INFORMÁTICA, FACULTAD DE INFORMÁTICA,
UNIVERSIDAD COMPLUTENSE DE MADRID

Trabajo de Fin de Master en Ingeniería Informática para la Industria
Curso académico: 2011/2012

Director / Colaborador:
Matilde Santos Peñas/José Antonio Martín Hernández

Calificación obtenida: 8/10

Autorización de Difusión

CÉSAR BYRON GUEVARA MALDONADO

21/06/2012

El abajo firmante, matriculado en el Máster en Investigación en Informática de la Facultad de Informática, autoriza a la Universidad Complutense de Madrid (UCM) a difundir y utilizar con fines académicos, no comerciales y mencionando expresamente a su autor el presente Trabajo Fin de Máster: "RECONOCIMIENTO DE PATRONES PARA IDENTIFICACIÓN DE USUARIOS EN ACCESOS INFORMÁTICOS", realizado durante el curso académico 2011-2012 bajo la dirección de Matilde Santos Peñas y con la colaboración externa de dirección de José Antonio Martín Hernández en el Departamento de Arquitectura de Computadores y Automática, y a la Biblioteca de la UCM a depositarlo en el Archivo Institucional E-Prints Complutense con el objeto de incrementar la difusión, uso e impacto del trabajo en Internet y garantizar su preservación y acceso a largo plazo.

Resumen en castellano

La detección y control de intrusos o accesos no autorizados en los sistemas informáticos ha sido desde siempre un tema a tener en cuenta en los sistemas de información donde la seguridad, integridad y privacidad de la información son aspectos fundamentales. El avance del conocimiento y la tecnología es cada vez mayor, lo que permite el desarrollo y aplicación de sistemas informáticos más sofisticados y eficientes, pero también aumenta la posibilidad de que sean vulnerados mediante accesos no legítimos.

En este trabajo se plantea el uso de diversas técnicas automáticas para identificar a los usuarios que acceden a datos fundamentales de los sistemas y comprobar si el acceso está o no permitido. Se han aplicado técnicas avanzadas e inteligentes para el análisis y aplicación de minería de datos, para obtener patrones de comportamiento; como son los árboles de decisión y las redes neuronales artificiales (RNA). Con ellas se obtienen perfiles dinámicos (patrones) de usuario. La hipótesis principal del trabajo es una solución efectiva para la detección de intrusos en los sistemas informáticos de información.

Palabras clave

Intrusión, reconocimiento de patrones, perfil de comportamiento, accesos no legítimos, seguridad, bases de información, detección intrusos.

Resumen en inglés

The detection and control of intruders or unauthorized access to computer systems has always been an issue to consider in information systems where security, integrity and privacy of information are key issues. The advancement of knowledge and technology is getting greater, allowing the development and application of more sophisticated and efficient computer systems, but also increases the possibility of being violated by illegitimate accesses.

The propose of this paper is to use different automated techniques to identify users accessing systems critical data, to check whether or not access is allowed. Advanced and intelligent analysis and application of data mining techniques have been applied for obtaining patterns of behavior, such as decision trees and artificial neural networks. Dynamic user profiles (patterns) are obtained with them. The main hypothesis of the work is an effective solution for the detection of intruders in computer information systems.

KEYWORDS

Intrusion, pattern recognition, behavioral profile, illegitimate access, security, information databases, detecting intruders.

Índice de contenidos

Contenido

Índice de figuras

Índice de tablas

Agradecimientos

Agradezco a mi Dios el que siempre ha estado conmigo, para brindarme la fortaleza y decisión para seguir adelante. A mi madre que nunca dudó de mí después de las circunstancias adversas que hemos pasado solos. A mí país que ha sido el forjador de esta oportunidad y con el cual estaré eternamente agradecido, y haré todo los posible para mantener una patria digna, soberana y en paz, mejorando cada día su entorno.

Agradezco mucho a mis directores Matilde Santos Peñas y José Antonio Martín Hernández, que me han ayudado con su conocimiento, experiencia y paciencia para realizar este trabajo. Sin ellos no habría sido posible realizar este estudio.

Capítulo 1- Introducción

En la actualidad, los sistemas informáticos a nivel mundial han constituido una gran red de información [1,2]. Los usuarios que acceden a esta red han generado datos sobre su comportamiento sin percatarse de que, al utilizar dichos sistemas, están brindando información, por ejemplo, de sus preferencias personales, posición geográfica, actividades realizadas, etc. Esa información personal describe un patrón de actividad que permite utilizarlo como una firma o un identificador único de cada persona y así facilitar sistemas de detección de intrusiones de una manera más efectiva.

La información personal obtenida se registra en diversas bases de manera sistemática, sin embargo, su análisis es muy complejo debido a sus enormes dimensiones. Por ello se han creado herramientas para el análisis de esa información de una manera eficiente y rápida, lo que permitirá obtener un patrón de comportamiento de cada persona y así generar un perfil de uso para cada usuario que utilice el sistema.

Teniendo en cuenta que el usuario posee un comportamiento regular (patrón) [15], este patrón debe ser dinámico para ajustarse a las actividades y necesidades diarias de cada persona. Por ello se ha realizado el análisis con técnicas ágiles, robustas y eficientes que puedan modelar o ajustarse a los continuos cambios de comportamiento de los usuarios. En el estudio se utilizaran

metodologías de clasificación e identificación de patrones para modelar el perfil de cada usuario. Las metodologías de clasificación de datos se describen como técnicas de aprendizaje automático que utilizan un conjunto de datos para construir un "*modelo*" o una "representación" de la regularidad existente (subyacente) en los datos, analizando una situación real con posibles opciones y, a partir de una determinada condición, ser excluyentes y así bridar un acceso seguro, adaptable y confiable en todos los sistemas informáticos.

En este estudio se van a utilizar varias técnicas para el análisis de información, dentro de la minería de datos, los árboles de decisión y las redes neuronales artificiales (RNA). Una de las técnicas que se utilizará son los árboles de decisión, los cuales representan la información de forma más comprensible, utilizando un conjunto de condiciones y reglas para obtener como resultado modelos inteligibles para los seres humanos. Esta técnica proporciona una visión gráfica para la toma de decisiones, así como también especifica las variables que son evaluadas, qué acciones deben ser tomadas y el orden en que la toma de decisión será efectuada.[25,27,31]

Se ha demostrado que los árboles de decisión son eficaces cuando es necesario describir problemas con más de una dimensión o condición, además de ser útiles para identificar los requerimientos de la información crítica que rodean al proceso de decisión [3].

De forma complementaria se utilizará la técnica de redes neuronales artificiales, ya que es una de las más utilizadas y efectivas para el reconocimiento de patrones, reducción de la dimensionalidad, agrupamiento y clasificación [3, 17]. Las redes neuronales, son un método de aprendizaje cuyo objetivo es imitar a las neuronas biológicas y la forma en la que éstas procesan la información.

En el presente estudio se aplicarán las técnicas mencionadas anteriormente para obtener un perfil de usuario dinámico, que permita ser un identificador único en el acceso a los sistemas informáticos y con ello mejorar la detección de intrusos o de agentes maliciosos en los sistemas.

1.1. Descripción del problema

El sector público de la República del Ecuador posee una red de sistemas de información de suma importancia, por lo cual es necesario tener un adecuado control de los usuarios que acceden a

dicha información. Los esfuerzos que se han realizado para mejorar la confidencialidad de estos datos han resultado en sistemas de seguridad complejos, que hacen el acceso a los usuarios más complicado y cada vez con mayores inconvenientes. Por ello se ha propuesto realizar una investigación sobre el comportamiento del usuario en la utilización de la red de información, para que se pueda controlar el acceso a los sistemas de una forma más automatizada, y así detectar intrusiones de agentes no autorizados y la fuga de información reservada de las entidades gubernamentales de este país.

Para llevar a cabo este estudio se tuvo acceso a la base de datos de una institución pública ecuatoriana, por lo que la información presentada tendrá la debida confidencialidad y reserva. Además, la información es específicamente de actividades que los funcionarios públicos realizan en el sistema. Así que se tomará una muestra para poder realizar el presente estudio y modelar perfiles de usuarios en una red de información.

La realización del modelado, conlleva desarrollar un método para el pre-procesamiento y análisis de los mecanismos automáticos de clasificación, siendo éstas las principales contribuciones de este trabajo de investigación.

1.2. Objetivos

El objetivo principal de este trabajo es la investigación, aplicación y combinación de técnicas de clasificación y reconocimiento de patrones, para procesar la información disponible sobre el comportamiento de usuarios para la identificación de accesos fraudulentos en sistemas informáticos.

Objetivos Específicos

- Abordar un problema con múltiples variables utilizando información real.
- Realizar un pre-procesamiento de la información recolectada para ser analizada con metodologías de clasificación y de reconocimiento de patrones, utilizando varias herramientas de procesamiento de datos.
- Aplicación y combinación de metodologías de reconocimiento de patrones para la obtención de perfiles de usuarios.

- Desarrollar una herramienta de simulación en Matlab para aplicar técnicas de reconocimiento de patrones.
- Asignación de probabilidades de acierto en la detección de accesos fraudulentos.
- Apertura de una línea de investigación de gran interés práctica para el desarrollo de una tesis doctoral.

1.3. Estructura de la memoria

Esta memoria está organizada en seis capítulos. En el *capítulo inicial*, se presenta una breve descripción del problema que se aborda y la motivación, información relevante a tratar y objetivos principales de la investigación. En el *capítulo dos*, se explica la fundamentación teórica de las metodologías y técnicas utilizadas en la investigación, las cuales son, dentro de la minería de datos, los árboles de decisión y las redes neuronales artificiales; y la metodología que se propone como solución al problema de identificación de patrones de comportamiento de usuario que será una combinación de ellas. En los *capítulos tres y cuatro*, se expone de manera detallada la aplicación de las metodologías seleccionadas para el estudio así como los resultados obtenidos.

En el *capítulo cinco*, se presenta una implementación de las metodologías estudiadas y aplicadas para desarrollar una herramienta la cual simula la identificación de intrusos en un sistema informático.

Finalmente, en el *capítulo seis*, se presentan las conclusiones y trabajos futuros que pueden sobrevenir de este proyecto.

Capítulo 2 - Estado del Arte

Hernández-Orallo [15, 27] define al KDD (*Knowledge Discovery in Database*) como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de datos". En esta definición se resumen las características primordiales del conocimiento que se extrae de la base de datos:

- **válido:** hace referencia a los patrones que deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no solo para aquellos que han sido usados para su obtención.
- **novedoso:** que aporte algo desconocido tanto para el sistema como preferiblemente para el usuario.
- **potencialmente útil:** la información debe conducir a acciones que aporten algún tipo de beneficio para el usuario.
- **comprensible:** la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, una información incomprensible no proporciona conocimiento[15, 27, 39].

El KDD, como se define como un proceso complejo, requiere no sólo la obtención de modelos o patrones sino también la evaluación e interpretación de los mismos, como se presenta en la siguiente figura 1:

Figura 1: Proceso de extracción del conocimiento en la minería de datos. Tomado de Hernández Orallo [27]

Para realizar el proceso de obtención del conocimiento se deben seguir las siguientes etapas, como se muestra en la siguiente figura 2:

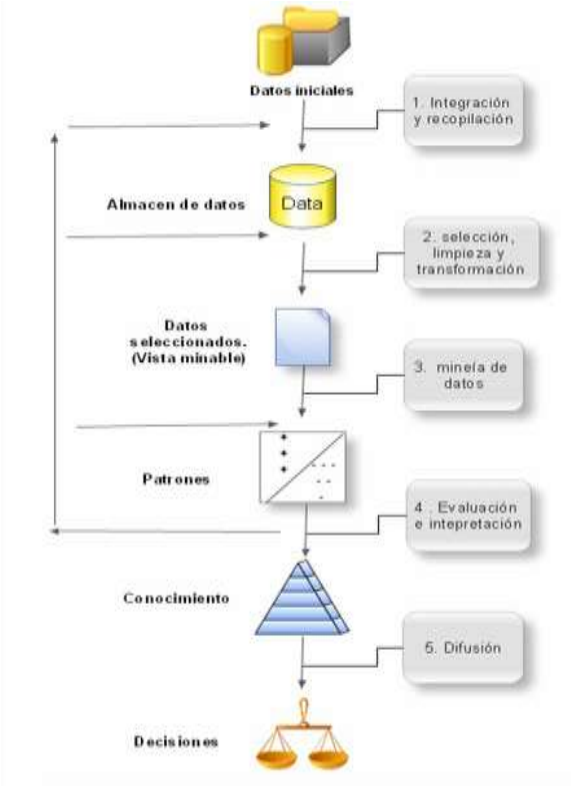

Figura 2: Fases del proceso de obtención del conocimiento de base de datos, KDD. Tomado de Hernández-Orallo [27]

Ya que el KDD es un proceso interactivo e iteractivo, en las salidas de algunas etapas requieren volver a etapas anteriores, y a menudo realizar varias iteraciones para obtener conocimiento de calidad.

De acuerdo a la bibliografía en el área [15, 27, 39], las técnicas más frecuentes pueden ser catalogadas en:

- **Descriptivas:** El objetivo de estos procedimientos es la búsqueda de la caracterización o discriminación de un conjunto de datos. Las técnicas más conocidas son: agrupamiento o clustering, reglas de asociación, análisis de patrones secuenciales, análisis de componentes principales, detección de desviación.
- **Predictivas:** El propósito de estos métodos es obtener una hipótesis la cual pueda clasificar a nuevos individuos. Los algoritmos principales son: regresión y clasificación (árboles de decisión, clasificación bayesiana, redes neuronales, algoritmos genéticos, conjuntos y lógica difusa).

Para tener una mejor comprensión de las técnicas enfocadas a la extracción del conocimiento a continuación se muestra en la figura 3 la clasificación de las técnicas de minería de datos según sus características:

Cuando realizamos un proceso de minería de datos, necesitamos tener en cuenta el conocimiento previo; este puede derivar del proceso mismo (elección de variables, técnicas, algoritmos, interpretación de resultados) o del dominio de aplicación.[4, 15, 27]

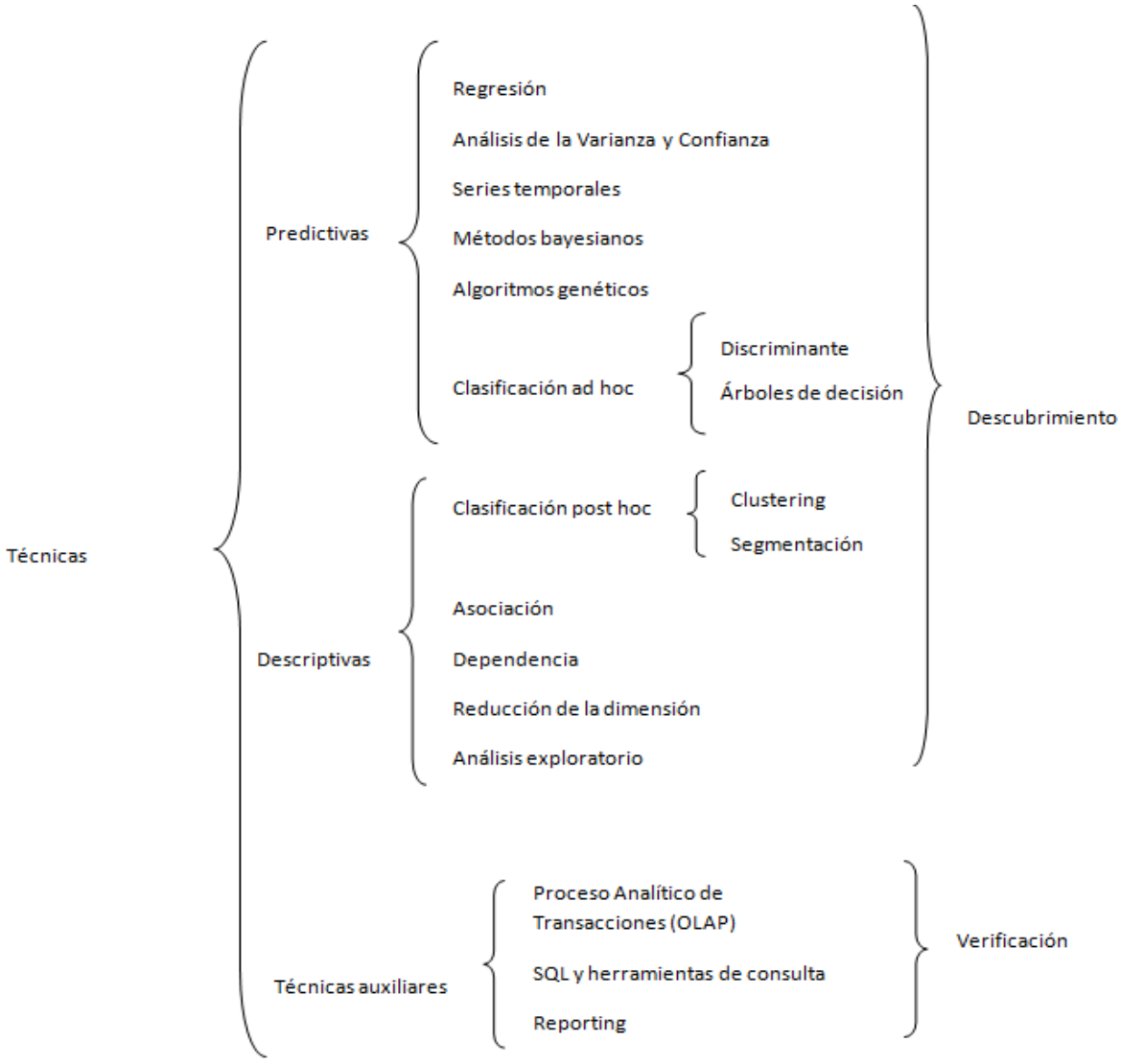

Figura 3: Técnicas de clasificación. Tomado de Pérez-López. [24]

En la figura anterior, se puede apreciar que existen diferentes técnicas para realizar la minería de datos, así como también su clasificación según las actividades que se requiere utilizar cada una de ellas.

2.1. Procesamiento de la información

En esta fase, se determinan las fuentes de información que pueden ser útiles para el estudio, además de transformar todos los datos a un formato común, y finalmente ingresarlo a un almacén de datos para que se consiga unificar toda la información obtenida para una mejor visualización

y análisis de la información. Por otra parte, la información recogida tiende a ser más operativa y se pueden detectar y solucionar inconsistencias que existan. El almacén de datos facilita la navegación y visualización de datos, lo que ayuda enormemente al análisis de los aspectos importantes del estudio. Este proceso tiene varias etapas, donde la información es sometida a filtros y transformaciones para que los datos brinden el conocimiento esperado y correcto, eliminando errores que pueden afectar enormemente al desenvolvimiento de las fases posteriores. [12,14, 27].

2.2. Integración y recopilación

Esta fase es una de las más importantes en el proceso de minería de datos ya que la información que se va a recolectar y depurar, por lo que debe ser lo más idónea para su estudio y análisis [15]. En este proceso será normal, en algunos casos, integrar bases de datos de distintos departamentos, sucursales de la institución, o también bases de datos externas (como censo poblacional, datos demográficos, climatológicos, etc.) para poder realizar un análisis más eficiente de los mismos. Se disponen de base de datos con diferentes formatos de los atributos, claves primarias, claves foráneas, índices de las tablas, etc. Por eso se debe en primer lugar integrar todos estos datos en un mismo repositorio, eliminando inconsistencias, redundancias, y utilizando procedimientos de almacenamiento que permitan integrar los datos de varias bases de datos en un único repositorio, como se puede ver en la figura 4.[10, 20, 27]

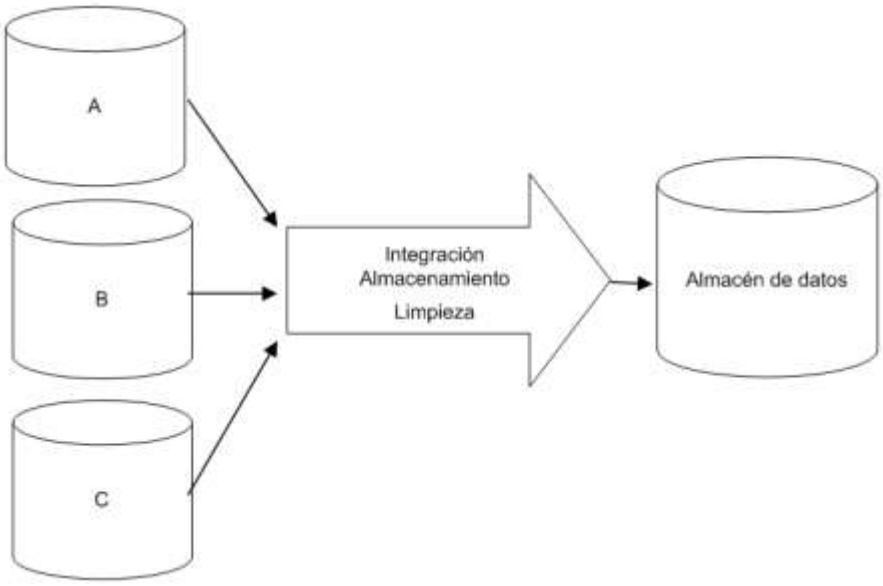

Figura 4: Integración en un almacén de datos. Tomado Hernández-Orallo [27]

Este repositorio de datos se utiliza para poder ingresar, agregar y comparar información de una manera sofisticada y eficiente. Por eso, el almacén de datos es diseñado con múltiples dimensiones, donde cada dimensión es un atributo o conjunto de atributos relacionados entre sí, llamado también con el termino de "hechos", como por ejemplo, ventas de un producto, temperatura en un país, etc. Esta transaccionalidad de los datos en múltiples dimensiones es adecuada para el procesamiento analítico en línea (*on-line analytical processing, OLAP*), que facilita el análisis de la información, realizando proyecciones, planificación, patrones y otras tareas de toma de decisiones en las empresas.[6, 21, 27]

Como se puede apreciar, la efectividad de los repositorios de datos en el proceso de minería de datos es amplia, además de ser una herramienta flexible para integrar varias bases de datos sin ningún problema.

2.2.1. Análisis de variables

El análisis de los atributos de cada una de las bases de datos que posteriormente van a ser parte de un único repositorio de información, tiene relación con el conocimiento que se va a obtener. Por eso se debe analizar de una forma más detallada cada uno de los atributos que intervienen en el estudio y su aporte en la extracción de la información.

Este proceso es muy importante ya que brinda información de cada uno de los atributos analizados de una manera más gráfica, además de proporcionar la tendencia, frecuencia y distribución de los datos y así decidir si son pertinentes para el estudio.

En el caso de la minería de datos este proceso puede ser parte de la selección y limpieza de los datos, pero puede ser beneficioso en cualquier etapa previa a la extracción de la información, ya que en algunos casos existen varios atributos que no brindan suficiente información para el proceso y deben ser retirados, o caso contrario, agregar atributos que no se tomaron en cuenta en el momento del diseño y construcción del repositorio de la información.

El proceso de análisis se inicia con una tabulación de la información con una muestra considerable de cada uno de los atributos que pretenden ser parte del repositorio de datos. Luego se muestran en una gráfica para observar la tendencia de cada una de las variables, así como su distribución según los periodos que se han considerado en la muestra. Como se puede apreciar, la figura 5 muestra la grafica de la edad en años de un conjunto de personas a modo de ejemplo.

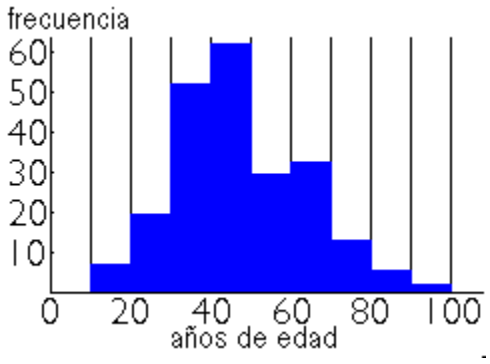

Figura 5: Diagrama de edades de un grupo de personas entre10 a 100 años. Tomado de Martín-Pliego [23]

La frecuencia en este atributo brinda la información adecuada para un estudio demográfico de un sector de la población. Por otra parte, es necesario analizar la distribución de los datos de cada atributo, y así obtener el conocimiento necesario del comportamiento cada uno de ellos. En la figura 6 se presenta la distribución de las variables genéricas, que permite visualizar la información de los atributos por medio de una grafica.[8, 16, 18,25]

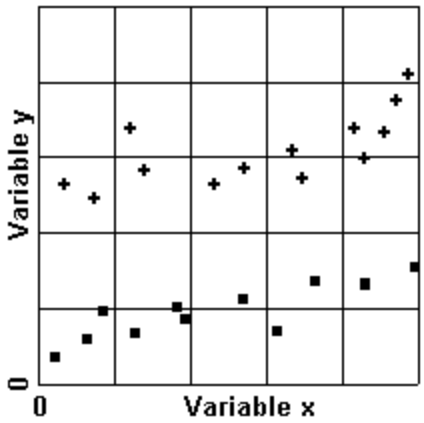

Figura 6: Diagrama de distribución de variables. Tomado de Martín-Pliego[18]

Finalmente, ya seleccionados los atributos que suministran información más relevante y una distribución de datos óptima, se procederá a diseñar y crear el repositorio de datos. [6, 10,16, 18]

2.2.2. Diseño del modelo multidimensional del repositorio de datos

La construcción del repositorio datos es una parte importante ya que debe ser diseñado con todos los atributos que intervienen en el análisis, ya que de esto depende un resultado eficiente al momento de utilizar las técnicas de extracción de la información [7]. La ventaja fundamental de los almacenes de datos es su diseño específico y su separación de las bases de datos transaccionales, además de:

- Facilitar el análisis en tiempo real.
- No interviene el procesamiento de las transacciones en línea de las demás bases de datos, es decir, que se trabaja sin que se realicen transacciones (consultas, inserciones, etc.) en la base de datos.

En la tabla 1, se presenta con más detalle las diferencias entre las bases de datos transaccionales y los almacenes de datos.

	BASE DE DATOS TRANSACCIONAL	ALMACEN DE DATOS
Propósito	Operaciones diarias. Soporte a las aplicaciones.	Recuperación de información, informes, análisis y minería de datos.
Tipo de datos	Datos de funcionamiento de la organización.	Datos útiles para el análisis, la sumarización, etc.
Características de los datos	Datos de funcionamiento, cambiantes, internos, incompletos…	Datos históricos, datos internos y externos, datos descriptivos…
Modelo de datos	Datos normalizados.	Datos en estrella, en copo de nieve, parcialmente desnormalizados, multidimensionales..
Número y tipo de usuarios	Cientos/miles: aplicaciones, operarios, administrador de la base de datos.	Decenas: directores, ejecutivos, analistas (granjeros, mineros).
Acceso	SQL. Lectura y escritura.	SQL y herramientas propias (slice & dice, dril/, rol/, pivot...). Lectura.

Tabla 1: Diferencias entre las bases de datos transaccionales y almacenes de datos. Tomado de Hernández-Orallo.[27]

Además, se debe tener en cuenta que los almacenes de datos incorporan la mayoría de la información en bases de datos transaccionales, lo que requiere un proceso de volcado de datos de las distintas bases hacia el repositorio para que pueda posteriormente realizar las tareas de análisis y extracción de la información.

El almacén de datos agrupa mayormente datos históricos o "hechos", que describen el comportamiento interno de la organización o del sistema del cual estamos obteniendo la información.

Para el diseño del repositorio de datos se debe crear un modelo multidimensional el cual permita organizar los datos según el objetivo de la extracción de la información. Por eso, para realizar el modelo se debe tener definido el objetivo principal del repositorio de datos, realizando las

preguntas más frecuentes: ¿cómo?, ¿cuándo?, ¿dónde?, que ayudará a crear el diseño de las dimensiones alineado al objetivo final de la investigación.[27, 38]

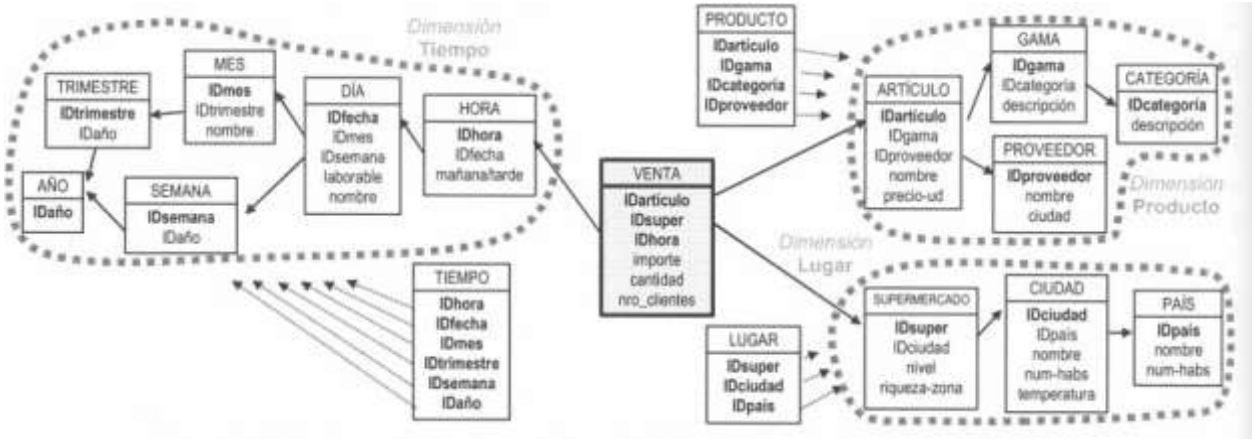

Figura 7: Implementación de un datamart utilizando tecnología relacional. Tomado de Hernández-Orallo. [27]

Para la construcción del almacén de datos se debe tener en cuenta las siguientes propiedades de las tablas que lo componen[13,27,27]:

- **Tablas copo de nieve (Snowflake tables):** para cada nivel de agregación de una dimensión se crea una tabla. Cada una de estas tablas tiene una clave primaria y tantas claves ajenas como sean necesarias para conectar con los niveles de agregación superiores.
- **Tabla de hechos (fact tables):** se crea una única tabla de hechos por datamart. En esta tabla se incluye un atributo para cada dimensión, que será clave ajena (*foreign key*) a cada una de las tablas copo de nieve de mayor detalle de cada dimensión.
- **Tablas estrella (star tables):** para cada dimensión se crea una tabla que tiene un atributo para cada nivel de agregación diferente en la dimensión. Cada uno de estos atributos es una clave ajena que hace referencia a tablas copo de nieve. Todos los atributos de la tabla forman la clave primaria.

Como se puede apreciar en la figura 7, los tres tipos de tablas están presentes en el diseño del almacén de datos, obteniendo una base de datos que puede ser objeto de análisis y aplicación de técnicas de extracción de la información, así como también es un método para facilitar el estudio de la información recabada.

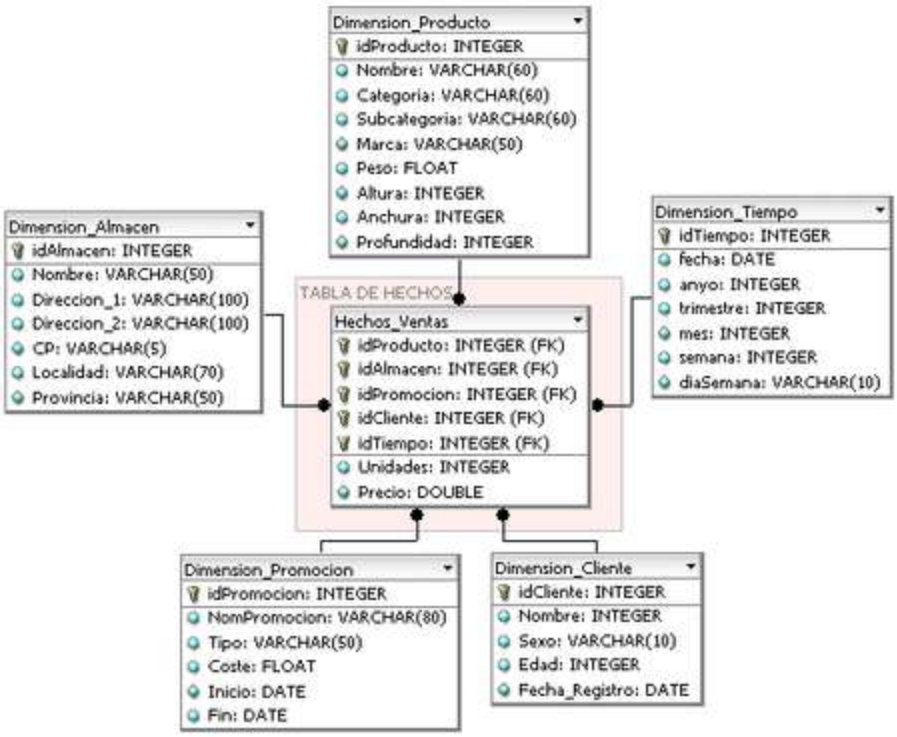

Figura 8: Esquema de un almacén de datos para su implementación. Tomado de [27,28]

En la figura 8, se muestra un almacén de datos que puede contener múltiples dimisiones de acuerdo a las necesidades que tenga la organización. En este ejemplo se muestra un almacén con distintos productos en venta, variedad de clientes, promociones, y fecha de las ventas.[7,26]

Finalmente, cuando ya se tiene el diseño del almacén de datos, lo único que se debe realizar el es volcado de información, utilizando procedimientos almacenados, consultas y otros tipos de herramientas para traspasar la información al repositorio.

2.3. Recopilación, limpieza y transformación

En esta fase lo que se quiere conseguir es la calidad de los datos recopilados en la fase anterior, por lo que seleccionarlos, limpiarlos y prepáralos para obtener una "*vista minable*". Este proceso es muy importante ya que algunos de los datos recopilados en la fase anterior no ofrecen información relevante para la tarea de minería de datos. Por otra parte, existen varios otros problemas que se presentan en la calidad de los datos, lo que puede ser perjudicial para el conjunto de datos, estos inconvenientes se definen como comportamientos anómalos de la

información. Los comportamientos anómalos pueden ser considerados como ruido y los algoritmos de minería de datos varias veces lo ignoran, pero esta información puede ser relevante para estudio, por lo que hay que realizar tareas de identificación de la importancia de esta información para poder observar su comportamiento relacionado con el resto de los datos.

Existe, por otra parte la falta de información o llamado *"datos en blanco"*, lo que constituye un problema dañino lo que puede conllevar a obtener resultados poco precisos o errados en el proceso de extracción del conocimiento.[27]

Para realizar esta tarea de selección, limpieza y transformación de la información, se debe utilizar distintas técnicas de identificación de datos para obtener los atributos más sobresalientes e importantes dentro del repositorio y así se procederá a eliminar inconsistencias, errores, vacios y si es necesario transformar la información existente en múltiples variables que permitan brindar un mejor resultado al momento de realizar la minería de datos.[27]

Por otra parte, en esta fase se utilizará técnicas estadísticas como son los histogramas, detección de valores anómalos y otros tipos de visualización, etc.

Integración y limpieza de datos

En el proceso de integración de datos es normal tener problemas en que la información recopilada de varias fuentes este incompleta o duplicada en la mayoría de los casos, es por eso que se realizan tareas de limpieza del repositorio de datos para detectar estas anomalías y corregirlas con varias herramientas que proporcionan una forma más eficiente de conseguir un conjunto de datos óptimo para la minería de datos.

Uno de los principales problemas que se presenta al momento de la integración de distintas fuentes de datos es la duplicidad de la información sobre un mismo objeto. Se presentan dos tipos de errores en la integración, los cuales son:

- **Unificación de datos:** dos o más objetos unifican patrones de diferentes individuos lo que es un problema al momento de la extracción del conocimiento.
- **Dos o más registros iguales de un mismo objeto:** este problema se presenta cuando la información de un mismo individuo se encuentra duplicada, lo que produce ciertos

problemas al momento de realizar un análisis, así también en la aplicación de las técnicas de minería de datos. Éste inconveniente produce ruido en el banco de datos.

Uno de los casos más frecuentes en la integración de datos es al momento de unir identificadores, como por ejemplo, un solo individuo puede poseer varios identificadores según el país donde se encuentre, como en el caso de España el DNI, NIE, NIF, Pasaporte; en Ecuador la CI, RUC, número de la seguridad social, etc. Es por eso que se debe agrupar de una manera eficiente esta información sin que pueda existir daños al resto de los registros vinculados. Como se puede ver en la figura 9.

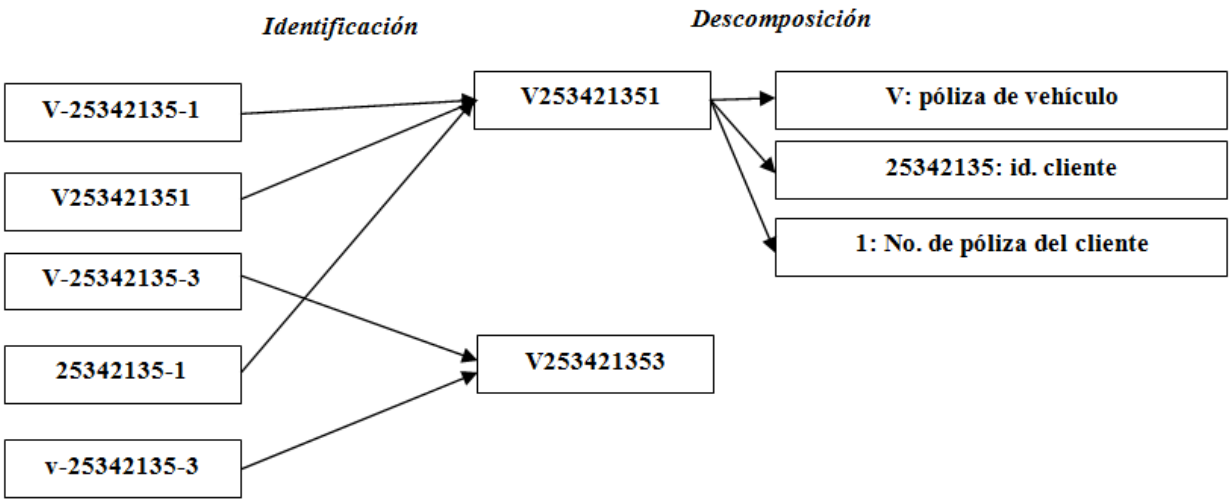

Figura 9: Ejemplos de integración: identificación y descomposición. Tomado de Hernández-Orallo[27]

En el caso de unificar varias fuentes de datos que posee registros faltantes o información incompleta, que demás esté relacionada entre sí, es un caso claro de inconsistencia de la información. La figura 10, muestra un ejemplo más claro este tipo de problema.

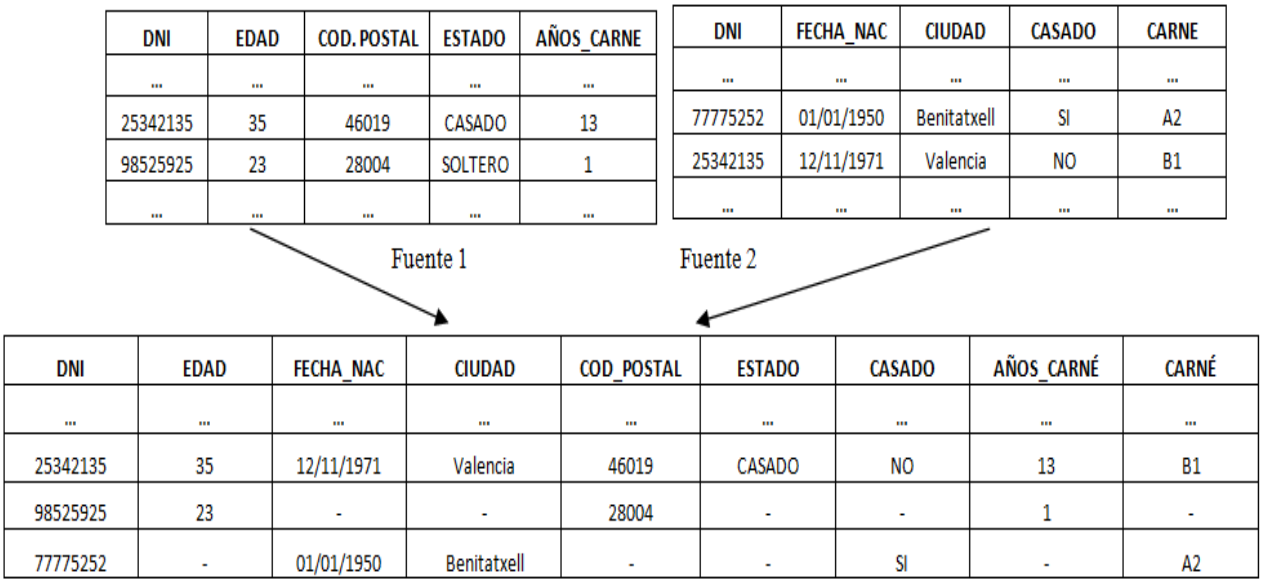

Figura 10: Ejemplos de integración de atributos de distintas fuentes. Tomado de Hernández-Orallo.[27]

Como indica en la figura anterior, existe información duplicada como también vacios en la información. La práctica normal, es dejar la información faltante y unificar los registros con duplicidad.

El caso más frecuente en la fusión de fuentes de información, donde se juntan los formatos, como fechas, estado civil, género, etc. Se puede ver de mejor manera en la figura 11.

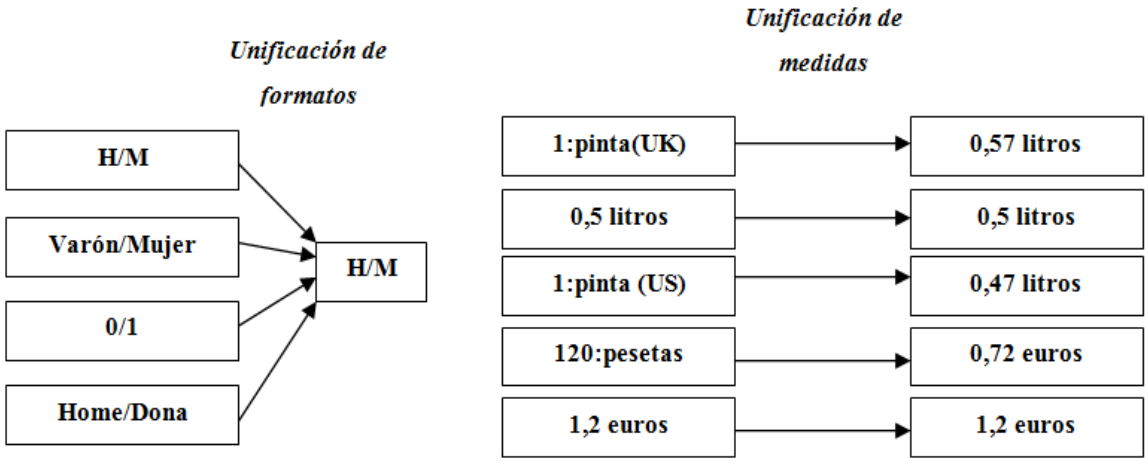

Figura 11: Ejemplos de integración: unificación de formatos y medidas. Tomado de Hernández-Orallo.[27]

En este proceso, cuando se tiene integrado todos los datos se realiza una tabla de resumen para visualizar características de los datos como máximos, mínimos, moda, etc. En la tabla 2, se puede ver un ejemplo de una tabla de resumen.

Atributo	Tabla	Tipo	#Total	#nulos	#dist	Media	Desv. E.	Moda	Min	Max
Código Postal	Cliente	Nominal	10320	150	1672	-	-	46003	01001	50312
Sexo	Cliente	Nominal	10320	23	6	-	-	V	E	M
Estado Civil	Cliente	Nominal	10320	317	8	-		Casado	Casado	Viudo
Edad	Cliente	Númerico	10320	4	66	42.3	12.5	37	18	87
Total póliza p/a	Póliza	Númerico	17523	1325	142	737.24 €	327.00 €	680.00 €	375.00 €	6,200.00 €
Asegurados	Póliza	Númerico	17523	0	7	1.31	0.25	1	0	10
Matrícula	Vehículo	Nonimal	16324	0	16324	-	-	-	A-0003-BF	Z-9835-AF
Modelo	Vehículo	Nonimal	16324	1321	2429	-	-	O. Astra	Audi A3	VW Polo
…	…	…	…	…	…	…	…	…	…	…

Tabla 2: Tabla de resumen de atributos. Tomado de Hernández-Orallo.[27]

Para un tipo especial de atributos que son muy utilizados como son las fechas y horas, que no proporcionan mucha información si están con distintos formatos al momento de la unificación de repositorios, por lo cual, se prefiere realizar un proceso de transformación de un solo atributo fecha y hora a varios, o también, crear una sola dimensión *"tiempo"*. Este proceso varía según la necesidad de cada uno de los objetivos que se requieren en la minería de datos. En este caso, se va a dividir la fecha en tres atributos: mes, día de la semana y festivo, como muestra en la figura 12.

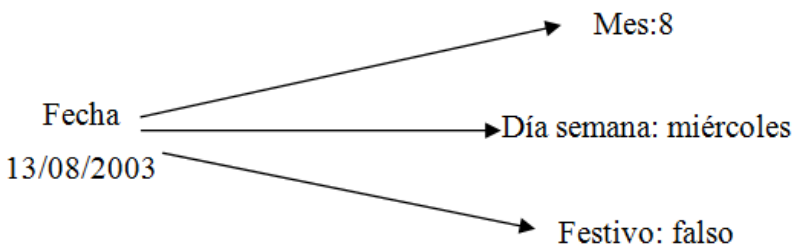

Tabla 3: Convirtiendo fechas en atributos más significativos. Tomado de Hernández-Orallo .[27]

Otra técnica para refinar la información recopilada es la discretización, que tiene por objetivo de intercambiar valores numéricos en valores nominales numerados y ordenados. Existen varios ejemplos prácticos donde se utiliza la discretización, como por ejemplo, en las notas obtenidas por los alumnos en universidades españolas, en la siguiente tabla 4, presenta como funciona el proceso.

Calificación 0 a 10
matrícula de honor (l0)
sobresaliente (8,5:9,99)
notable (7:8,49)
aprobado (5:6,99)
suspenso (0:4,99)

Tabla 4: Calificaciones de educación superior española.

Este proceso es muy apropiado para nombrar a periodos de valores con mucha información o valores numéricos que poseen decimales.

Como se ha visto, el proceso de transformación y limpieza de los datos se puede realizar de varias formas, hay que tener en cuenta la información que se quiere obtener y así utilizar todas la herramientas, técnicas y metodologías existentes para sacar un mejor partido de la información.

Por otra parte existen herramientas libres que pueden ayudar a estas tareas, como es weka y otras de pago como matlab, clementine, etc.

2.4. Exploración y análisis de datos

Al llegar a esta fase se puede tener la mayor parte de la información integrada, limpia y formalizada, pero aún así, en ciertos casos se debe realizar ciertos procedimientos para alcanzar a tener una materia prima para aplicar metodologías de minería de datos. Es por eso en esta fase

se utilizaran varias técnicas para visualizar y seleccionar variables, que puedan ayudar a la obtención del resultado esperado, una "*vista minable*", y así poder extraer información de calidad y efectiva para el presente estudio.

Para entender de mejor manera el termino de "*vista minable*" se podrá definir de la siguiente manera:

Una vista minable es la selección de variables relevantes o que aportan información importante para que posteriormente se pueda aplicar técnicas de minería de datos. Esta vista evita la pérdida de calidad del modelo de conocimiento obtenido del proceso de minería.[27]

Para realizar un estudio de los datos importantes que se debe extraer del repositorio es necesario saber el objetivo principal para el cual se extrae la información, para eso, se debe realizar las siguientes preguntas:

- ¿Qué parte de los datos es pertinente analizar?
- ¿Qué tipo de conocimiento se desea extraer y cómo se debe presentar?
- ¿Qué conocimiento puede ser válido, novedoso e interesante?
- ¿Qué conocimiento previo me hace falta para realizar esta tarea?

Solo teniendo el conocimiento de ¿qué es lo importante de los datos?, se podrá seleccionar atributos relevantes que pueden proporcionar información importante. Las cuatro preguntas anteriores son, en realidad, una manera de clasificar al conjunto de datos que se podrían utilizar, ya que, en el fondo, son preguntas que están interrelacionadas.[27, 9]. En la siguiente figura 12, se muestra de mejor manera el proceso de obtener una vista minable.

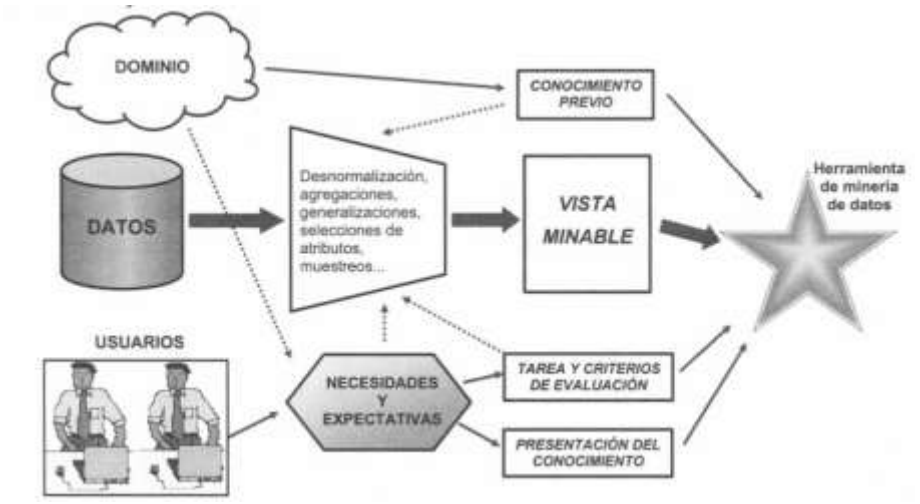

Figura 12: Diagrama de datos, dominio y usuarios a la vista minable y elementos asociados. Tomado de Hernández-Orallo. [27]

En la figura 12, presenta el proceso de obtención de una vista minable, donde es necesario tener como base un conocimiento previo de lo que se desea obtener como resultado de la información, así también como las tareas que se deben realizar para conseguir este objetivo, esas tareas son:

- desnormalización,
- agregación,
- generaciones,
- selecciones de atributos,
- muestreos, etc.

Además, se debe tener en cuenta las necesidades y expectativas que el usuario tiene dentro de este proceso de selección y exploración de la información.

Como se puede apreciar, no basta con obtener una vista minable, sino también, que va acompañada de una tarea de evaluación y selección. En la figura 13, se presenta un esquema general de las tareas que se van a realizar utilizando las preguntas que se plantearon con anterioridad.

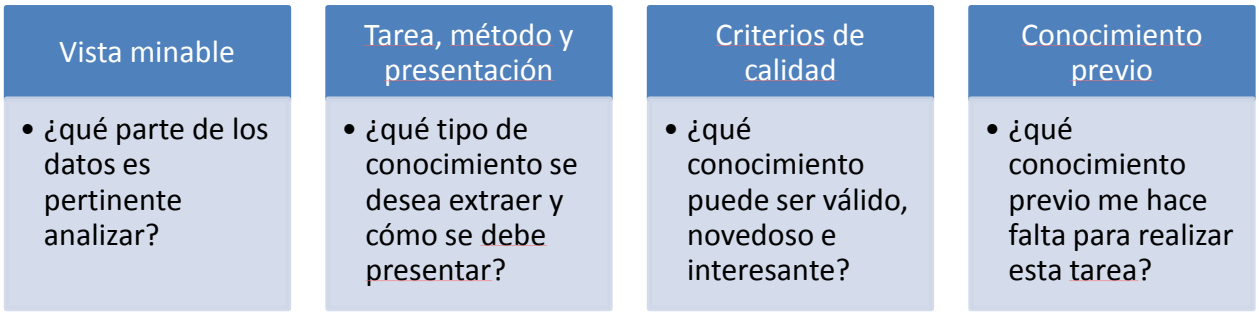

Figura 13: Proceso de selección y exploración de la información

Vista minable: Recoge la información imprescindible para realizar la minería de datos.

Tarea, método y presentación: Indica que métodos se van a utilizar como regresión, clasificación, agrupamiento, etc. Así también, la información de entrada y de salidas para la aplicación de técnicas de extracción de conocimiento como redes neuronales, árboles de decisión, regresión logística y demás.

Criterios de Calidad: En muchos casos hay que aplicar criterios para comprobar la eficiencia de los datos resultantes, se pueden aplicar criterios de comprensibilidad, fiabilidad, confianza, utilidad, interés o novedad.

Conocimiento previo: Se aplica el conocimiento que se ha obtenido en procesos anteriores para extraer la información por medio de OLAP, las dimensiones creadas y agregación de nuevos atributos en el caso de que sea necesario.

2.4.1. Exploración mediante la visualización

En esta tarea se va a realizar gráficas de los atributos para detectar patrones, y posibles técnicas que se puedan aplicar. En sí, lo que se va a utilizar es una técnica de exploración llamada "*minería de datos visual*" (***visual data mining***) [27], lo que permite por medio de graficas de uno o varios atributos a la vez, es detectar patrones de la información de una forma visual por el usuario. Es por eso que los objetivos principales de esta técnica son:

- Aprovechar la gran capacidad humana de ver patrones, anomalías y tendencias a partir de imágenes y facilitar la comprensión de los datos[27, 38].

- Ayudar al usuario a comprender más rápidamente patrones descubiertos automáticamente por un sistema de KDD[23, 38].

Para utilizar esta técnica de descubrimiento de patrones se puede clasificar en dos tipos de momentos de visualización.

- **Visualización previa:** Se utiliza para detectar patrones y posibles herramientas que se pueden aplicar, además, se encarga de presentar resúmenes para determinar información que se debe investigar.[23]
- **Visualización posterior:** Se detecta patrones y se los estudia de una mejor manera. Por otra parte se visualiza y se valida los resultados obtenidos en el proceso de extracción del conocimiento.[23]

Como se presenta en la figura 14, la distribución de variables en un periodo determinado, obtenido de una muestra definida.[8, 23]

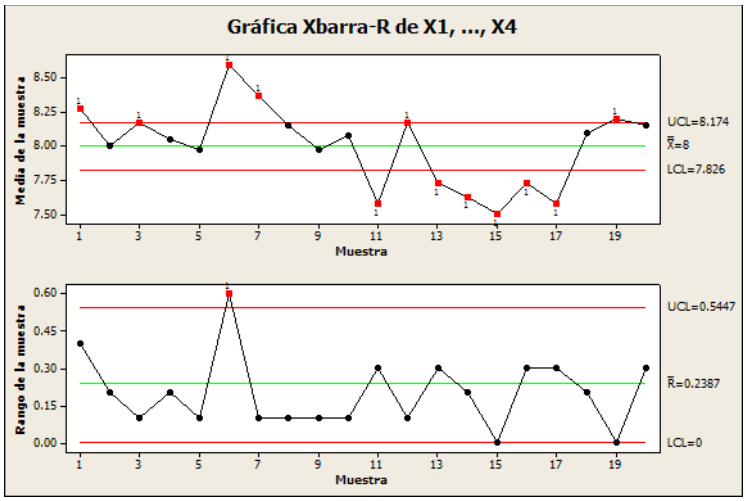

Figura 14: Grafica de variables con una muestra determinada. Tomado de Martín-Pliego.[24]

2.4.2. Selección de datos

En la fase de selección de datos, se determinan que atributos son importantes y deben ser introducidos en la vista minable para la aplicación de la minería de datos. La tarea de la selección

no es necesariamente la reducción del tamaño de los datos, sino mejorar el resultado de la información.

Para realizar esta selección se procederá a realizar técnicas de muestreo para determinar el grupo de datos eficiente para el estudio.

Técnicas de muestreo

La mejor manera de seleccionar un conjunto de datos óptimo para el estudio es realizar un muestro. Existen varias técnicas estadísticas que se basan en una población, conjunto o subconjunto de datos.[24, 36]

En la minería de datos se plantean dos casos que depende de la población de la información:

- **Se dispone de la población:** Se determina que datos son imprescindibles y no es recomendable una muestra aleatoria.
- **Datos de realidad:** Son datos recolectados en una base de datos y son una parte considerable de la realidad.

Por otra parte existen varios otros tipos de muestreos que se detallan a continuación:

- **Muestreo simple:** Consiste en extraer todos los individuos al azar de una lista (marco de la encuesta). En la práctica, a menos que se trate de poblaciones pequeñas o de estructura muy simple, es difícil de llevar a cabo de forma eficaz.[6, 27, 30]. Lo indica en la figura 15.

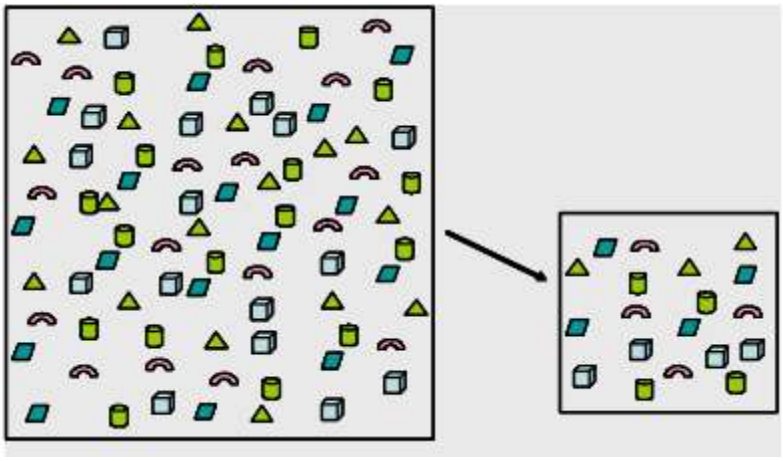

Figura 15: Representación gráfica del muestreo aleatorio simple. Tomado de Casal-Mateu.[5]

- **Muestreo sistemático:** En este caso se elige el primer individuo al azar y el resto viene condicionado por aquél. Este método es muy simple de aplicar en la práctica y tiene la ventaja de que no hace falta disponer de un marco de encuesta elaborado. Puede aplicarse en la mayoría de las situaciones, la única precaución que debe tenerse en cuenta es comprobar que la característica que estudiamos no tenga una periodicidad que coincida con la del muestreo.[6, 27], como lo indica en la figura 16.

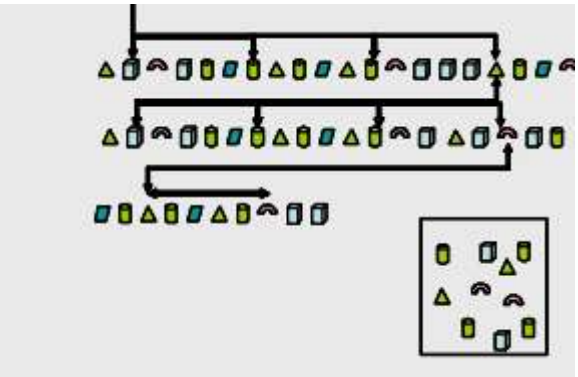

Figura 16. Representación gráfica del muestreo aleatorio sistemático. Tomado de Casal-Mateu.[5]

- **Muestreo aleatorio estratificado:** Se divide la población en grupos en función de un carácter determinado y después se muestrea cada grupo aleatoriamente, para obtener la parte proporcional de la muestra. Este método se aplica para evitar que por azar algún grupo de animales este menos representado que los otros.[6, 27], como se puede apreciar en la figura 17.

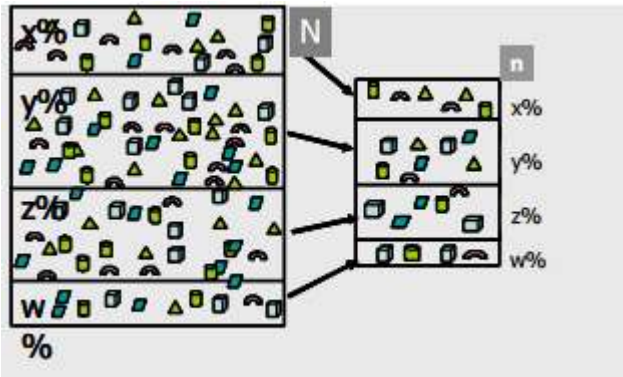

Figura 17: Representación gráfica del muestreo aleatorio estratificado. Tomado de Casal-Mateu.[5]

- **Muestreo aleatorio por conglomerados:** Se divide la población en varios grupos de características parecidas entre ellos y luego se analizan completamente algunos de los grupos, descartando los demás. Dentro de cada conglomerado existe una variación importante, pero los distintos conglomerados son parecidos. Requiere una muestra más grande, pero suele simplificar la recogida de muestras. Frecuentemente los conglomerados se aplican a zonas geográficas.[4,30], como presenta en la figura 19.

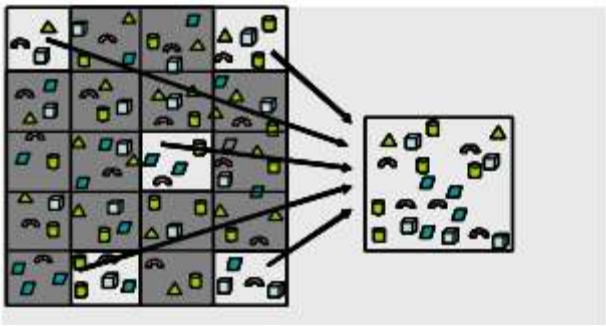

Figura 18: Representación gráfica del muestreo aleatorio por conglomerados. Tomado de Casal-Mateu.[5]

- **Muestreo mixto:** Cuando la población es compleja, cualquiera de los métodos descritos puede ser difícil de aplicar, en estos casos se aplica un muestreo mixto que combina dos o más de los anteriores sobre distintas unidades de la encuesta.[6, 27, 30]

Como se ha presentado, las formas de seleccionar un grupo de datos es muy variada y depende de cada caso para su aplicación, es por eso que se existen herramientas que se pueden utilizar para ayudar a la selección del conjunto de datos, como por ejemplo weka que permite realizar estas tareas de una forma automatizada y utilizando código SQL, que permite filtrar de mejor forma el conjunto de datos.

Técnicas de reconocimiento de patrones y minería de datos

Existen una variedad de técnicas para el reconocimiento de patrones a partir de un repositorio de datos, se ha visto la necesidad de no solo utilizar una sola sino varias para expresar con total claridad el conocimiento ingresado en el almacén de datos. Es por eso que se utilizará dos técnicas que son las más eficientes para detectar patrones de comportamiento de datos, las cuales son:

- Redes neuronales artificiales (RNA),
- Árboles de decisión.

Combinando varias técnicas se podrá analizar la información y tener un resultado más cercano a la realidad. Obteniendo una mejor calidad en los datos resultantes para el estudio.[25]

2.5. Redes neuronales Artificiales

Las redes de neuronas artificiales (RNA) son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. En inteligencia artificial es frecuente referirse a ellas como redes de neuronas o redes neuronales. [1]

2.5.1. Fundamentos biológicos de las redes neuronales.

Las neuronas son un tipo de células del sistema nervioso cuya principal característica es la excitabilidad eléctrica de su membrana plasmática; están especializadas en la recepción de estímulos y conducción del impulso nervioso (en forma de potencial de acción) entre ellas o con otros tipos celulares, como por ejemplo las fibras musculares de la placa motora. [1, 31]

Las neuronas presentan unas características morfológicas típicas que sustentan sus funciones: un cuerpo celular llamado soma o «pericarion», central; una o varias prolongaciones cortas que generalmente transmiten impulsos hacia el soma celular, denominadas dendritas; y una prolongación larga, denominada axón o «cilindroeje», que conduce los impulsos desde el soma hacia otra neurona u órgano diana. Como lo indica en la figura 19. [1, 18, 37]

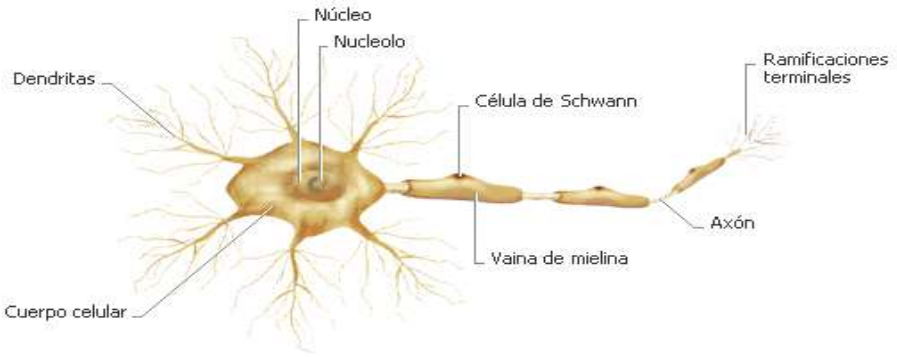

Figura 19: Descripción de una célula nerviosa típica. Tomado de Viñuela.[38]

2.5.2.Componentes de las neuronas

- **Dentritas:** conjunto de fibras en un extremo de una neurona que recibe mensajes de otras neuronas.
- **Axón:** parte de la neurona que transmite mensajes destinados a otras neuronas.
- **Botones terminales:** pequeñas protuberancias en el extremo de los axones que envían mensajes a otras neuronas. Los mensajes que viajan por la neurona son de naturaleza eléctrica.
- **Vaina de mielina:** evita que las neuronas entren en corto circuito, cubierta de grasa y proteínas que envuelve al axón. También contribuye en la velocidad de la transmisión del mensaje.[24,27].

Modelo computacional

Existen mucha diferencia entre las redes neuronales y los programas de computador, ya que no se trata de una aplicación que ejecuta un algoritmo definido, sino que en cierta medida la red neuronal procesa la información para obtener una salida o respuesta. Dicha información depende de varias características tanto estructurales y funcionales de la propia red para su procesamiento.

En la actualidad, existen varios modelos de redes de neuronas que siguen distintas filosofías de diseño, reglas de aprendizaje, y muchas muy variadas funciones de construcción de respuestas;

lo que genera una amplia gama de posibilidades al momento de la construcción de cada una de ellas.

A continuación se describirá el modelo computacional genérico que se utiliza para el desarrollo de diferentes sistemas de Redes Neuronales Artificiales.

2.5.3. La Neurona Artificial

La neurona artificial o también llamada célula o autómata, es un elemento que posee un estado interno, denominado nivel de activación, el cual recibe señales que permiten cambiar de estado.

Las neuronas tienen una función que les facilita cambiar de nivel de activación, lo realizan utilizando las señales que reciben; a dichas funciones se las conoce como función de activación y función de transición respectivamente, como se indicó anteriormente. Las señales recibidas por la neurona pueden provenir del exterior o de neuronas que están conectadas.

El nivel de activación de una neurona depende de las entradas recibidas y de valores sinápticos, pero, no de valores anteriores de estados de activación. Para conocer el valor del estado de activación de una neurona se calcula la entrada total de la célula, Ei. Este valor se calcula como la suma de todas las entradas pandeadas por ciertos valores. En la figura 20, presenta la idea de la unida típica de una neurona.[33]

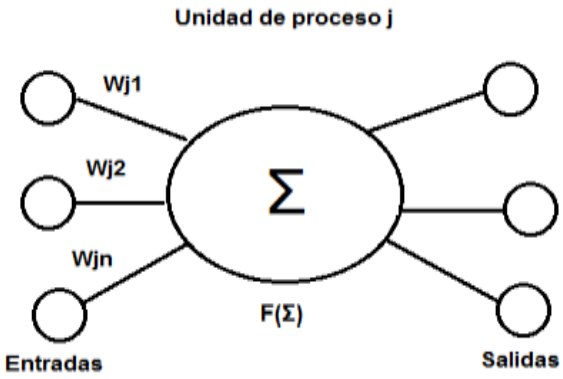

Figura 20. Esquema de una unidad de proceso típica. Tomado de Vinuela-Galván.[37]

Estructura de una red básica

En la figura 21, se muestra una unidad de red de neuronas artificiales, donde se puede apreciar en la parte superior una serie de entradas a la neurona; donde cada una de las entradas llegan a una salida de otra neurona de la red. Una vez que se ha calculado la salida de una neurona, como se presento anteriormente, se propaga vía conexiones de salida a las células destino. Donde todas las conexiones reciben el mismo valor de salida.[9, 33, 37]

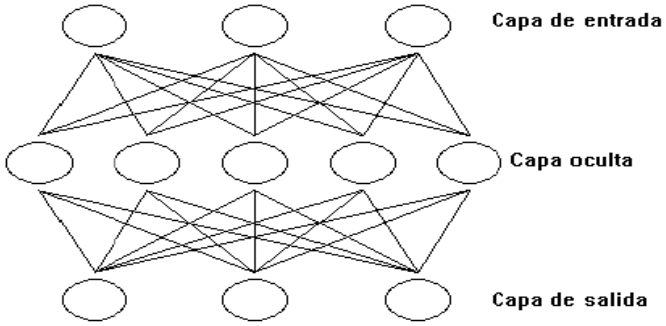

Figura 21: Esquema de una red de tres capas totalmente interconectadas. Tomado de Vinuela-Galván.[37]

Esta forma de conexión entre sí de las células se la denomina patrón de conectividad o arquitectura de red. Donde la arquitectura básica de una red multicapa es la figura 21. Presenta que en el primer nivel lo constituyen las células de entrada; estas unidades reciben valores de unos patrones que se representan como vectores que se utilizan para ingresarlos en la red. En la nivel intermedio, existen múltiples capas, las que dependen de rasgos particulares presentes en los patrones de entrada. Pueden existir, uno o varios niveles ocultos. Finalmente, el último nivel es el de salida, donde reciben los resultados de toda la red.

Cada interconexión entre unidades de proceso actúan como una ruta de comunicación, por medio de ellas viajan los valores numéricos de una célula a otra. Estos valores son evaluados por los pesos de las conexiones, donde se ajustan en el periodo de aprendizaje para producir la red neuronal artificial final.

Como se puede apreciar, la red neuronal tiene el aspecto de un grafo, el cual cada una de las unidades de proceso son idénticas y transmiten su información a través de arcos. Este grafo distingue nodos de entrada, salida e intermedios.

Funcionamiento de la red neuronal artificial

Su funcionamiento es realmente simple, para cada vector de entrada, éste introducido por la red copiando cada valor de dicho vector en la célula de entrada correspondiente. Cada célula de la red, una vez recibida la totalidad de sus entradas, procesa y genera una salida que es propagada a través de las conexiones entre las células, llegando como entrada a la célula destino. Una vez que la entrada ha sido completamente propagada por toda la red, se producirá un vector de salida, cuyos componentes son cada uno de los valores de salida de las células de salida.[1,2,18, 27]

2.5.4. Aprendizaje de la red neuronal artificial

La parte fundamental de las redes neuronales artificiales es el aprendizaje, ya que el esquema de como aprende la red determina el tipo de problemas que puede solucionar, además, las redes neuronales son sistemas que se basan en ejemplos, ya que según el tipo de ejemplos ingresados se procesara la información para que el sistema pueda aprender. El punto de vista de los ejemplos, el conjunto de aprendizaje presenta las siguientes características:

- Mediante un número fijo de ciclos. Se define a priori cuantas veces será introducido todo el conjunto, y una vez superado el presente número se detiene el proceso y se da como construida la red resultante.
- Cuando el error descienda por debajo de una cantidad preestablecida, habrá que establecer una función de error a nivel de patrón individual o a nivel de la totalidad del conjunto de entrenamiento. Posteriormente, ingresar un criterio adicional de parada para no bajar por menos del nivel prefijado de error, que indicará el número de ciclos que deben realizar hasta conseguir el modelo de red deseado y sea la solución óptima al problema.
- Cuando la modificación de pesos sea irrelevante. En algunos modelos se define un esquema de aprendizaje que hace que las conexiones vayan modificándose cada vez con menor intensidad. El proceso de aprendizaje continúa, no se producirán variaciones a los valores de los pesos de ninguna conexión, en ese momento se dice que la red ha terminado el proceso de aprendizaje.

Esquemas de aprendizaje de las redes neuronales

Existen varios esquemas de aprendizaje y del problema a solucionar, es por eso que se presentan tres tipos de esquemas de aprendizaje:

- Aprendizaje supervisado (Entrada-Salida-Objetivo)
- Aprendizaje no supervisado (Entrada-Salida)
- Aprendizaje por reforzamiento (Recompensa/castigo)

Para el presente estudio solo se presentará el aprendizaje supervisado ya que es el que se utilizará en la etapa de implementación.

Aprendizaje Supervisado

El aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento son conjunto de objetos en pares, normalmente estos pares de objetos son vectores, que se componen de un par de datos de entrada y otros de salida llamados resultados deseados) La salida de la función puede ser un valor numérico o una etiqueta de clase (en el caso de la clasificación). El objetivo del aprendizaje supervisado es el de crear una función capaz de predecir el valor correspondiente a cualquier objeto de entrada válida después de haber visto una serie de ejemplos, los datos de entrenamiento. Para ello, tiene que generalizar a partir de los datos presentados a las situaciones no vistas previamente. [27].

Las algoritmos del aprendizaje supervisado son:

- Aprendizaje por Corrección de error.
- Aprendizaje por Refuerzo.
- Aprendizaje Estocástico.

Aprendizaje por Corrección de error

Consiste en ajustar los pesos de la conexiones de la red en función de la diferencia entre los valores deseados y obtenidos en la salida de la red en función del error cometido.

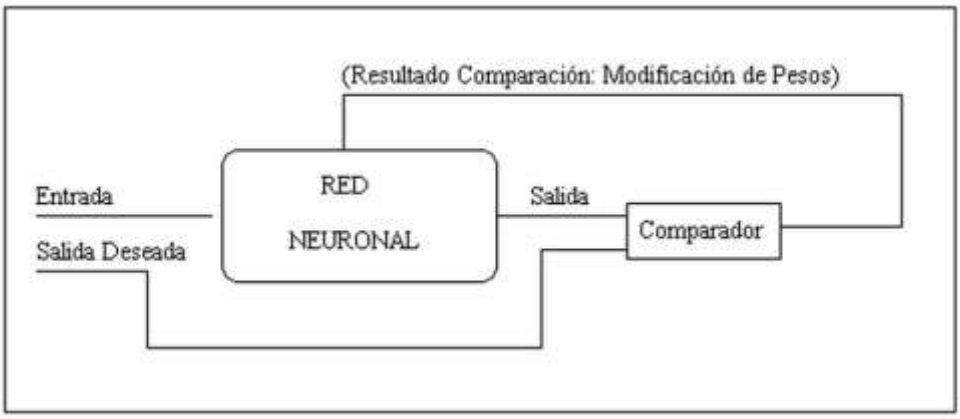

Figura 22: Proceso de algoritmos de aprendizaje de corrección de error. Tomado de Vinuela-Galván.[37]

El objetivo de este algoritmo de aprendizaje, que se muestra en la figura 22, es de minimizar el error entre la salida deseada y la actual, además de ser una aprendizaje fuera de línea. Se presenta el proceso para el aprendizaje:

1. Inicializar aleatoriamente los pesos
2. Presentación del conjunto de entrenamiento (CE)
3. Obtención de las salidas para el CE
4. Comparación de salidas deseadas con actuales.
5. Si se verifica el criterio de finalización ir al siguiente paso, sino ir al paso 2.
6. Finalizar el proceso.

Aprendizaje por Refuerzo

Este tipo de aprendizaje es más lento que el expuesto anteriormente. No se dispone de un ejemplo completo del comportamiento deseado. No se conoce la salida deseada exacta para cada entrada. Se conoce como debería de ser el comportamiento de manera general ante diferentes entradas. Es un aprendizaje en línea. Relación entre entrada y salida a través de un proceso de éxito o fracaso, produciendo una señal (Señal de Refuerzo) que mide el buen funcionamiento del sistema.

Esta "señal de refuerzo" está caracterizada por el hecho de que es menos informativa que en el caso de aprendizaje supervisado mediante ejemplos.

Aprendizaje Estocástico

Este tipo de aprendizaje consiste básicamente en realizar cambios aleatorios en los valores de los pesos y evaluar su efecto a partir del objetivo deseado y de distribuciones de probabilidad.

Símil: Red Neuronal ------- Sólido Físico (Estados Energéticos)

Estado de mínima energía: Valores de Pesos con los que la estructura se ajusta al objetivo deseado. El proceso es el siguiente:

1. Se realiza un cambio aleatorio en los Pesos.
2. Se determina la nueva energía de la red
3. Si la energía no decrece: se aceptaría el cambio en función de una determinada y preestablecida distribución de probabilidades.[1]

2.6. Árboles de decisión

En esta técnica se centra más en métodos para el aprendizaje de modelos comprensibles, que se basan en sistemas de reglas. Este método es uno de los más fáciles de utilizar y entender, ya que está organizado de una manera jerárquica, donde la decisión final a tomar se puede determinar seleccionando condiciones que se cumplan desde el nodo raíz del árbol hasta alguna de las hojas, simulando el razonamiento normal del ser humano.[17]

Unas de las ventajas de los árboles es la decisión es dependiendo de las opciones posibles a partir de una determinada condición son excluyentes. Lo que permite analizar una situación y siguiendo el camino del árbol adecuadamente, llegar a una sola acción o decisión a tomar.

2.6.1. Árboles de decisión para clasificación

La tarea en la cual los árboles de decisión se comporta mejor con clasificación, ya que clasificar es determinar de entre varias clases a qué clase pertenece un objeto; la estructura de la condición y la ramificación de un árbol de decisión, es perfecta para solucionar este problema.[9]

La característica más importante del problema de clasificar es que asume que las clases son disjuntas, es decir, diferentes entre ellas. Esta propiedad es exhaustiva, lo que quiere decir es que una de las dos condiciones se debe cumplir, además, esto dio lugar al esquema básico de los primeros algoritmos de aprendizaje de árboles de decisión, estos algoritmos se llaman de partición o algoritmos de "*divide y vencerás*". [17, 27]. El esquema del algoritmo es el siguiente:

ALGORITMO Partición(N:nodoE,:conjuntodeejemplos)
SI todos los ejemplos E son de la misma clase e **ENTONCES**
Asignar la clase e al nodo **N**.
SALIR; // Esta rama es pura, ya no hay que seguir partiendo. N es hoja.
SI NO:
Particiones:=generar posibles particiones.
Mejor Partición:= seleccionarla mejor partición según el criterio de partición.
PARA CADA condición i de la partición elegida.
Añadir un nodo hijo i a N y asignar los ejemplos consistentes a cada hijo(E).
Partición(i,E). // Realizar el mismo procedimiento global con cada hijo.
FIN PARA
FIN SI
FIN ALGORITMO
Para clasificar un conjunto de ejemplos E, se invoca con la llamada Partición(R,E),
Donde R es un nodo raíz de un árbol por empezar.

Como se pudo observar, los dos puntos principales para el algoritmo anterior tenga un funcionamiento adecuado son:

- Particiones a considerar
- Criterio de selección de particiones.

Estos criterios son los que diferencian entre sí, a los algoritmos de partición existente, como son CART, ID3, C4.5, ASSISTANT, etc. Se pueden mencionar las siguientes ventajas de los árboles de clasificación y/o regresión [17, 31, 33]:

1. Se obtiene conocimiento estructurado en forma de reglas de clasificación o de los valores de una variable de intervalo. Esto facilita interpretar en un lenguaje llano la caracterización de las clases o los valores de una variable de intervalo.
2. Al ser un procedimiento de análisis no paramétrico (*distribution free procedure*) no se requiere validar supuestos distribucionales de probabilidad.
3. Permite trabajar con todo tipo de variables predictoras: binarias, nominales, ordinales y de intervalo o razón.
4. Permite valores desconocidos para las variables predictoras en los individuos, tanto en la fase de construcción del árbol como en la de predicción.
5. En el caso de Clasificación se puede establecer probabilidad a priori de las clases.
6. Se puede ponderar las observaciones usando una variable ad-hoc.[3, 17,35].

2.6.2. Construcción del Árbol

Un árbol es un conjunto de nodos y arcos. Cada uno de los nodos es un parte o subconjunto de toda una población de datos. Dentro del árbol se muestra el nodo raíz, el cual representa toda la población y no tiene arcos entrantes. Nodos terminales, son los que representan la partición final. Nodos intermedios cuyos arcos salientes apuntan a los nodos hijos.

El algoritmos de clasificación traduce la información en un diagrama recursivo que se traduce en los siguientes pasos:

1. El nodo raíz es dividido en subgrupos (dos o más) determinados por la partición de una variable predictora elegida, generando nodos hijos.
2. Los nodos hijos son divididos usando la partición de una nueva variable. El proceso recursivo se repite para los nuevos nodos hijos sucesivamente hasta que se cumpla alguna condición de parada.
3. Algunos de los nodos resultantes son terminales, mientras que otros nodos continúan dividiéndose hasta llegar a un nodo terminal.
4. En cada árbol se cumple la propiedad de tener un camino único entre el nodo raíz y cada uno de los demás nodos del árbol.

Se presenta el siguiente ejemplo de 215 pacientes que sufrieron un ataque al corazón, donde se evaluaron variables socio demográficas, historia médica, y exámenes de laboratorio. A los 30 días 37 pacientes murieron. Se presenta el Árbol de Clasificación desarrollado con el fin de estimar "El riesgo de un segundo ataque". [12, 29]. Como se presenta en la siguiente figura 23:

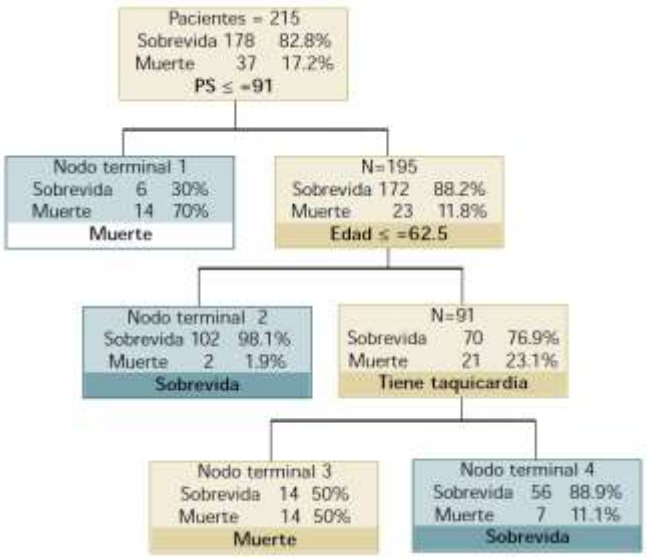

Figura 23: Ejemplo de árbol de clasificación. Tomado de Broadley-Utgoff.[3].

En el proceso recursivo descrito se deben establecer algunos criterios:

1. Cómo son los cortes posibles y un número máximo de cortes determinados por un predictor desde el nodo. Los cortes que se establecen para variables ordinales y de intervalo se realizan por intervalos consecutivos.

2. Una condición de admisibilidad para los cortes posibles.

3. Una medida de contenido de información del árbol respecto al conjunto de individuos o un criterio de optimización de los cortes; es decir, obtener la mejor combinación de cortes admisibles respecto a una variable predictora.

4. Determinar la descripción de la variable objetivo en los nodos del árbol.

Para clasificación: El grupo con la mayor representación determina la clase a la que asigna el nodo. En caso de empates se puede elegir cualquiera.

Para regresión: En los nodos se estiman las medias muestrales de la variable respuesta condicionadas a los nodos.

5. Una condición de parada para un nodo de un árbol. Por ejemplo, si el número de individuos en el nodo es inferior a un valor pre especificado, si la contribución del nodo a la calidad del árbol es mayor que otro umbral, si la profundidad del nodo es igual a un parámetro pre-especificado.[6, 35].

El criterio más importante en la construcción del árbol es la elección de una medida de contenido de información del árbol con respecto a las clases o variable de intervalo de interés ya que la elección de este criterio diferencia los algoritmos de selección. Medida de contenido de la información. Es la suma ponderada de una medida de contenido de la información H(.) de las hojas del árbol. H(.) es una función de incertidumbre o entropía aplicada a una distribución de probabilidad. Al ser ésta una medida aditiva en los nodos, en un paso del algoritmo es suficiente con optimizar el incremento de la medida de contenido de información del árbol en el nodo que se está explorando. En este caso, se obtiene la combinación de cortes que hace máxima la reducción de la incertidumbre en los nodos del árbol. Se han propuesto distintas definiciones de H(.), entre ellas [6, 17]:

- Entropía de Shannon
- Índice de diversidad de Gini

2.6.3. Particiones posibles

En este apartado se va a presentar las particiones que se van a considerar. Como se dijo anteriormente las particiones son conjunto de datos que son exhaustivas y excluyentes. Principalmente, mientras más tipos de condiciones se permitan, más posibilidades de encontrar los patrones que hay detrás de los datos. Pero mientras más particiones se trabajé más complejo será el algoritmo, es por eso el reto para el investigador es encontrar un equilibro entre las particiones que puedan ser eficientes y puedan mostrar la información de una forma expresiva.

Los tipos de particiones con las que trabajan los algoritmos de clasificación son la siguientes:

- Particiones nominales: si un atributo Xi es nominal, y tiene posibles valores {VI,V2,...,Vk}, sólo existirá una partición posible para dicho atributo y dicha partición será (Xi= VI,Xi = V2,...,Xi = Vk), es decir, una condición con la igualdad entre el atributo y cada posible valor. Muchos algoritmo s siguen esta partición, mientras otros exigen que los árboles sean binarios (sólo dos hijos por nodo) y, por tanto, que las particiones sean binarias (sólo dos condiciones). Para ello, consideran k particiones del estilo (Xi=VI, Xi *-VI), (Xi=V2, Xi < >V2),etc. Nótese que las dos variantes permiten obtener prácticamente los mismos árboles de decisión (equivalentes, al fin y al cabo).
- Particiones numéricas: si un atributo Xi es numérico y continuo, puede haber tomado muchos valores diferentes en los ejemplos y puede tomar infinitos posibles valores en general. Por esta razón, se intentan obtener particiones que separen los ejemplos en intervalos. Para ello, las particiones numéricas admitidas son de la forma (Xi<=a, Xi> a) donde a es una constante numérica elegida entre un conjunto finito de constantes que discriminen los ejemplos vistos. Por ejemplo, si tenemos diez ejemplos y en ellos aparecen los siguientes valores para el atributo Xi:{0,2, 0,3,0,7,0,1, 0,8, 0,45, 0,33, 0,1, 0,8, O}, muchos algoritmo s realizan el siguiente procedimiento: ordenan los valores (eliminando repetidos), es decir, {O,0,1, 0,2, 0,3, 0,33,0,45, 0,7, 0,8 }Y después obtienen el valor intermedio entre cada par de valores. Para el ejemplo anterior, tendríamos que nuestro conjunto finito de constantes sería: {0,05, 0,15, 0,25, 0,315, 0,39, 0,575, 0,75 }.Con estos siete valores, tenemos, por tanto, siete particiones posibles: (Xi<=0,05,Xi > 0,05), (Xi<=0,15, Xi > 0,15), (Xi<=0,25,Xi > 0,25), etc. Generalmente, si existen muchos ejemplos se suele seleccionar un subconjunto de los valores anteriores (mediante análisis de densidad u otras técnicas, véase por ejemplo[3, 17, 27]), con el objetivo de reducir el número de particiones posibles.

La forma de expresar las particiones descritas anteriormente tiene el nombre de expresividad proporcional cuadricular. Esta definición indica que son particiones que afectan solamente a un atributo de un ejemplo cada vez.

2.6.4. Agrupamiento

Aunque existen varias definiciones de la tarea de agrupamiento en todos los casos se plantea que la idea fundamental es formar grupos en los cuales se maximice la similitud entre los elementos que lo conforman y se minimice la similitud entre los distintos grupos. Es decir, se forman grupos tales que, los objetos que conformen un grupo sean muy similares entre sí y a la vez sean muy distintos a los objetos que formen otro grupo. Partiendo de esta idea se desarrollan las técnicas de agrupamiento.

Entre las clasificaciones de técnicas de agrupamiento planteadas en Marín-Palma [27,35] se encuentra:

Agrupamiento particional: El objetivo final del proceso de agrupamiento es obtener un conjunto de clases o grupos. Cuando todos los grupos que se obtienen son disjuntos cubriendo todo el conjunto de elementos se dice que el agrupamiento es particional.

Agrupamiento jerárquico: Cuando se obtiene una jerarquía de agrupamientos particionales "anidados", de manera tal que cada grupo de un nivel se divide en varios en el nivel siguiente, se dice que el agrupamiento es jerárquico.

El problema del agrupamiento particional se puede formalizar como sigue: "Dados n elementos representados en un espacio d-dimensional en el que hay definida una función de distancia, determinar una partición de los mismos en k subconjuntos o grupos, tales que los elementos incluidos en un grupo se parezcan más entre ellos de los que se parecen a los clasificados en otros grupos" [3, 17, 27, 35].

Las técnicas de agrupamiento jerárquico se basan en la construcción de un árbol en el que las hojas son los elementos del conjunto de ejemplos, y el resto de los nodos son subconjuntos de ejemplos que pueden ser utilizados como particiones (grupos) del espacio.

Capítulo 3 - Procesamiento de la información

Previo a la aplicación de las técnicas de identificación de patrones y clasificación, se debe precisar el origen de la información y realizar un proceso de selección de las variables que son más relevantes para el estudio.

La información se ha tomado de una base de datos de una entidad pública de la República del Ecuador, la que se posee múltiples usuarios que interactúan con un sistema y realizan una variedad de tareas dentro de esa base de datos.

En esta fase se integrará la información de las distintas bases de datos que están disponibles, para unificarlas en una sola y poder así realizar la tarea de seleccionar, limpiar y obtener la información necesaria para aplicar las técnicas de minería de datos. Como se muestra en las secciones siguientes, se detallan paso a paso todas las actividades que se realizan con los datos iniciales, para obtener un resultado óptimo en el posterior análisis.

3.1. Integración y recopilación

Para realizar la integración de las bases de datos de la entidad pública ecuatoriana, se han construido procedimientos de almacenamiento (*store procedures*), que permiten que la información de varias bases de datos se pueda almacenar en una sola llamada "tfmdata". Este diseño de una única base de datos facilitará al análisis de la información.

El diseño de la base de datos "tfmdata" será multidimensional, para que en cada dimensión se describa un conjunto de hechos con relevancia, como es el ¿cuándo?, ¿cómo?, ¿dónde?, etc. del acceso de cada usuario al sistema informático, obteniendo finalmente lo que se denomina OLAP (*Procesamiento Analítico en Línea*). Se utiliza un OLAP para centrar la atención en variables importantes, identificar excepciones, o encontrar iteraciones. Además, favorece la comprensión de los datos de una manera efectiva y ayuda al proceso de extracción del conocimiento.

Análisis de variables

El análisis de las variables debe ir acorde al comportamiento del usuario en la utilización del sistema, ya que describen sus actividades de una manera detallada. Por eso, se describirán las tareas en las que el usuario ha sido más recurrente y se analizarán una por una.

En la base de datos existen varios atributos que permitirán definir un patrón de comportamiento del usuario que ingresa al sistema informático. Esta información describirá las actividades que realiza dentro del sistema, como también fechas, horas, y la estación de trabajo que utiliza regularmente en su acceso. Los siguientes son los atributos que se pueden obtener de dicha base de datos, como se presenta en la tabla 5.

Nombre atributo	*Tipo de dato*	*Descripción*
Fecha de ingreso	*DateTime*	*Indica la fecha y hora en la cual el usuario ingresa al sistema.*
Ip	*Varchar*	*Proporciona la IP de la estación de trabajo en la cual realiza el ingreso al sistema.*
País	*Varchar*	*Muestra el país en el cual realiza el ingreso al sistema.*
Fecha de Actividad	*DateTime*	*Presenta la fecha y hora en la cual el usuario realiza tareas de Insertar, Actualizar y Eliminar en la base de datos.*
Tabla afectada	*Varchar*	*Muestra la tabla en la cual realizo interacciones con la base de datos.*
Operación	*Varchar*	*Proporciona las operaciones que realiza en la base de datos.*
Usuario	*Varchar*	*Presenta el usuario que realiza las actividades en la base de datos.*

Tabla 5: Tabla de atributos

Para este análisis se ha tomado una muestra de cada una de las actividades y tareas nombradas anteriormente, y se han creado gráficas que brindan una visualización de la información de una manera más simple y sencilla. La muestra es de un periodo de tres meses y de un solo usuario.

Variable "Operación"

Las operaciones realizadas dentro del sistema se describen de una manera general en la siguiente figura 24:

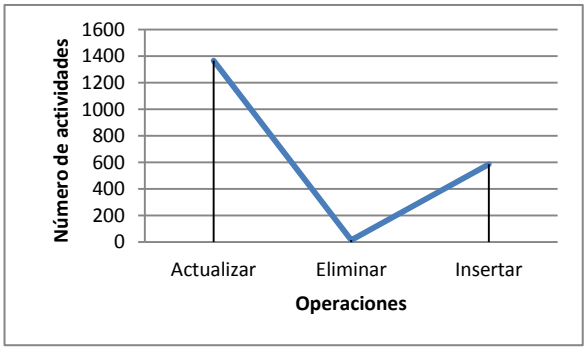

Figura 24: Operaciones

En las figuras 25 A, B y C, presentan las operaciones que un usuario concreto ha realizado sobre las bases de datos y su frecuencia temporal.

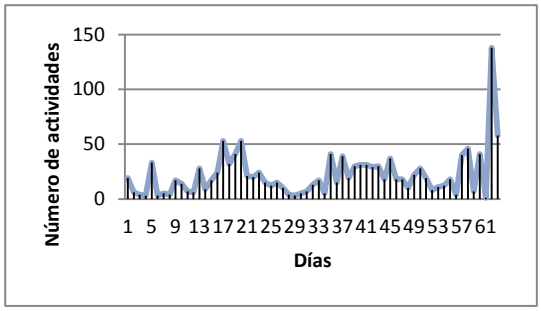

Figura A

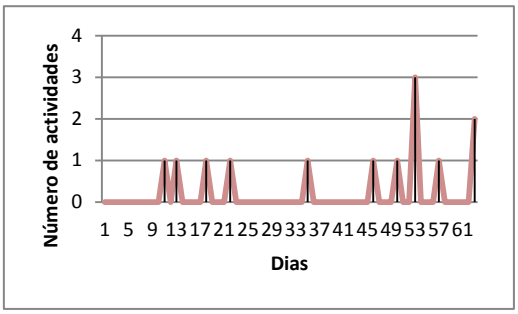

Figura B

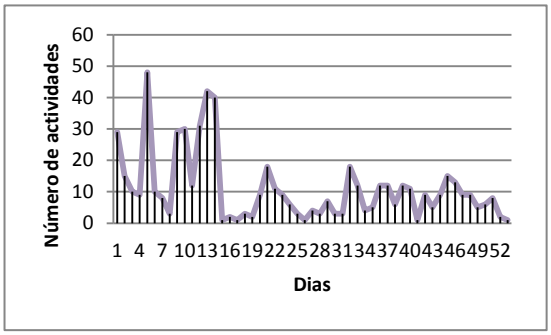

Figura C

Figura 25: Operación de Actualizar (Fig A), Eliminar (Fig B) e Insertar (Fig C)

Como se puede observar, la mayor tendencia de utilización de este usuario es realizar más operaciones de actualizar y de insertar que las de eliminar, pero no se descarta del estudio ésta última ya que la combinación de estas tres operaciones puede ser muy importante para un futuro patrón de comportamiento.

Variable "Tabla"

Las tablas en las que se realizan actividades dentro del sistema se describen de manera general en la siguiente Figura 26. Se han denominado Tabla 1, 2, 3 y 4 para mantener la confidencialidad de la información que está registrada en cada una de ellas, como se muestra en las figuras 27,28,29 y 30, respectivamente.

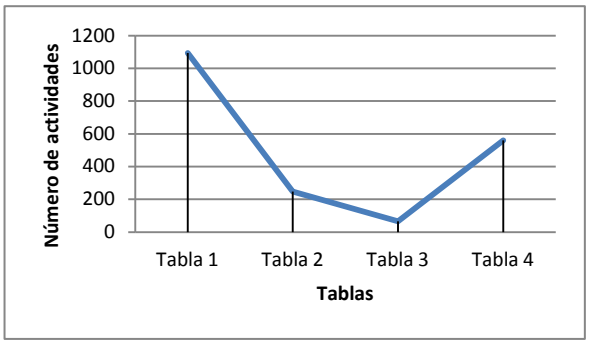

Figura 26: Gráfica de registro de información por tablas del usuario.

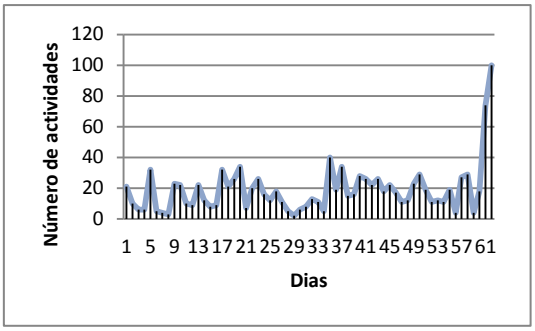

Figura 27: Tabla 1, 61 días

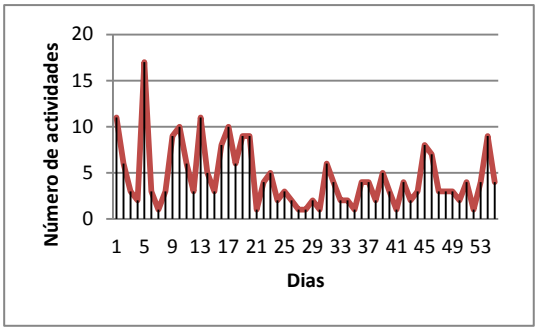

Figura 28: Tabla 2, 61 días.

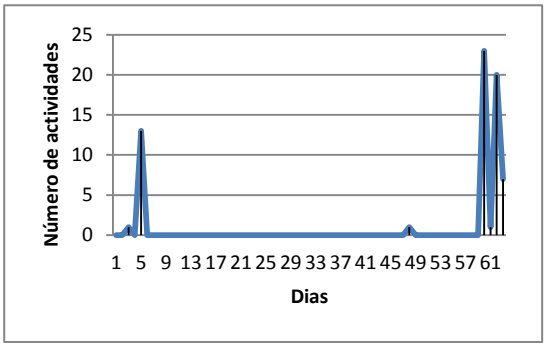

Figura 29: Tabla 3, 61 días

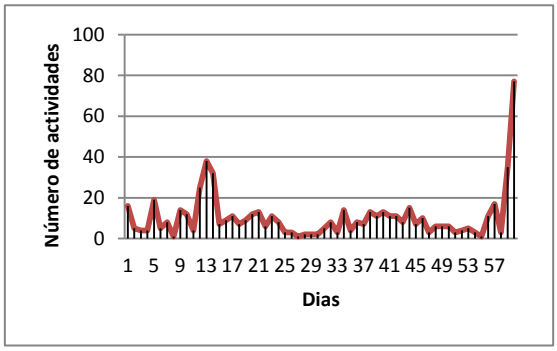

Figura 30: Tabla 3, 61 días.

En el análisis de las de actividades de las tablas por parte de un usuario se puede apreciar que en las tablas 1,2 y 4 el acceso es más frecuente; por otra parte, en la tabla 3 su actividad es casi nula, pero es un comportamiento que no deja de ser recurrente durante algunos días del mes.

Variable "Fecha"

Los días de la semana en las que se realizan operaciones dentro del sistema se representa en la figura 31.

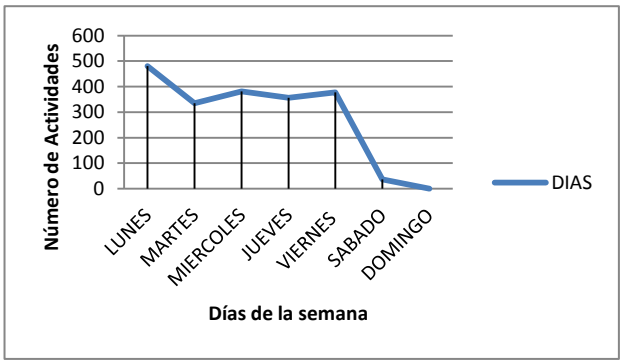

Figura 31: Número de accesos por día de la semana.

Variable "Hora"

De igual forma, en la Figura 32, se muestran los accesos en horas desde estaciones de trabajo desde las que se realizan las operaciones en el sistema.

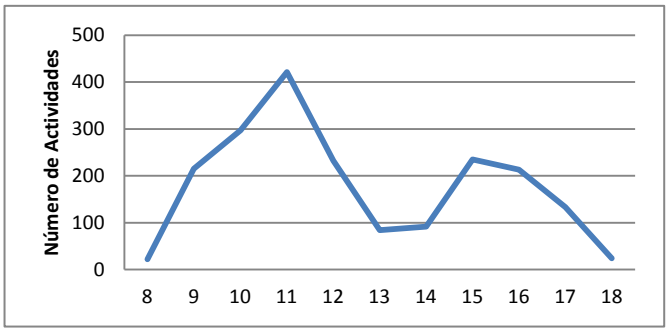

Figura 32: Número de accesos en horas a lo largo de un mes.

Variable "Estación de Trabajo"

La figura 33, representa la dirección IP de la estación de trabajo desde la que realiza operaciones dentro del sistema.

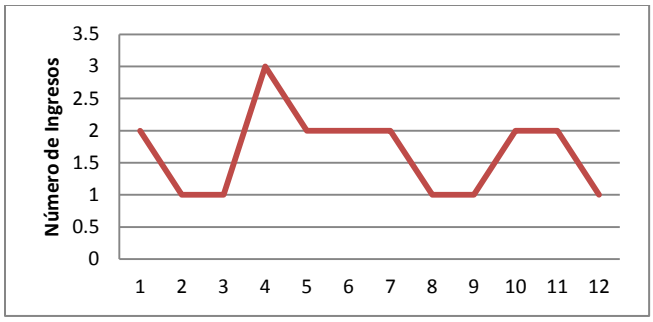

Figura 33: Número de accesos desde cada estación de trabajo a lo largo de un mes.

Esta representación ha permitido el análisis gráfico de las variables de interés, al menos de forma general. Después del análisis de la información de las anteriores figuras se han podido encontrar los hechos o registros del repositorio de datos (OLAP) que son más significativos para obtener un patrón relevante del comportamiento de cada usuario. El siguiente paso es realizar el diseño del repositorio de datos con los que son relevantes y el volcado de la información para la posterior etapa de selección, limpieza y transformación de los mismos.

3.2.Diseño del modelo multidimensional del repositorio de datos

Para generar el repositorio de datos debemos diseñar cada una de las variables o atributos a considerar, los cuales deberán estar relacionados entre sí, para un mejor entendimiento y análisis de la información.

El modelo del OLAP se divide en cuatro dimensiones o variables, que son:

1. dimensión tiempo,
2. dimensión usuario,
3. dimensión actividades,
4. dimensión estación de trabajo.

Cada una de las variables o atributos se han generado para que se pueda formalizar la información y así realizar todas las actividades de análisis con las diferentes técnicas de clasificación.

Dimensión "Tiempo"

Para la creación de la dimensión tiempo en repositorio de datos se ha dividido la variable fecha en varias tablas, que son: año, trimestre, mes, semana, día, y hora, con el fin de extraer y visualizar la información de mejor manera.

El diagrama es el siguiente (Figura 34):

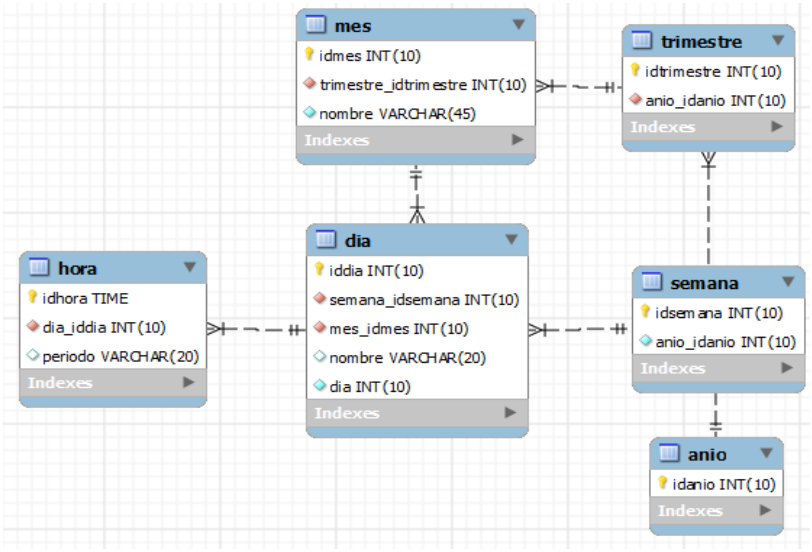

Figura 34: Diagrama de la dimensión tiempo en el repositorio de datos OLAP

Dimensión "Usuario"

La dimensión o atributo usuario se ha subdividido en varias tablas, para describir el comportamiento de un grupo de usuarios específicos o de un usuario individual. Para ello se han creado tres tablas, las cuales son: cargo, usuario, y departamento. El diagrama es el siguiente (Figura 35):

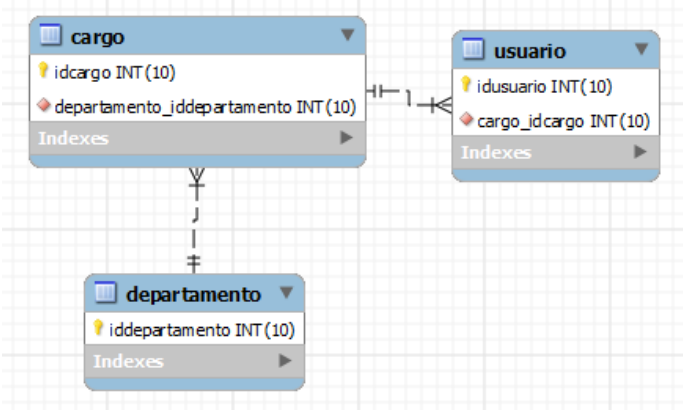

Figura 35: Diagrama de la dimensión usuario del repositorio de datos OLAP

En estas tablas no se han especificado nombres de los cargos, departamentos y usuarios para conservar la confidencialidad de la información que ha proporcionado la entidad colaboradora.

Dimensión "Actividades"

Esta parte del OLAP permite describir las operaciones y detalla las tablas en las que trabaja el usuario al manejar el sistema. Esta variable contiene dos tablas que son: operaciones y actividades (Figura 36).

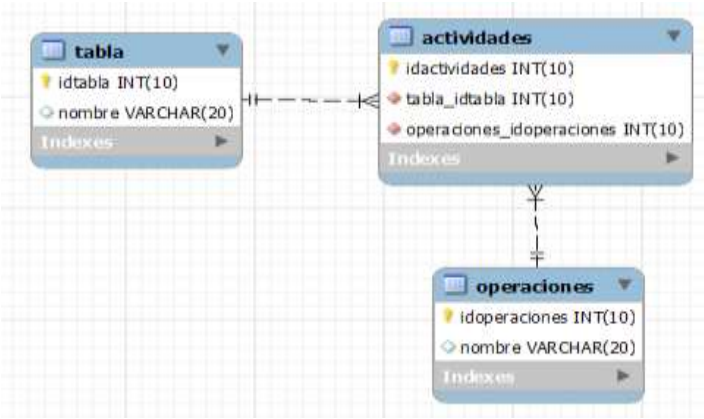

Figura 36: Diagrama de la dimensión actividades del repositorio de datos OLAP.

Dimensión "Estación del trabajo"

Este atributo es uno de los más simples debido a que consiste en una sola tabla con los datos de la IP del dispositivo de conexión que utiliza el usuario. El diagrama es el siguiente (Figura 37):

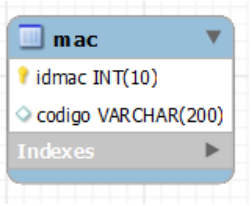

Figura 37: Diagrama de la dimensión estación de trabajo del repositorio de datos OLAP

Finalmente, para unir toda la información de las variables del repositorio debe existir una tabla que pueda relacionar todas ellas. Se ha creado la tabla de hechos (*fact table*), para unirlas y así tener un único OLAP. La tabla de hechos es la "*accesousuario*", que unifica las claves primarias de las demás dimensiones. En la siguiente figura 38, se puede apreciar el modelo final del OLAP.

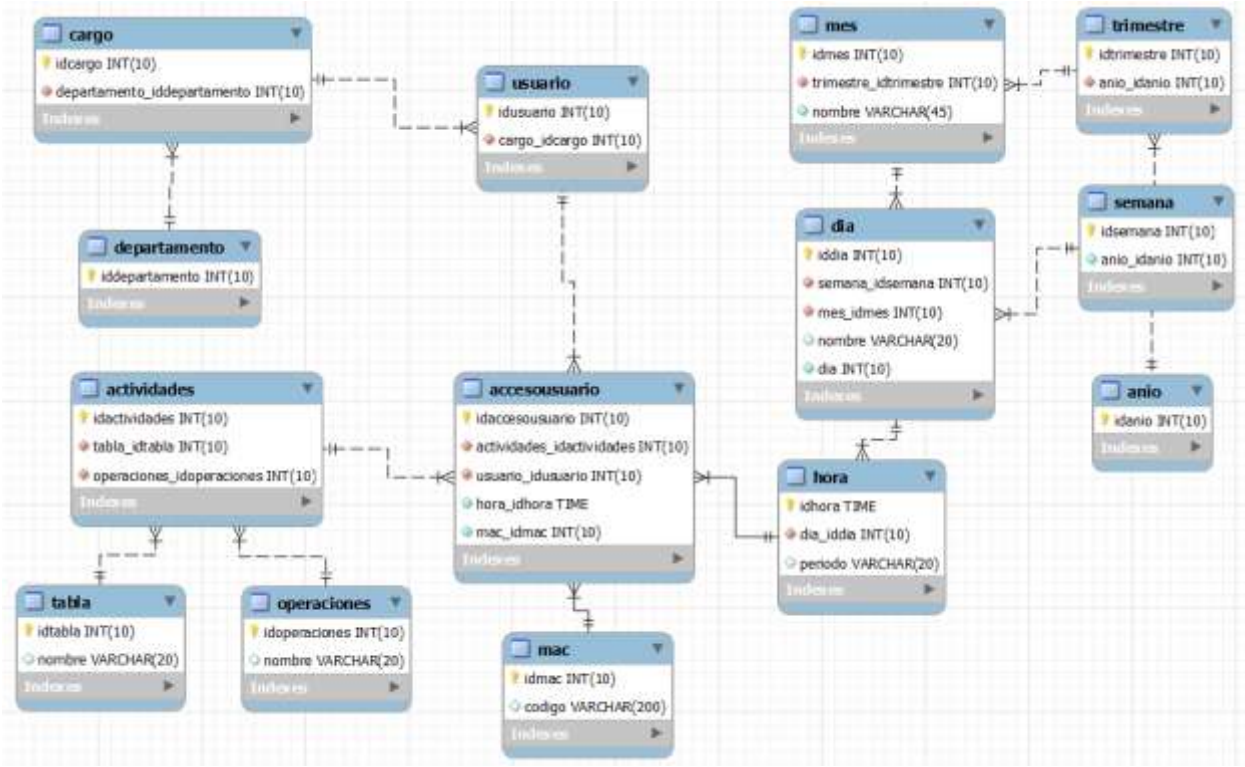

Figura 38: Diagrama con dimensiones del repositorio de datos OLAP

Con la estructura de datos creada se debe proceder a un volcado de información, utilizando procedimientos almacenados (*store procedures*), para posteriormente llenar las dimensiones del OLAP. Hay que realizar conjuntamente esta tarea con la de limpieza y transformación de la información, para conseguir que los datos almacenados en el repositorio sean correctos.

3.3. Reconocimiento, limpieza y transformación

Esta etapa se realiza cuando ya se tiene una estructura de repositorio de datos, para poder, solucionar los inconvenientes que tenga la información en las bases de datos de origen o mejorarla para su posterior procesamiento, y para ser transportadas al OLAP.

En esta tarea se han encontrado varias inconsistencias que se han solucionado con algunas técnicas muy sencillas, para adecuarse al formato del nuevo contenedor de la información unificada. Una de estas técnicas es la descomposición de la información y la unificación de formatos.

3.3.1. Descomposición de la información

En varias de las tablas de las bases de datos de origen existían algunos atributos que poseían un formato muy extenso, que al analizarlo brindaba poca información, por lo que se ha optado por dividirlo en varias partes para que sea más útil cuando se analice.

Uno de estos atributos es el de la fecha de las actividades del usuario, que tenía un formato de tipo fecha y hora conjuntamente. Dicho atributo se dividió en seis variables para poder ser tenidos en cuenta en la dimensión tiempo. Se realizó de la siguiente manera (Figura 39):

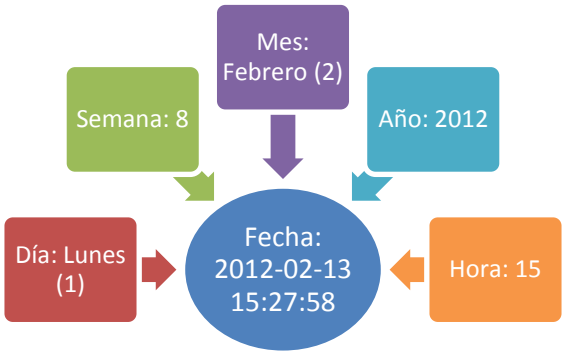

Figura 39: Diagrama de descomposición del atributo fecha

Otros de los atributos que requerían descomponerse para poder ser parte del OLAP era la dimensión de usuario, que comprendía un código de identificación personal, nombres, apellidos, cargo y departamento. Estos datos, al ser información confidencial, se han codificado de la siguiente manera (figura 40). Se han considerado XX cargos, YY departamentos, y un solo usuario.

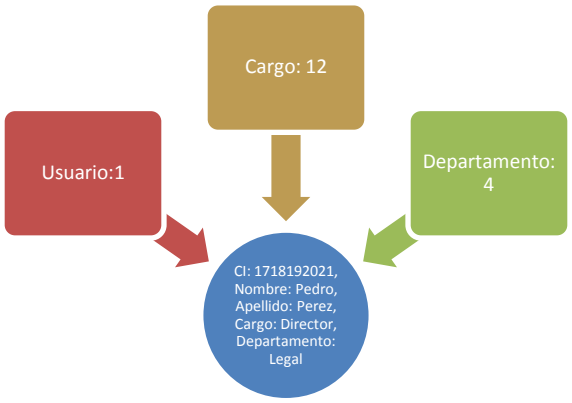

Figura 40: Diagrama de descomposición del atributo usuario

El atributo que requiere utilizar en este procedimiento es el de la información de tablas accedidas por el usuario, y las operaciones que se realizan sobre ellas, ya que esa información es reservada para la institución, figura 41.

Figura 41: Diagrama de descomposición del atributo operación y tabla

En la figura 41, muestra esta manera se obtiene el formato adecuado para el repositorio de datos que permite, por un lado, ocultar la información privada de la entidad estatal, y por otro trabajar de forma eficiente con las bases de datos.

3.3.2. Reconocimiento de la información

Una vez integrados los datos en el repositorio de datos, el primer paso que se debe realizar es hacer un resumen de la información o de las características (informe de estado) de atributos o de todo el almacén de datos. También puede ser interesante mostrar las características importantes como: valores máximos, mínimos y medios, distinguir entre valores nominales y numéricos, e integrarlo todo en una misma tabla. Para realizar este análisis de la información recopilada se va a utilizar la herramienta weka, que ayudará a hacerlo de una manera fácil, eficiente y automatizada.

La herramienta weka, ayuda a preparar de la información, lo que facilita a su compresión y visualización de manera gráfica. La herramienta realiza un barrido de los datos con algoritmos de análisis y brida la información con datos exactos y eficientes de cada variable. Como se puede apreciar en la figura 42, existen diversidad de valores en cada uno de los atributos de la base de datos, de los cuales se detallarán por separado los más significativos a continuación.

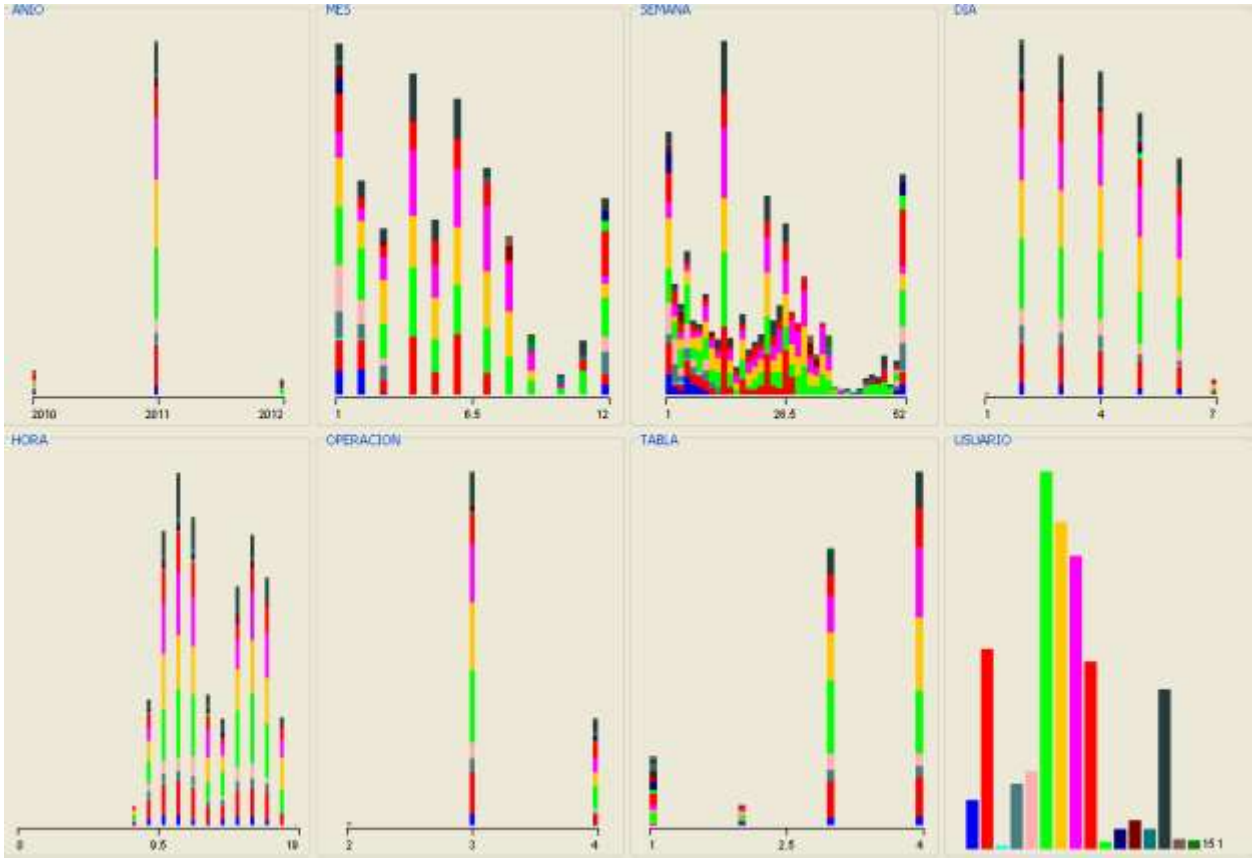

Figura 42: Diagrama de distribución de variables de la base de datos OLAP.

Cada uno de los colores representados en las graficas corresponden a los valores de cada uno de los usuarios en su respectivo atributo como año, mes, semana, día, etc.

Variable "Usuario"

La variable usuario presenta los 18 usuarios y la cantidad de registros que han ingresado cada uno. Por ejemplo, en la gráfica se puede observar que el usuario representado por el color verde es el que realiza más veces la tarea de introducir información, seguido por el usuario amarillo, lo que se debe tomar en cuenta para realizar la creación de un patrón de comportamiento con suficientes datos del usuario, para que brinde mayor información. Como muestra en la figura 43.

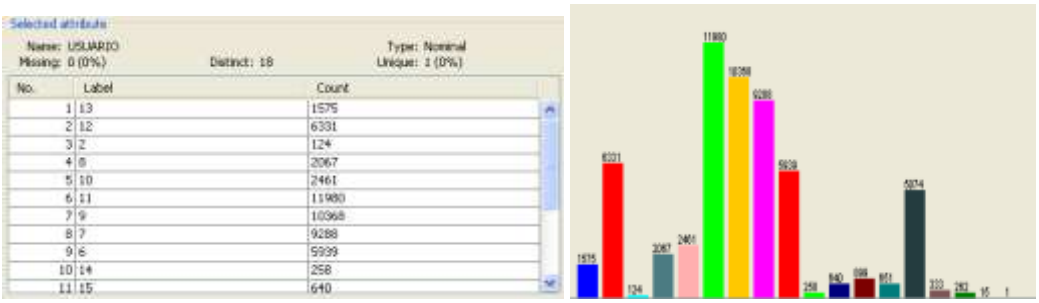

Figura 43: Gráfica de variable usuario y descripción de los datos.

Variable "Año"

Esta variable como todas las que conforman la dimensión del tiempo tienen una regularidad según la muestra tomada, dado que los valores tomados son desde el 2010 hasta febrero del 2012. De esta forma, solo presentará más información del año 2011 como se puede observar en la gráfica que muestra la tendencia de los datos. Como indica en la figura 44.

Además, señala la descripción estadística de la información, que es la siguiente:

- máximo: 2012,
- mínimo: 2010,
- media: 2010.97,

desviación estándar: 0.329.

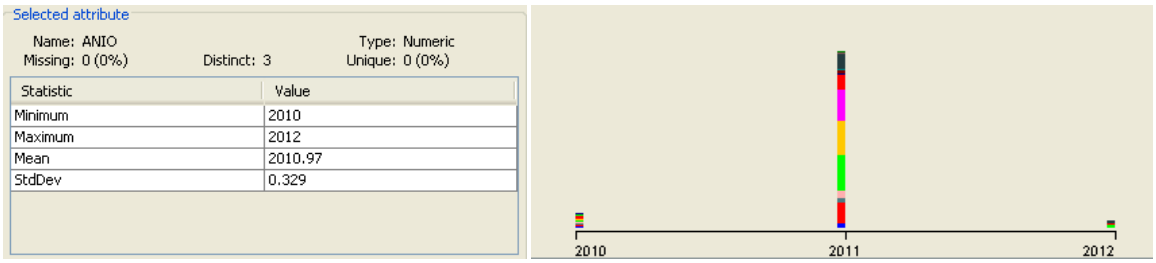

Figura 44: Gráfica de variable año y descripción de los datos.

Variable "Mes"

De igual forma, la variable mes contiene la información clasificada por el usuario para tener una idea de que usuario y en qué mes se ha ingresado la mayor parte de los datos. La herramienta de análisis de datos muestra la siguiente gráfica con su respectiva descripción estadística:

- máximo: 12,
- mínimo: 1,
- media: 5.25,
- desviación estándar: 3.28.

Se puede apreciar, que los meses cerca de fin de año son en los cuales se ingresa menor cantidad de información en la base de datos por parte de los usuarios, como se presenta en la figura 45.

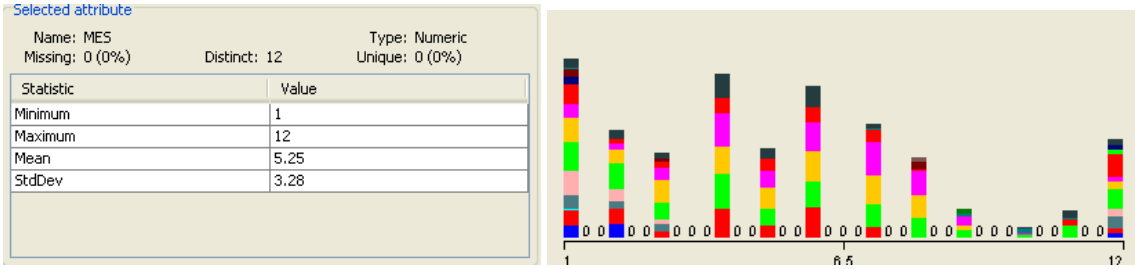

Figura 45: Gráfica de variable mes y descripción de los datos.

Variable "Semana"

En la variable semana, se presenta una tendencia similar a la del mes, como muestra en la figura 46, la cual indica que las semanas cercanas al final del año, el ingreso de información es casi mínimo. Por eso, se debe realizar un análisis con más información que contemple este inconveniente. El análisis estadístico de la información es la siguiente:

- máximo: 52,
- mínimo: 1,
- media: 20.767,
- desviación estándar: 14.523.

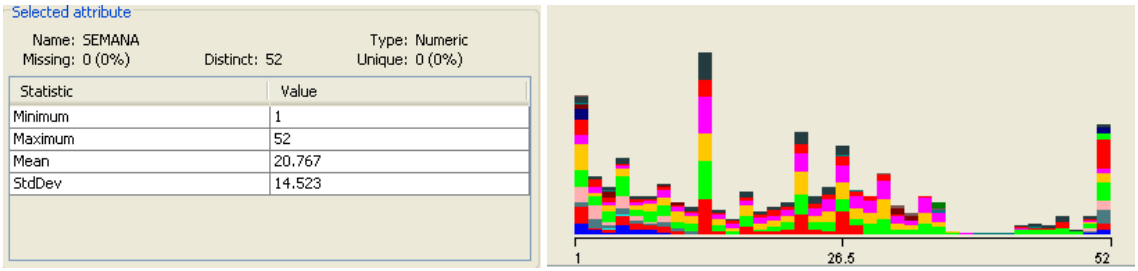

Figura 46: Gráfica de variable semana y descripción de los datos.

Variable "Día"

En el análisis de esta variable podemos observar en la figura 47, que el día menos productivo de los usuarios para el ingreso de datos es el día domingo con el número 1, seguido por el día sábado con el día número 7. El análisis estadístico de la información es la siguiente:

- máximo: 7,
- mínimo: 1,
- media: 3.84,
- desviación estándar: 1.412.

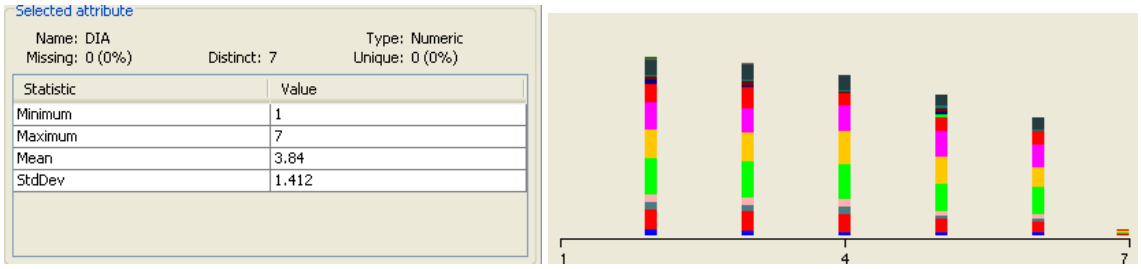

Figura 47: Gráfica de variable día y descripción de los datos.

Variable "Hora"

En la variable hora es donde se puede detectar con mayor facilidad el comportamiento del usuario en la interacción con la base de datos, como presenta en la figura 48, ya que establece

horas para el ingreso de información. Como se puede apreciar en la gráfica, las horas donde el usuario realiza el mayor ingreso de datos es desde las 8h30 hasta las 18h00, siendo el mayor ingreso de información entre las 10h00 y 16h00. Además existe un descenso en el ingreso de datos de la gráfica, en el periodo de la hora de almuerzo y receso de medio día. El análisis estadístico de la información, es la siguiente:

- máximo: 19,
- mínimo: 0,
- media: 13.232,
- desviación estándar: 2.779.

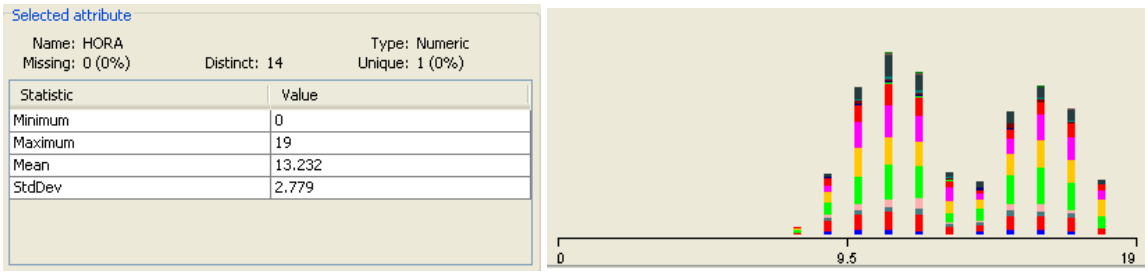

Figura 48: Gráfica de variable hora y descripción de los datos.

Variable "Operación"

La variable operación indica en la figura 49, que en la operación 3 y 4 son las que más registros se han realizado en la base de datos, por otra parte, la operación 2 tiene pocos datos ingresados. El análisis estadístico de la información, es la siguiente:

- máximo: 4,
- mínimo: 2,
- media: 3.224,
- desviación estándar: 0.437.

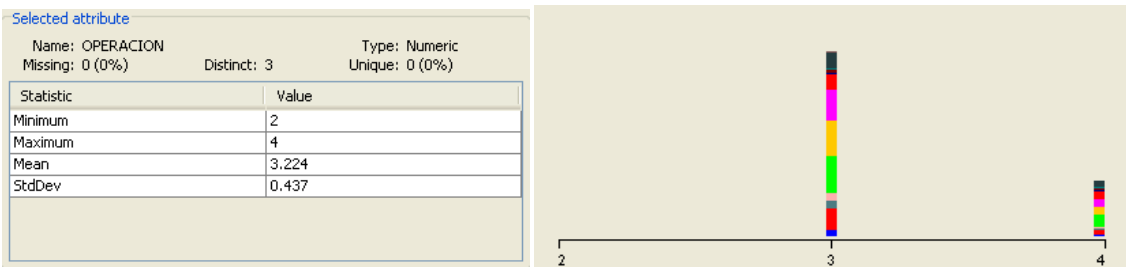

Figura 49: Gráfica de variable operación y descripción de los datos.

Variable "Tabla"

La variable tabla presenta la información de en qué tabla se ha ingresado la mayor parte de los registros en la base de datos. Como indica la figura 50 las operaciones 4 y 3 son las que tienen mayor cantidad de registros, y las operaciones 1 y 2 son las que menos son utilizadas por los usuarios. El análisis estadístico de la información, es la siguiente:

- máximo: 4,
- mínimo: 1,
- media: 3.27,
- desviación estándar: 0.912.

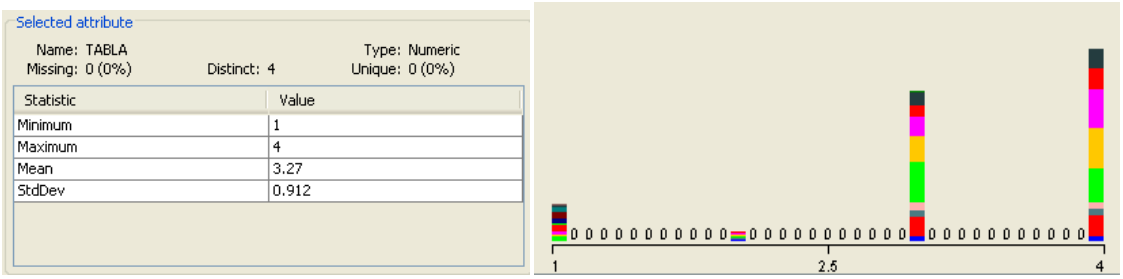

Figura 50: Gráfica de variable tabla y descripción de los datos.

3.4. Exploración y selección de datos

Una vez que están los datos ya recopilados, integrados y limpios, se debe realizar una tarea de análisis exploratorio para extraer la máxima información posible. Posteriormente se utilizarán las distintas metodologías de clasificación y búsqueda de patrones. Se debe identificar el

conocimiento que va a ser significativo para la obtención de patrones, para conseguir así un resultado más afín con la realidad.

El objetivo principal de esta fase es obtener una "*tabla de datos*", consistente en una tabla clásica en el sentido de base de datos. Esta vista debe contener toda la información relevante y necesaria para realizar las tareas de análisis.

Para realizar esta selección de la información se debe cumplir con otros subprocesos para canalizar de la mejor manera la información recolectada, para lo que se debe reconocer el objetivo del estudio y proceder a la exploración de los datos. Estos se detallarán a continuación.

3.4.1. Reconocimiento y objetivo del negocio

Para dotar de una directriz a las tareas que se van a realizar posteriormente, se van a describir aspectos importantes y reglas para la creación de un escenario de toma de decisiones lo más real posible. Para ello se van a plantear varias preguntas y así conocer el comportamiento de la información y su uso.

1. ¿Qué desea obtener de la información recopilada?

Obtener un patrón de comportamiento de cada usuario de una red de sistemas de información, para detectar el acceso no autorizado de intrusos.

2. ¿Qué reglas están definidas para modelar el ingreso de la información?

La información es recopilada individualmente por cada sistema, tanto del acceso como de la utilización. Las reglas definidas para el acceso a los usuarios son las estándares de todo sistema, usuario y clave, pero en cada sistema son diferentes y no están vinculadas entre sí.

3. ¿Existe documentación para la recopilación de esta información?

No existe documentación que respalde la información recopilada, ya que fue realizada de manera emergente a las necesidades de cada institución.

Como se puede apreciar, la información proviene de varias bases de datos y de sistemas distintos, por lo que hay que definir un patrón de comportamiento de cada uno de los usuarios a

partir de los datos recopilados y, en el caso de que esté fuera de ese patrón, detectar una intrusión.

3.4.2. Análisis exploratorio de datos

El objetivo principal, como ya se había señalado, es obtener una vista de la información que facilite su exploración. Para eso se van a utilizar varias técnicas de exploración de datos, de generalización, agregación y selección de la información.

Los objetivos de la exploración y visualización de datos son:

- Detección de patrones, anomalías y tendencias a partir de imágenes o gráficos, y facilitar la mejor comprensión de los datos.
- Facilitar al analista el discernimiento de los patrones descubiertos por las herramientas utilizadas.

Para este análisis de variables se debe tener en cuenta que se tiene un conjunto inicial de usuarios muy grande, por lo que se va a utilizar una muestra de tres usuarios que permita evaluar mejor el comportamiento de cada una de las variables.

Variable "Año"

En el atributo año, como indica en la figura 51, se posee información de tres usuarios distintos el usuario 6,8 y 11; donde cada gráfica presenta una distribución similar en cada uno de los usuarios, más aun entre los usuarios 6 y 11, que casi son idénticas. Es por ello que este atributo no proporciona información muy relevante para el estudio y se lo excluye del proceso.

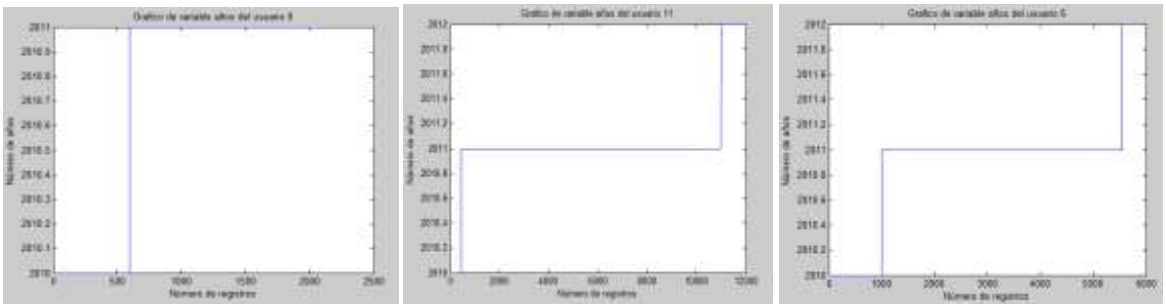

Figura 51: Diagramas de datos de la variable año de los usuarios 8,11 y 6.

Variable "Mes"

En el atributo mes, como se muestra en la figura 52, el comportamiento de los usuarios 6,8 y 11 durante un período de 12 meses, lo que se puede determinar que la similitud de las gráficas no sugiere un patrón distintivo para la identificación de un usuario definido, es por ello que se lo excluye del estudio.

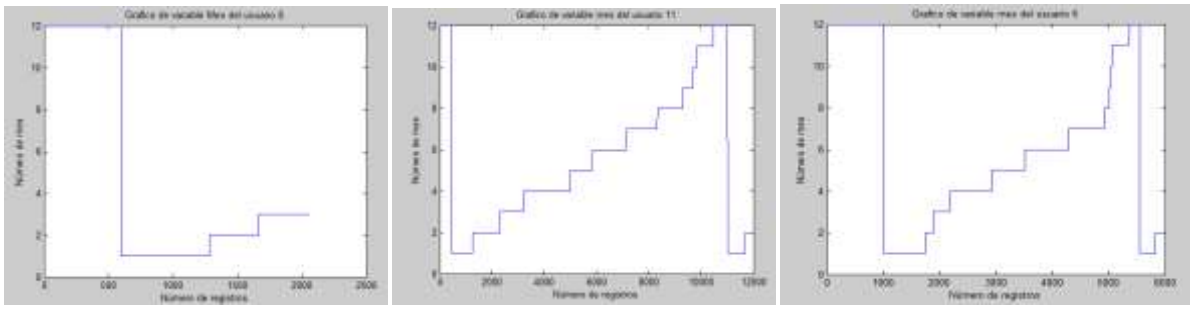

Figura 52: Diagrama de datos de la variable mes de los usuarios 8,11 y 6.

Variable "Semana"

En el atributo semana, como se muestra en la figura 53, las graficas de distribución de la información obtenida de los usuarios 6,8 y 11 es entre sí similar, lo que no indica un patrón para la identificación particular de cada usuario, por ello el atributo se excluye del estudio.

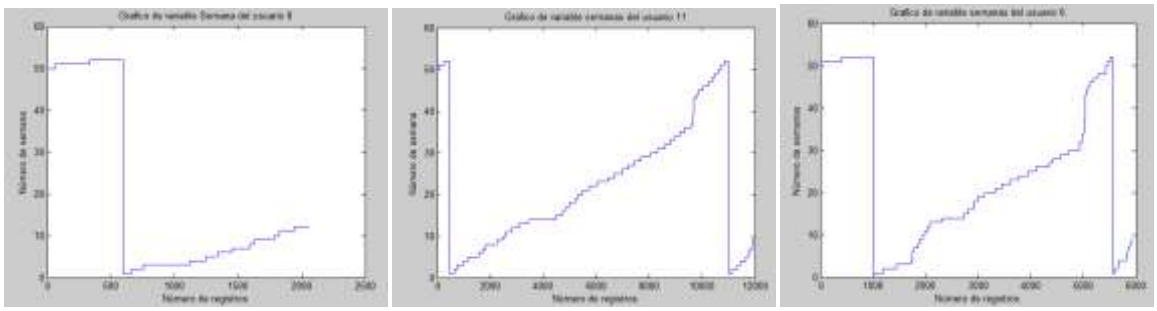

Figura 53: Diagrama de datos de la variable semana de los usuarios 8,11 y 6.

Variable "Día"

En el atributo día, como muestra la figura 54, una distribución de los días en los que los usuarios 6,8 y 11 realizaron alguna actividad dentro del sistema, en este caso, cada gráfica difiere mucho entre sí, además de tener una variación y comportamientos distintivos para cada uno de los

usuarios. La característica de este atributo puede ser fundamental para el estudio de la identificación de un patrón de comportamiento.

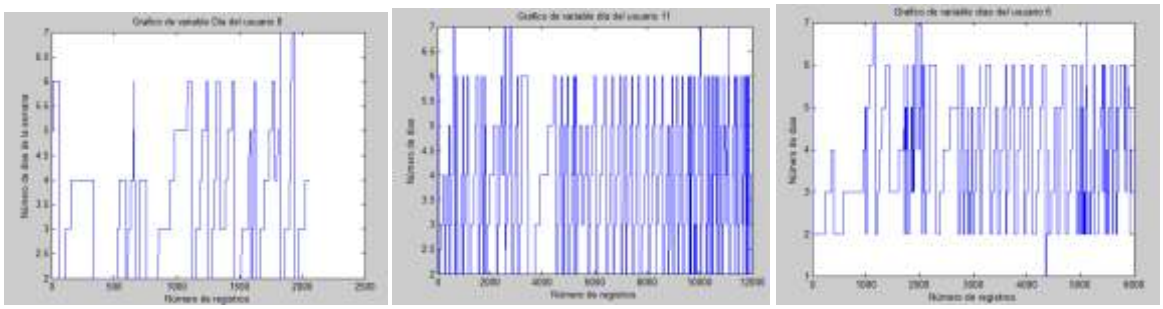

Figura 54: Diagrama de datos de la variable día de los usuarios 8,11 y 6.

Variable "Hora"

En el atributo hora, como presenta en la figura 55, indica una distribución diferente en cada uno de los usuarios que realizaron tareas en el sistema, además de una periodicidad única en cada uno de las gráficas. El atributo hora es el que ofrece más cantidad de información para la creación de un patrón de comportamiento, es por ello que es fundamental tomarlo en cuenta para el estudio.

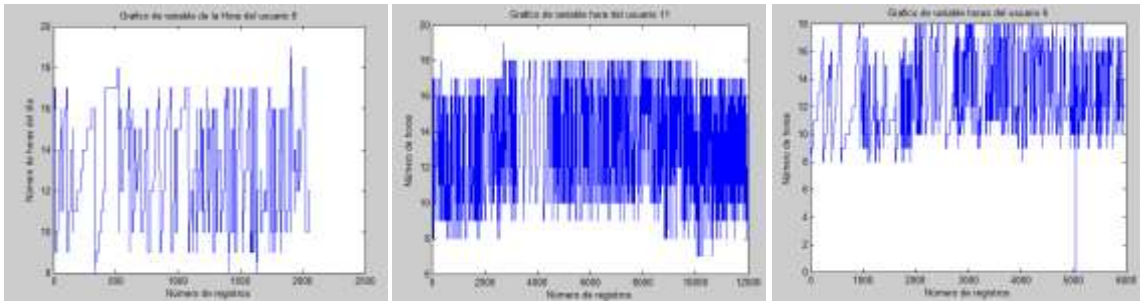

Figura 55: Diagrama de datos de la variable hora de los usuarios 8,11 y 6.

Variable "Operación"

En el atributo operación, como se observa en la figura 56, tiene una clara gráfica de distribución de datos muy particular para cada uno de los usuarios, además es uno de los atributos que reflejan de una manera detallada el comportamiento del usuario dentro del sistema y combinándolo con otros atributos se puede obtener un resultado más eficiente y real.

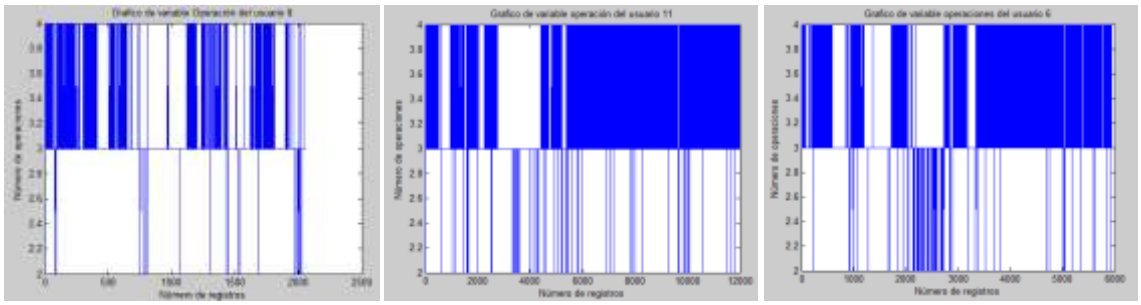

Figura 56: Diagrama de datos de la variable operación de los usuarios 8,11 y 6.

Variable "Tabla"

El atributo tabla, como se señala en la figura 57, es uno de los atributos que más aportan información del comportamiento de cada uno de los usuarios, en las graficas se puede observar que cada una tiene una periodicidad y distribución muy diferentes, por ello este atributo es un pilar fundamental para el estudio.

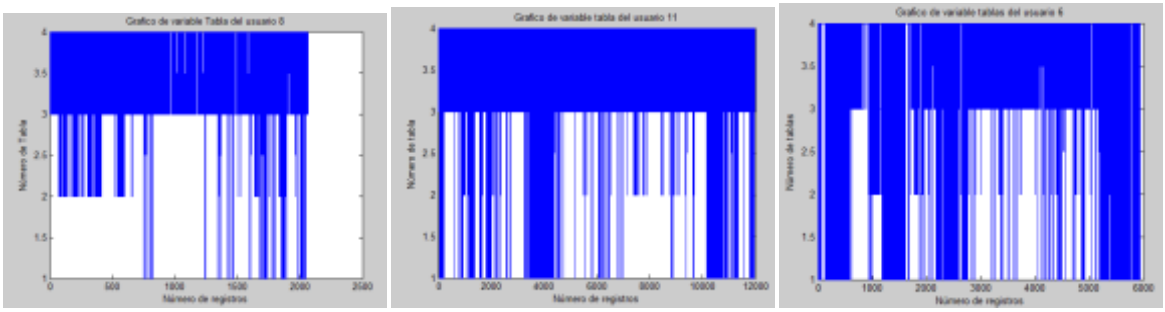

Figura 57: Diagrama de datos de la variable tabla de los usuarios 8,11 y 6.

Después de realizar las graficas de cada una de las variables , se han observado varias tendencias en cada una de ellas según el usuario.

DÍA	HORA	OPERACIÓN	TABLA	USUARIO
5	16	4	4	13
5	16	4	4	12
5	16	4	3	11
5	16	4	3	13
5	16	3	3	11

Tabla 6: Tabla que representa una muestra de la vista de información relevante

Se ha decidido no utilizar las variables año, mes, y semana, ya que no aportan información relevante para la obtención del patrón del comportamiento del usuario. Esto se puede apreciar en las gráficas de dos de los tres usuarios tienen curvas muy similares, por lo que no identifican de forma única a cada usuario, ya que los datos de los usuarios coinciden en el mismo periodo de tiempo. Por otra parte, las variables día, hora, operación y tabla, son variables que según la curvas de cada usuario sí son diferentes en cada caso y pueden ser utilizadas para obtener un patrón único de cada usuario. Por lo tanto, la tabla 6, con una muestra de la información relevante.

Capítulo 4 -Aplicación de técnicas inteligentes de reconocimiento de patrones

La extracción del conocimiento partiendo de datos ya procesados tiene como objetivo obtener un único un patrón de comportamiento de cada usuario que, entre otras cosas, es necesario que sea único, dinámico y lo más cercano a la realidad. Por esa razón aspectos fundamentales del estudio son la expresividad y la comprensibilidad de los modelos.

Se va a realizar una tarea de clasificación (o discriminación), para lo que se definen un conjunto de patrones (clases), en base a la correlación existente entre los datos. La clasificación incluye no sólo determinar la clase para cada nuevo ejemplo que se evalúe sino además, hallar el grado de certeza o de fiabilidad de dichas predicciones, lo que ayudará a mejorar el mejoramiento de la clasificación cada vez que sea modelada.

Conjuntamente se van a realizar tareas de detección de valores e instancias anómalas (outlayer), para precisamente detectar valores atípicos que pueden sugerir fraudes, fallos, intrusos o comportamientos diferenciados. En general esta técnica no utiliza una sola variable sino que toma en cuenta todas, con el objetivo de encontrar diferencias entre las clases. Para detectar este

patrón de comportamiento, es necesario aplicar estimadores de probabilidad, ya que si un ejemplo tiene baja probabilidad de ocurrir se puede considerar un caso aislado y por ello anómalo.

Para la clasificación se van a utilizar conjuntamente dos técnicas conocidas, los árboles de decisión y las redes neuronales. Además, dentro de los árboles de decisión se expresan probabilidades de ocurrencias en cada una de las hojas del árbol, para que sea más eficiente y se puedan detectar las anomalías en el recorrido del modelo.

La idea de fusionar las dos técnicas y así detectar de una manera más eficiente los fraudes, fallos o intrusiones en los sistemas que los usuarios utilizan es una de las principales contribuciones de este trabajo que ha resultado muy eficiente.

4.1. Árboles de decisión

En el proyecto se utiliza esta técnica de clasificación ya que es una de las más comprensibles, por su representación grafica, ya que es más fácil entenderlo sin tener un conocimiento previo. Este modelo permite expresar reglas que relacionan varios atributos a la vez y organiza la información disponible de una manera jerárquica y ordenada.

4.1.1. Aplicación de árboles de decisión

Para la aplicación de esta técnica es necesario utilizar la vista obtenida anteriormente tabla 6, para poder modelar el comportamiento de un usuario en un periodo de tiempo determinado.

Para el desarrollo de los arboles de decisión se tomarán cuatro periodos de tiempo, de uno, tres, seis y doce meses de un sólo usuario. Se realizarán los diagramas con la herramienta de Matlab y la función que tiene para construir el árbol a partir de la información proporcionada. La definición de reglas para el árbol de decisión fue generada por el algoritmo que utiliza la herramienta.

Usuario 6

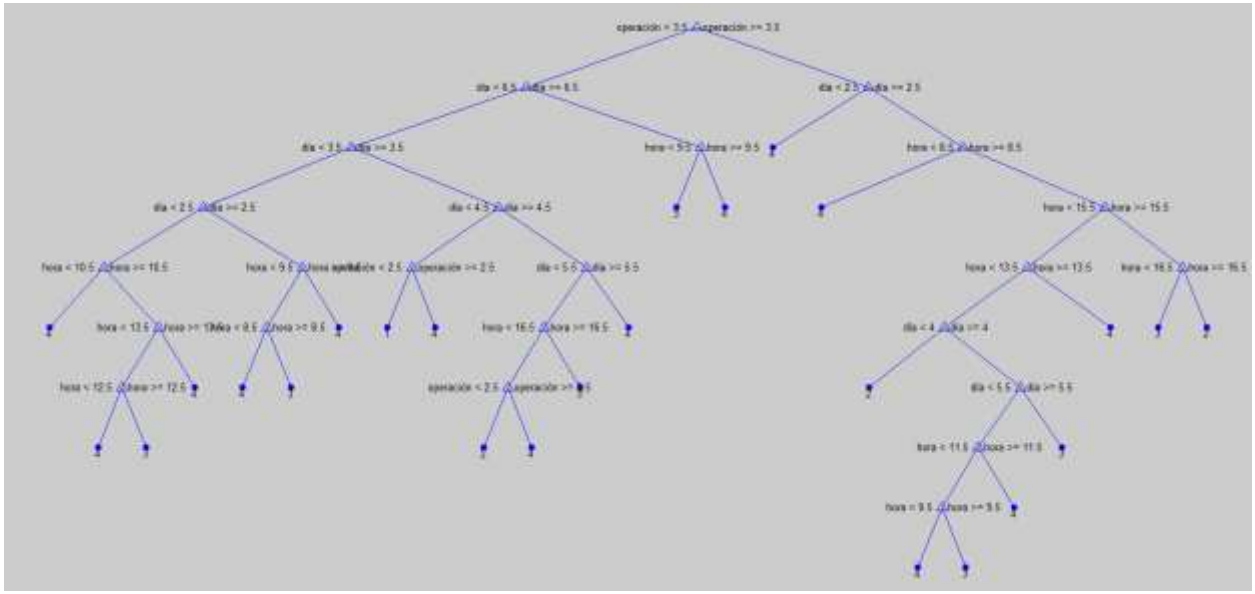

Figura 58: Diagrama de usuario 6 en un período de 1 mes

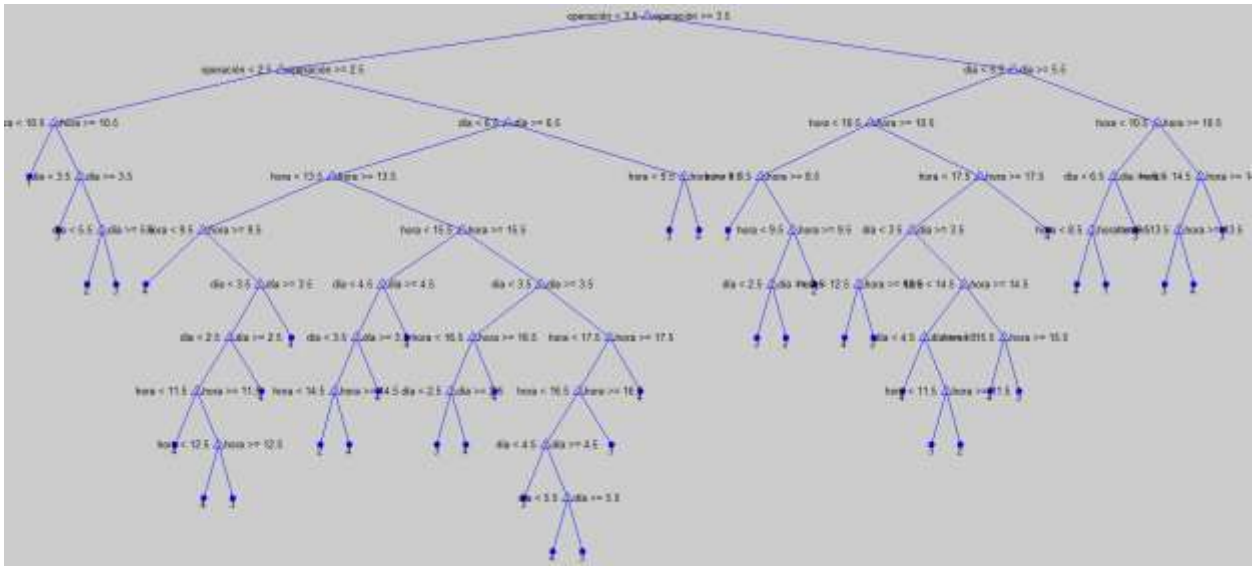

Figura 59: Diagrama de usuario 6 en un período de 3 meses

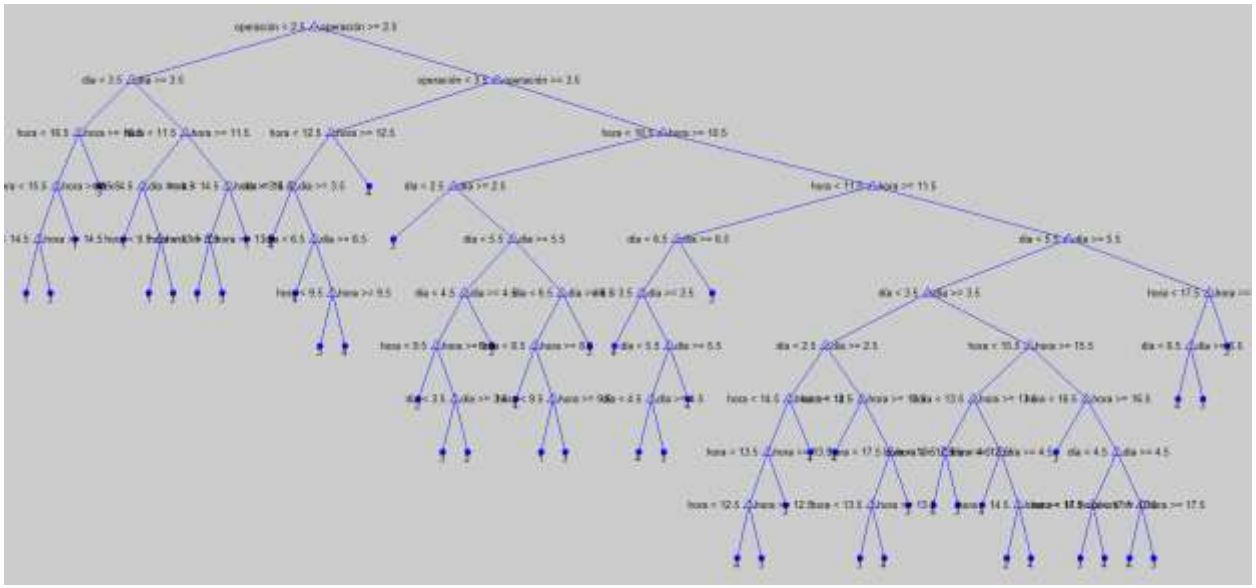

Figura 60: Diagrama de usuario 6 en un período de 6 meses

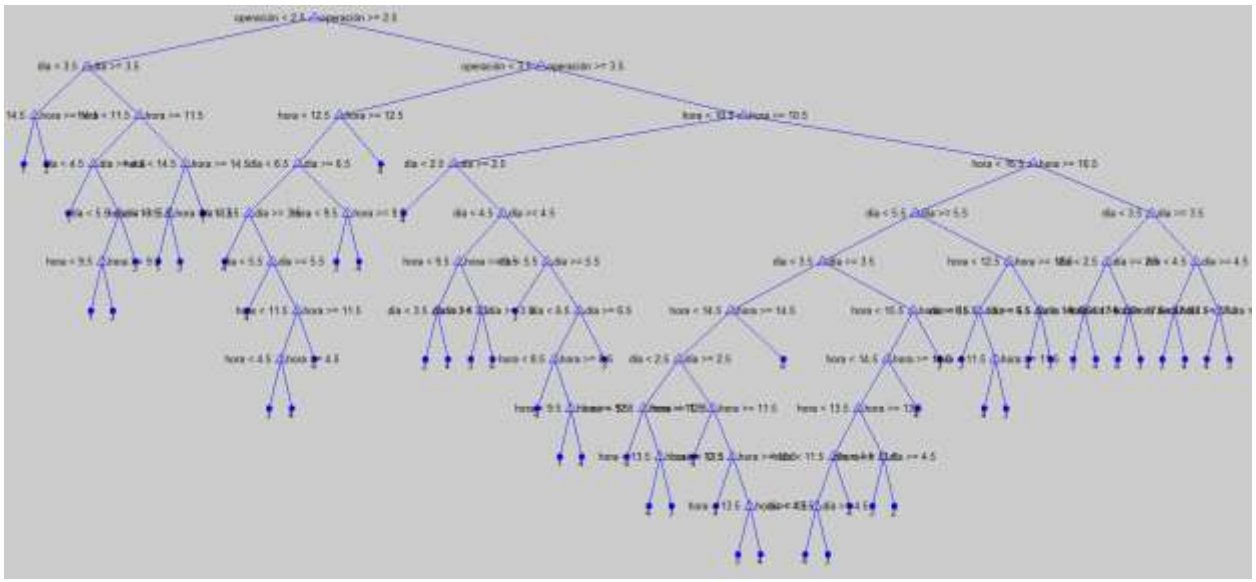

Figura 61: Diagrama de usuario 6 en un período de 12 meses

Como se puede apreciar en las gráficas anteriores 58, 59, 60 y 61, se muestra de una forma cada vez más detallada y con más nodos según el paso del tiempo, por lo tanto, para el estudio se tomará el periodo de tres meses, figura 59, para la aplicación de esta técnica, ya que muestra la información. Al seleccionar un periodo de tiempo y obtener el modelo en cual se va a trabajar, aplicaremos las técnicas estadísticas y así mostrar la probabilidad de suceso de cada una de las ramas hoja.

4.1.2. Probabilidad de sucesos en el árbol de decisión

Para indicar la probabilidad en cada rama y hoja del árbol se tomará la concurrencia de datos en cada una de las reglas generadas por la herramienta de Matlab, y se mostrará el porcentaje. Se define con la siguiente fórmula:

$$P(A)=valA/N$$

$$P(B)=valB/N$$

$$P(T)= P(A1)*P(A2)*P(A3).....*P(An)$$

Donde N es el número total de casos. Además se describen a continuación las variables de cada una de las ecuaciones:

- **valA** es el número de casos en la rama A.
- **valB** es el número de casos en la rama B.
- **P(A)**, probabilidad de sucesos en la hoja A.
- **P(B)**, probabilidad de sucesos en la hoja B.
- **P(T)** es la probabilidad final en un punto n.
- **n** es el número total de hojas que ha recorrido.

P(T) Es la probabilidad en la hoja final del árbol, de una acción realizada por el usuario según el criterio de comportamiento que describe según sus actividades realizadas. Ya que el árbol del usuario generado es demasiado extenso para describirlo en detalle, se tomará las dos ramas principales por separado para explicarlo paso por paso.

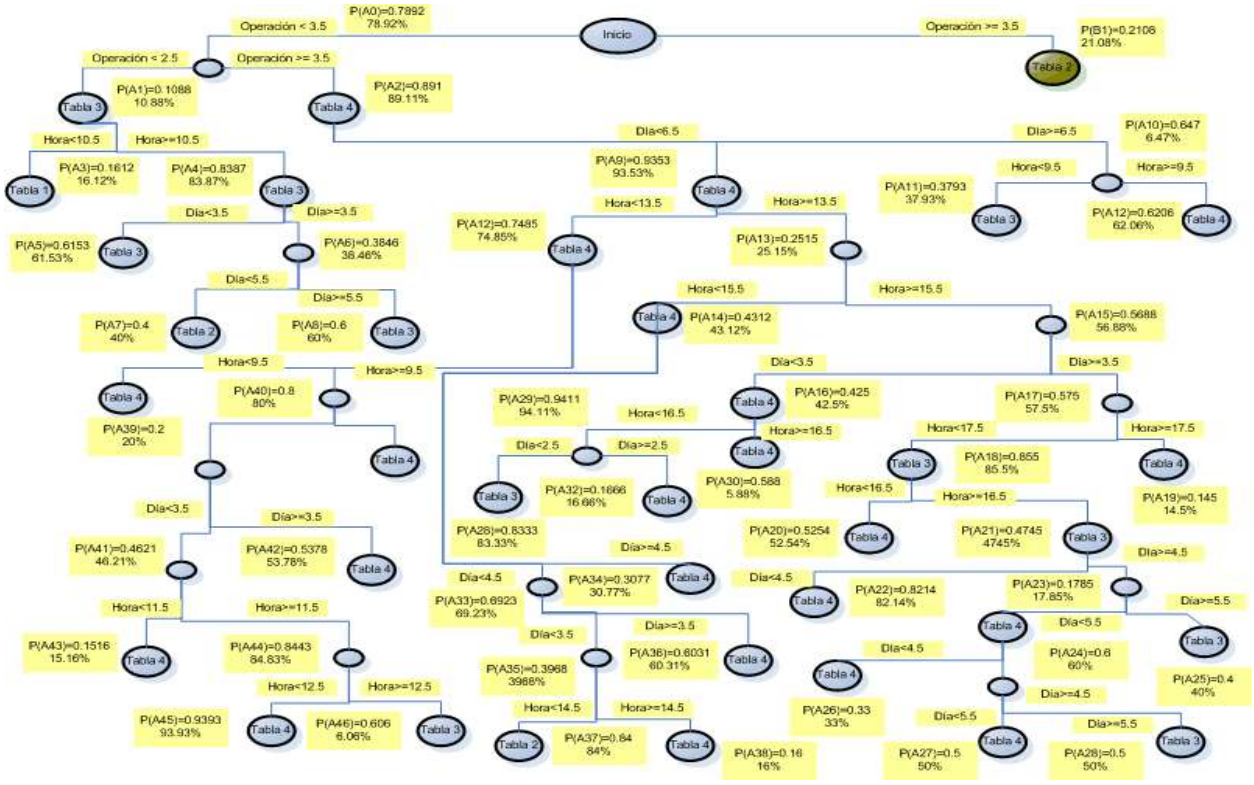

Figura 62: Diagrama de árbol de decisión de la rama A con las probabilidades y porcentajes

En la figura 62, se muestra la rama derecha (rama A), de una árbol de decisión creado con información de un usuario y obteniendo en cada punto una probabilidad, utilizando las ecuaciones anteriormente descritas. Además presenta el porcentaje de datos que intervinieron en la creación del árbol y las condiciones con la que se construyó.

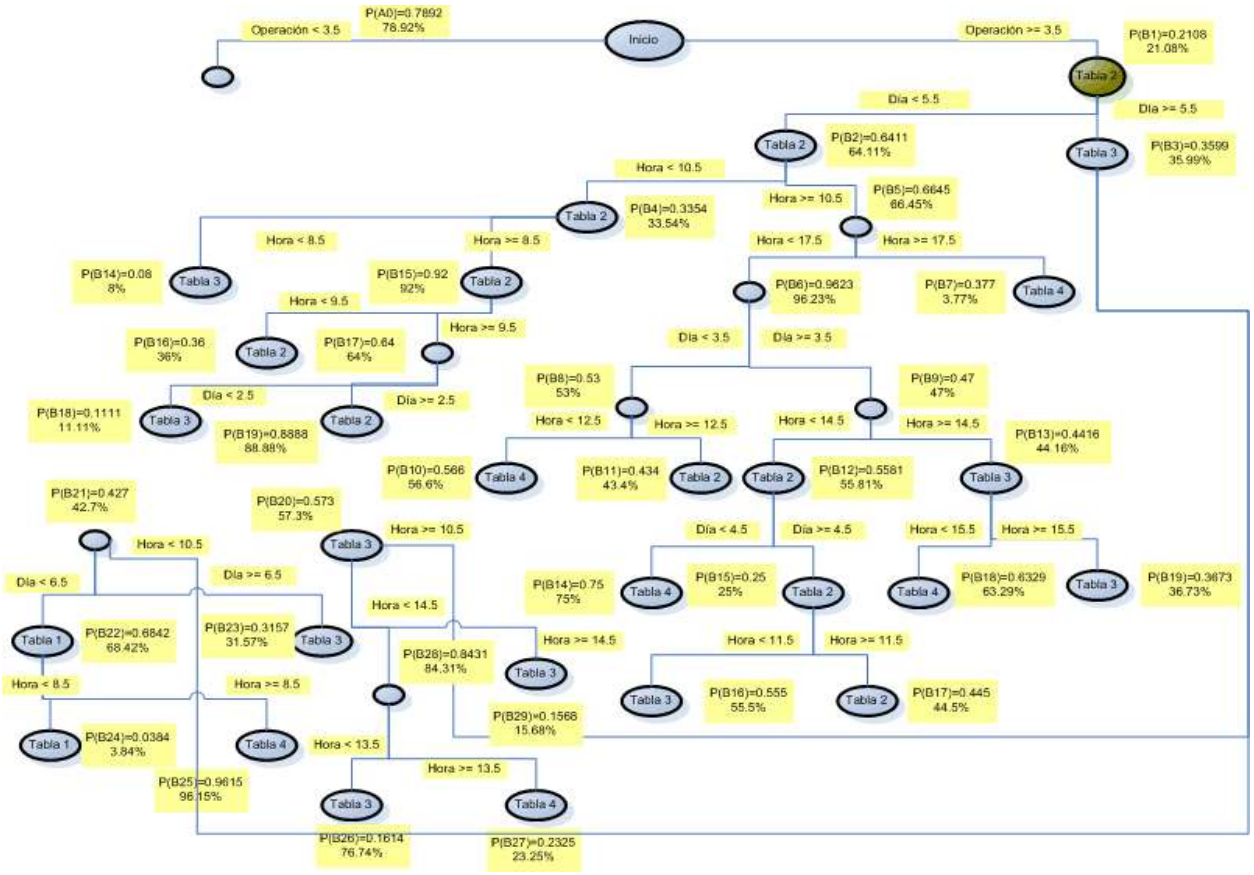

Figura 63: Diagrama de árbol de decisión de la rama B con las probabilidades y porcentajes

En la figura 63, presenta rama derecha (rama B), donde se puede apreciar que cada uno de sus puntos tiene su probabilidad de ocurrencia como también su porcentaje. Estas graficas indican que para un evento realizado por el usuario puede ser más o menos probable y con este valor se puede detectar de una manera más rápida y eficaz la intrusión de un usuario no autorizado.

4.1.3. Casos comportamiento de usuario con árboles de decisión

Para la aplicación de casos de prueba se tomarán varias de las hojas del árbol, y se utilizará la fórmula probabilística descrita anteriormente para calcular la probabilidad de ocurrencia de un evento dentro del árbol de comportamiento del usuario. Con esa probabilidad se medirá cual es la actividad del usuario dentro del sistema informático, más factible y cuál es la que no.

Para realizar los casos de prueba se toman las siguientes condiciones:

1. El usuario 6, realiza una operación 3 en la tabla 1 a las 9:15 h del día 1 (Lunes).

Desarrollo matemático

P1(T)=P(A0)*P(A1)*P(A3)

P1(T)=(0.7892)*(0.1088)*(0.1612)=0.0138

Porcentaje: 1.38%

2. El usuario 6, realiza una operación 2 en la tabla 3 a las 16:00 del día 4 (Jueves).

Desarrollo matemático

P2(T)=P(A0)*P(A1)*P(A4)*P(A6)*P(A8)

P2(T)=(0.7892)*(0.1088)*(0.8387)*(0.3646)*(0.6)=0.0157

Porcentaje: 1.575%

3. El usuario 6, realiza una operación 4 en la tabla 4 a las 12:20 del día 7 (Domingo).

Desarrollo matemático

P3(T)=P(B1)*P(B3)*P(B20)*P(B28)*P(B26)

P3(T)=(0.2108)*(0.3599)*(0.573)*(0.8431)*(0.7674)=0.0281

Porcentaje: 2.81%

En los tres casos anteriores los comportamientos del usuario 6 son distintos y cada uno tiene su propia probabilidad. El caso tres, tiene una probabilidad de valor PT(3) = 0.0281 (2.81%), lo que significa que es más probable que el usuario 6 realice la actividad tres que las restantes dos actividades. Las actividades uno y dos son descritas como comportamientos singulares, eventuales, o como una posible intrusión. Es por eso que si un usuario realiza la misma tarea la el mismo día, a la misma hora tiene una alta probabilidad de que el usuario sea el autorizado y no un intruso.

El comportamiento con menor probabilidad será un límite para considerarlo como una posible intrusión. Por otra parte, con ésta técnica de reconocimiento de patrones se puede visualizar de manera eficiente todos los comportamientos de los usuarios en detalle, pero, no describe el comportamiento que no realiza habitualmente el usuario, por lo que se complementará con la técnica de redes neuronales para tener un mejor método de detección de intrusos.

4.2. Aplicación de redes neuronales

Para la aplicación de redes neuronales se ha tomando en cuenta la vista de la Tabla 7, pero con la información de todos los usuarios que interactúan en el sistema , con el objetivo de poder identificar de forma completa si un usuario es legal o es un intruso.

Para la aplicación de redes neuronales se utiliza de nuevo la herramienta Matlab. Se ha diseñado considerando las siguientes entradas y salidas:

Entradas (Input): Contiene las columnas que representan los atributos día, hora, operación, y tabla accedida por el usuario.

DIA	HORA	OPERACION	TABLA
2	8	3	1
2	8	4	1
2	9	3	1
2	9	3	1
2	9	4	1
2	9	4	1
2	9	4	1
2	9	4	1
2	9	4	1
...	...	...	...

Tabla 7: Muestra de la tabla de entrada de datos de la red neuronal

Salidas (Output): Muestra la columna del usuario que corresponde a la actividades correspondientes a la tabla 8 de entrada.

USUARIO
10
9
9
10
13
13
11
9
9
...

Tabla 8: Muestra de tabla de salidas de datos de la red neuronal

Red(net): Red neuronal generada con la información de Entradas y Salidas.

Para la generación de la red neuronal ha sido necesario realizar varias pruebas hasta obtener el resultado esperado. La estructura final se muestra en la siguiente figura 64:

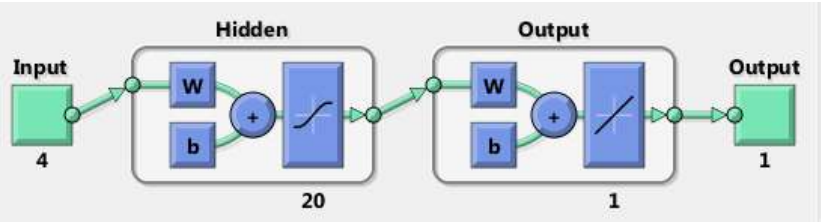

Figura 64: Diagrama de red neuronal de usuarios

4.2.1. Estructura y diseño de la red neuronal

Para la generación de la red neuronal se debe realizar una estructura de datos de entrada y datos de salida, es por eso que se va a detallar el proceso paso a paso para tener una mayor información

de cómo está diseñada la red y como se genera y que resultados brinda. Los proceso consta de los siguientes pasos:

4.2.2. Preparación de los datos

Los datos correspondientes a los problemas de clasificación se establecen para una red neuronal mediante la organización de los datos en dos matrices, la matriz de entrada X y el objetivo de la matriz T.

Cada columna i de la matriz de insumo contará con cuatro elementos que representan una día, hora, operación y tabla donde el usuario realizó las tareas dentro del sistema.

Cada columna correspondiente de la matriz de destino tendrá un solo elemento. Los usuarios por número son los que representan al primer elemento.

4.2.3. Construir el clasificador de red neuronal

El siguiente paso es crear una red neuronal que aprende a identificar el usuario según las características de cada uno de ellos.

Dado que la red neuronal se inicia con pesos iniciales aleatorios, los resultados de esta implementación variaran ligeramente cada vez que se ejecute. La semilla aleatoria se establece para evitar esta aleatoriedad. Sin embargo, esto no es necesario para sus propias aplicaciones.

La implementación será de dos capas, es decir de una sola capa oculta. Capas que no son capas de salida se denomina capas ocultas.

Además se va a trabajar con una sola capa oculta de 20 neuronas para este aplicación. En general, los problemas más difíciles requieren más neuronas, y tal vez más capas. Simplificación de los problemas requieren un menor número de neuronas.

La entrada y salida tienen tamaños de 0 debido a que la red todavía no ha sido configurado para que coincida con nuestra entrada y datos de destino. Esto sucederá cuando la red se entrena. Como se puede visualizar en la figura 66, presentada anteriormente.

Una vez que ya se tiene el diseño de la red neuronal, ya se puede proceder a entrenarla. Las muestras se dividen automáticamente en conjuntos de entrenamiento, validación y prueba. El

conjunto de entrenamiento se utiliza para enseñar a la red. Formación continúa mientras la red continúa mejorando en el conjunto de validación. El conjunto de prueba proporciona una medida completamente independiente de la exactitud de la red.

Para conocer el rendimiento de la red mejoró durante el entrenamiento, de debe observar la ventana de "rendimiento" en la ventana de ejecución, una vez que haya terminado el proceso de entrenamiento.

El rendimiento se mide en términos de error cuadrático medio, y se muestra en escala logarítmica. Rápidamente se disminuyó a medida que la red se formó.

El rendimiento se muestra para cada uno de los conjuntos de entrenamiento, validación y prueba. La versión de la red que mejor se hizo en el conjunto de validación se fue después del entrenamiento. Como lo expone en la figura 65.

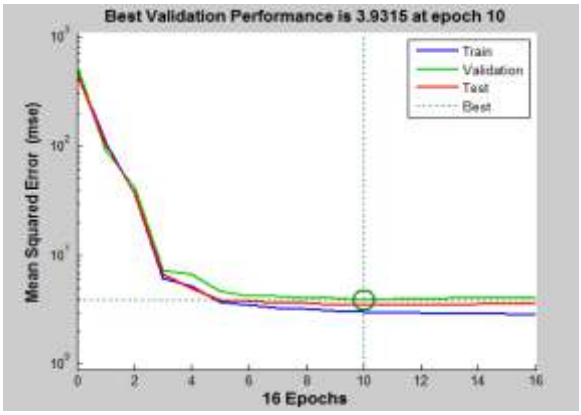

Figura 65: Diagrama de estados de regresión lineal y errores de la red neuronal

4.2.4. Comprobación del clasificador

La red neuronal entrenada ahora se puede probar con las muestras de prueba, esto dará una idea de lo bien que la red va a hacer cuando se aplica a los datos del mundo real. Una medida de qué tan bien la red neuronal se ha ajustado a los datos es la matriz de confusión, ya que traza o divide los datos a través de todas las muestras.

La matriz de confusión muestra los porcentajes de clasificaciones correctas e incorrectas. Clasificaciones correctas son las plazas verdes de las matrices diagonales. Clasificaciones incorrectas formar los cuadrados rojos.

Si la red ha aprendido a clasificar adecuadamente, los porcentajes en los cuadros rojos debe ser muy pequeña, lo que indica errores de clasificación pocos. Si esto no es el caso, entonces la formación continua, o la formación de una red con más neuronas ocultas, sería aconsejable, como indica la figura 66, en la matriz de confusión en la implementación.

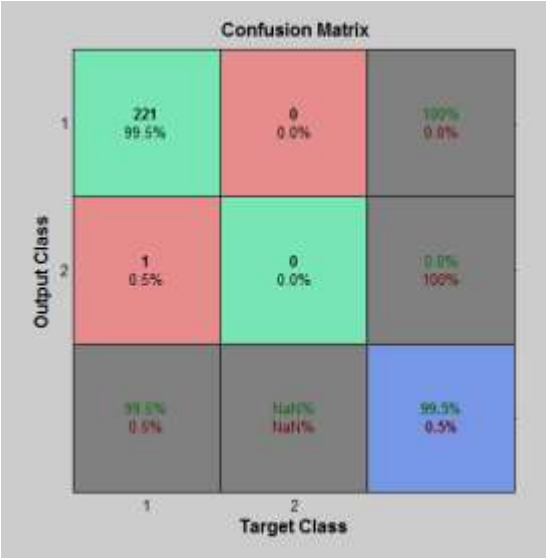

Figura 66: Matriz de confusión de la red neuronal entrenada

Otra medida para conocer el optimo entrenamiento de la red neuronal es visualizar los datos de ajuste en el diagrama de funcionamiento del receptor característico. Esto muestra cómo las tasas de falsos positivos y verdaderos positivos se refieren como el umbral de salidas se variaron de 0 a 1.

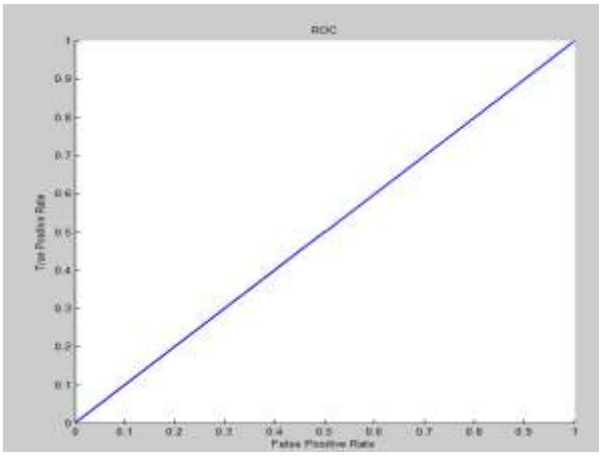

Figura 67: Diagrama de funcionamiento del receptor característico

En la figura 67, se muestra en la parte superior izquierda la línea de los falsos positivos menos que tienen ser aceptados con el fin de obtener una alta tasa de verdaderos positivos. Los mejores clasificadores tendrán una línea que va desde la esquina inferior izquierda, en la esquina superior izquierda, a la esquina superior derecha, o cerca de eso, como lo indica en la figura 70.

Capítulo 5 -Herramienta de simulación para la detección de intrusos

En este capítulo se presenta la herramienta final desarrollada para la simulación del acceso de usuarios en un sistema informático, que permite la detección de intrusos o usuarios no autorizados.

Se desarrolló una interface en Matlab, para presentar de mejor manera los datos. Esta aplicación utiliza los datos recopilados con anterioridad y emplea la información de los usuarios necesaria para realizar pruebas cruzadas de individuos distintos. Para manejar la aplicación debemos seleccionar un usuario y detallar su comportamiento en las variables día, hora, operación y tabla. Al introducir estos comportamientos la red neuronal identifica el patrón de usuario al que corresponde y muestra gráficas de las variables, como también su árbol de decisión. Así puede poder observar de manera gráfica las diferencias en el comportamiento de cada uno de los usuarios. Las líneas dentro de las gráficas representan las variables con las que se están trabajando día (color azul), hora (color verde), operación (color celeste) y tabla (color rojo).

Existen tres casos en los que la herramienta presentar una respuesta:

- **Usuarios iguales:** determina que el usuario y la información introducida es auténtica y determina que es un acceso correcto.
- **Usuarios diferentes:** presenta un mensaje de "***Usuario Intruso***", cuando la información es diferente de lo que cabría esperar para ese patrón.
- **Patrón no reconocido:** muestra en la pantalla una X cuando no se puede detectar el patrón de comportamiento, lo que puede resultar en una alerta en el sistema por una posible intrusión.

5.1. Simulación de comportamientos normales de usuarios

Las condiciones normales de un usuario se presentan cuando introduce información que pertenece a su comportamiento habitual en el sistema. Por ejemplo, el usuario 8 accede al sistema con las siguientes variables:

Día: Martes

Hora: 13:00 h

Operación: Operación 4

Tabla: Tabla 3

La aplicación da como resultado acceso autentico y muestra la siguiente pantalla (figura 71):

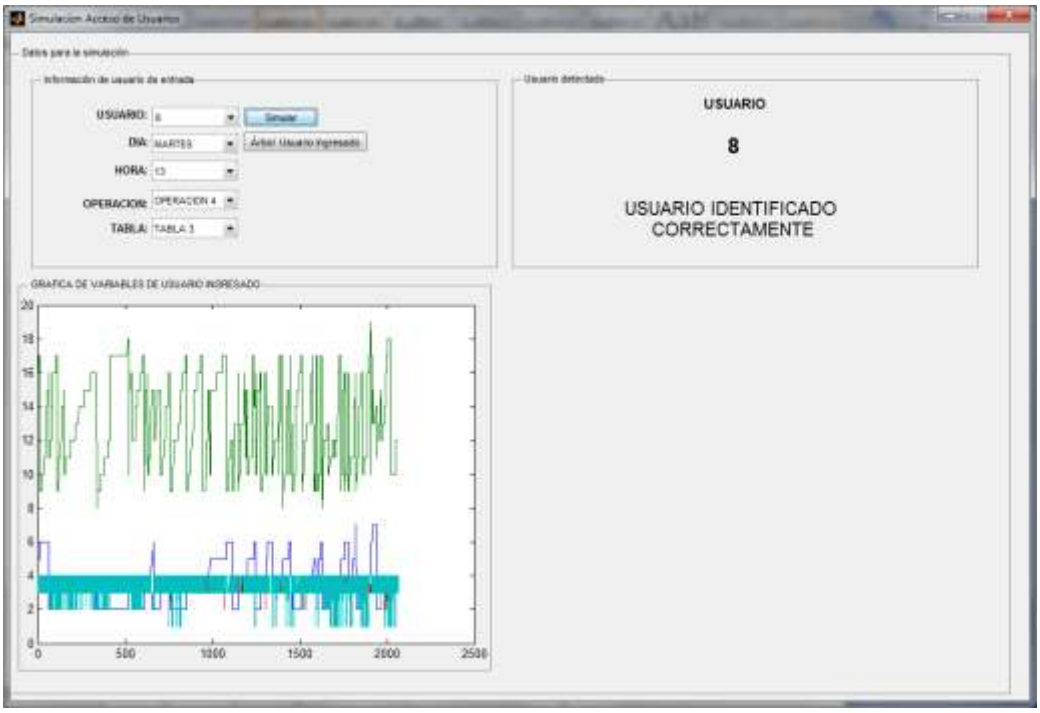

Figura 68: Aplicación de detección del usuario 8 (auténtico)

En la aplicación se puede apreciar que las gráficas de las variables de comportamiento son las mismas, además se puede verificar que sus árboles de comportamiento son los mismos. Eso sugiere que el comportamiento del usuario ingresado (usuario 8), corresponde a su desenvolvimiento normal y la red neuronal y el árbol de decisión (figura 72), lo clasifica como un usuario que accedió correctamente al sistema.

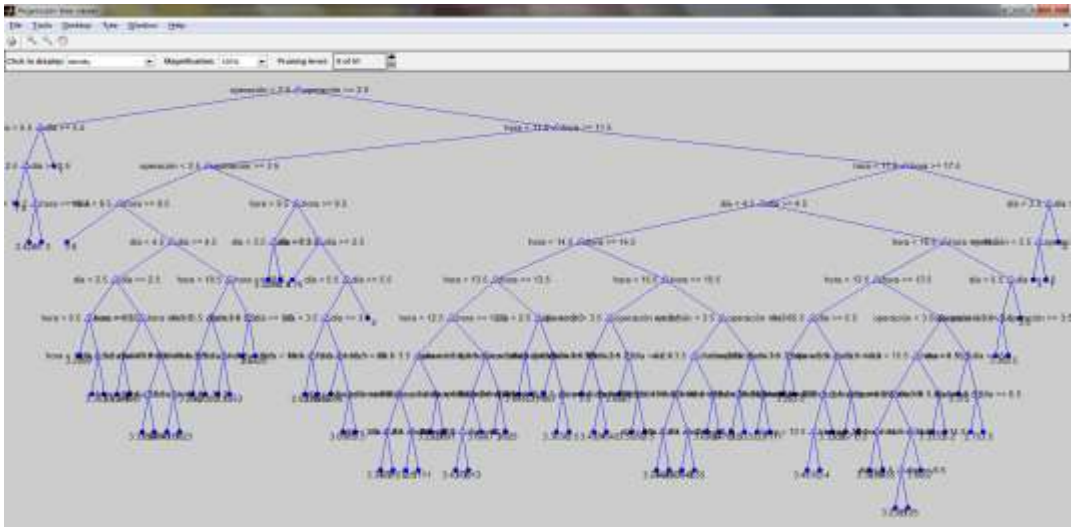

Figura 69: Árbol de decisión del usuario 8 (auténtico)

5.2. Simulación de comportamientos anormales de usuarios

Para detectar el acceso de un intruso se introduce información de un usuario con acciones diferentes a lo que sería su patrón normal de comportamiento. La herramienta presenta el usuario que más se asemeja a dicho comportamiento. En este caso se utilizó el usuario 10 con las siguientes variables de comportamiento:

Día: Viernes

Hora: 10:00 h

Operación: Operación 4

Tabla: Tabla 1

En la figura 73, se puede observar cómo los perfiles son diferentes.

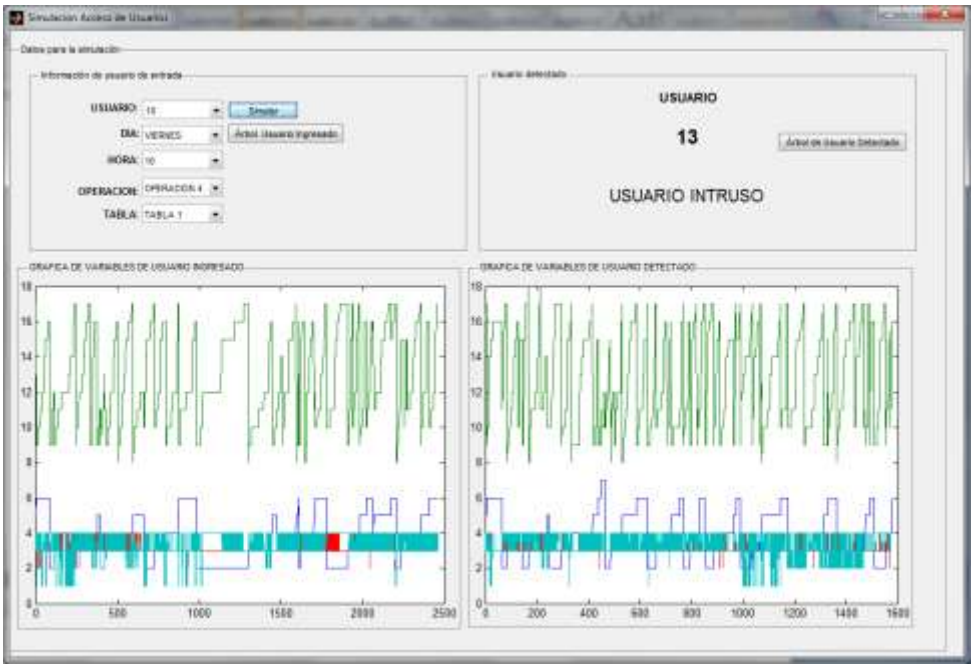

Figura 70: Aplicación de detección del usuario 10 no autentico

Por otra parte, los diagramas de los correspondientes árboles de decisión no coinciden, como muestran las siguientes figuras 74 y 75:

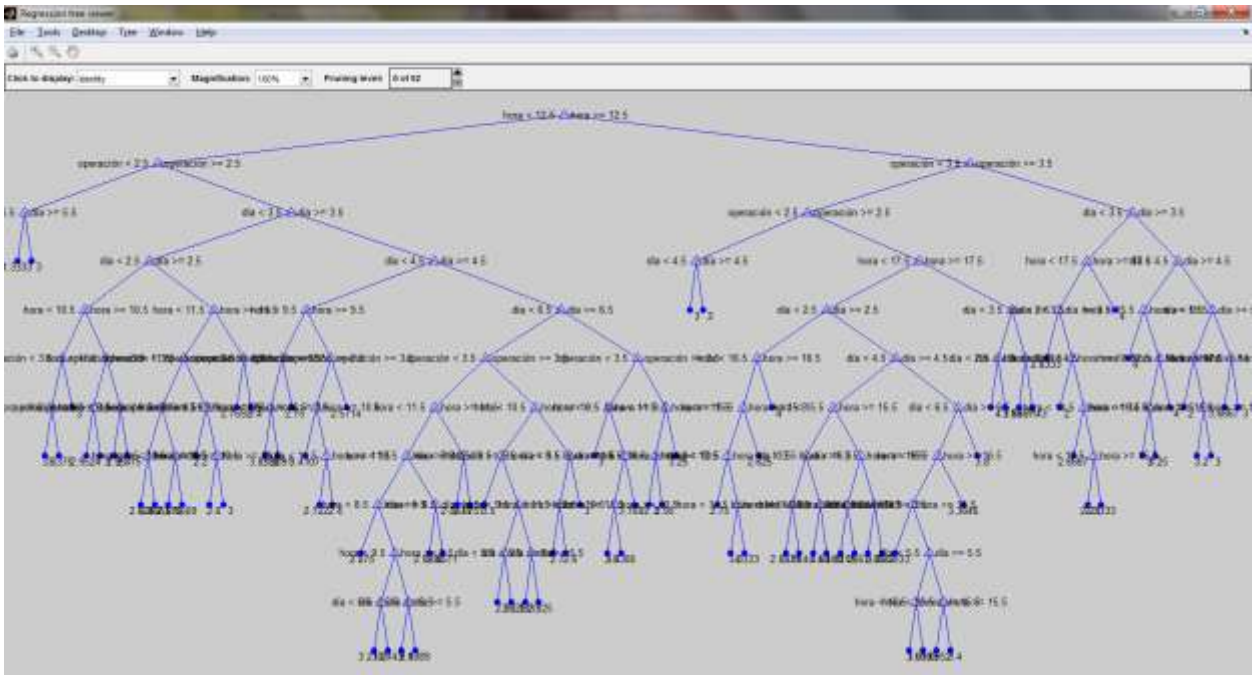

Figura 71: Árbol de decisión de usuario 10 introducido

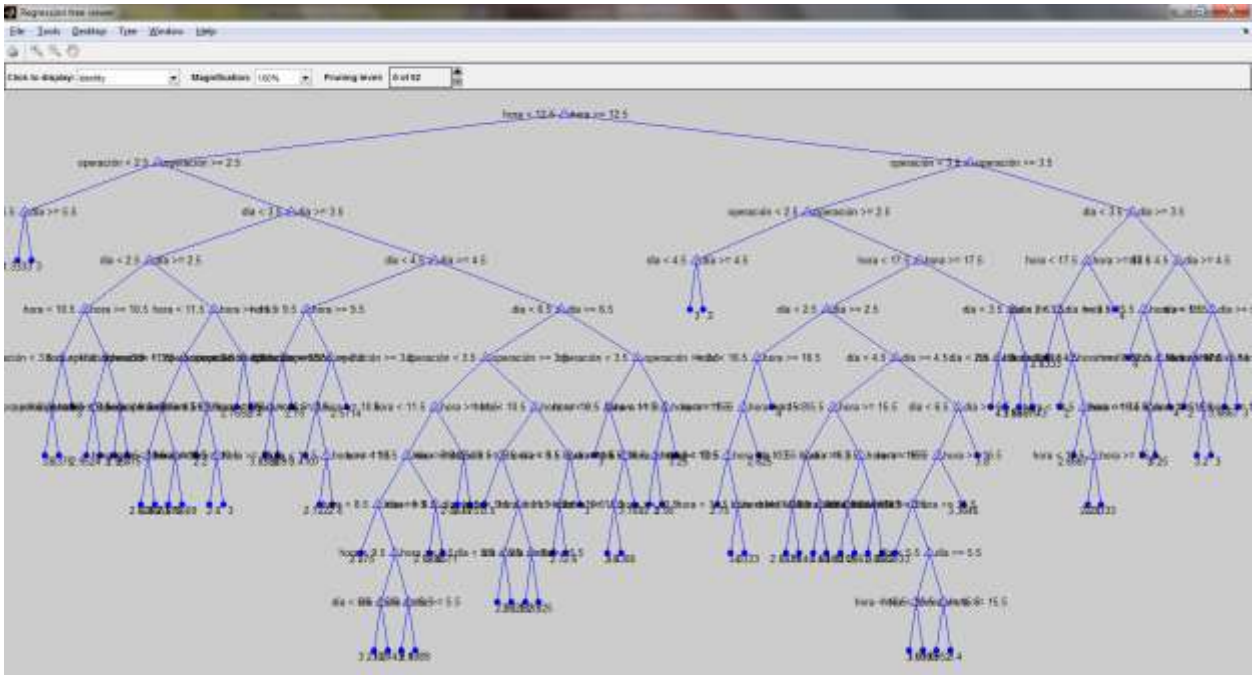

Figura 72: Árbol de decisión de usuario 13 detectado

5.3. Patrón desconocido

Finalmente, la aplicación puede también detectar un acceso de un usuario del que no tenga registrado un comportamiento, lo cual se presume que es una alerta de intrusión, pero también puede ser un comportamiento del usuario fuera de lo común. Por ejemplo, se introducen las siguiente variables al usuario 1:

Día: Miércoles

Hora: 3:00 h

Operación: Operación 2

Tabla: Tabla 1

En la aplicación se ingresará la información de un usuario en la cual no se tiene ningún comportamiento asignado a ningún individuo, por lo que presentará en la aplicación una X en el usuario, lo que significa que ese comportamiento es nuevo o una alerta de intrusión. En la figura 76, muestra la ventana con este comportamiento.

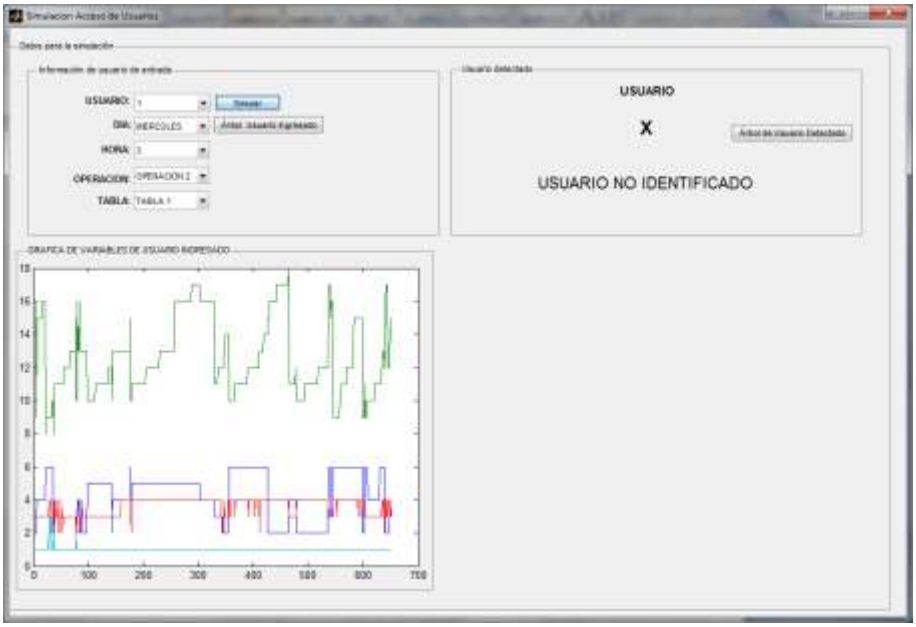

Figura 73: Árbol de decisión de usuario 1 comportamiento no detectado.

En este caso la herramienta sólo presenta una alerta "***Posible Intrusión***", que puede significar un comportamiento poco común en un usuario autorizado o una posible intrusión en el sistema.

Capítulo 6 - Conclusiones y trabajos futuros

En este capítulo se presentan los principales resultados obtenidos durante la investigación realizada sobre el uso técnicas de reconocimiento de patrones para la detección de intrusos en sistemas informáticos, se discuten los resultados obtenidos, y por último se muestran las futuras direcciones de investigación.

6.1. Conclusiones

Las técnicas de minerías de datos como la de árboles de decisión y redes neuronales, son unas de las más eficientes al momento de realizar una tarea de identificación de patrones y clasificación, es por eso, que al ser utilizarlas de manera conjunta se puede obtener resultados reales y más ligados a la realidad. Es el caso del presente estudio, la detección de intrusos dentro de sistemas

informáticos, con estas técnicas se ha obtenido patrones de comportamiento únicos para cada usuario y la posibilidad de revelar su identidad con una eficacia notable.

Para poder aplicar estas técnicas se realizó un proceso de depuración de la información proporcionada por la institución pública de la república del Ecuador. Se procedió a para refinar los datos, y así alcanzar un mejor partido de la información para identificar los patrones de comportamiento de cada uno de los usuarios del sistema y así obtener como resultado una forma automatizada de control de intrusiones en los sistemas de dicha institución, finalmente creando perfiles de comportamiento para cada uno de los usuarios.

En este trabajo se han realizado comparaciones entre las dos técnicas utilizadas, dando como resultado final la aplicación conjunta de las mismas para obtener el resultado deseado. Además se ha desarrollado una herramienta de simulación, que muestra la aplicación de las técnicas y así poder visualizar una detección de intrusiones.

Se realizaron pruebas cruzadas con los usuarios que están presentes en la base de datos e indicaron resultados satisfactorios al momento de su utilización.

6.2. Trabajos futuros

Para dar continuidad al trabajo aquí expuesto se propone analizar otras diferentes metodologías y técnicas de reconocimiento de patrones que puedan aportar una mejor forma de detección de intrusos en los sistemas informáticos, como por ejemplo la obtención de un sistema experto cuyas reglas se obtienen con la aplicación de algoritmos genéticos.

Otra posible línea de investigación sería el estudio de todas las variables que intervienen dentro del proceso de uso de sistemas informáticos, como señales que emite cada estación de trabajo, interacción con el medio, movimientos físicos por parte del usuario y otros comportamientos que faciliten información para determinar un perfil exacto del usuario en el sistema informático. Además de la creación y aplicación de técnicas que permitan diseñar y evaluar el comportamiento humano.

Por otra parte la aplicación de equipos informáticos y herramientas que faciliten a la identificación de usuarios, como también nuevas metodologías de minería de datos que permitan la obtención de la información de una manera más fácil y confiable.

Bibliografía

[1]. Anderson, James A.. An introduction to neural networks. Cambridge, Mass.: MIT Press, 1995.

[2]. Bramer, Max, Frans Coenen, and Miltos Petridis. Research and Development in Intelligent Systems XXIV. Dordrecht: Springer, 2008.

[3]. Brodley, Carla E., and Paul E. Utgoff. Multivariate decision trees. Amherst, Mass.: University of Massachusetts at Amherst, Dept. of Computer Science, 1992.

[4]. Cantos A J ,Santos Peñas M, "Classification of plasma signals by genetic algorithms", Fusion Science and Technology, 58, 706-713, Oct 2010.

[5]. Casal , Jordi , and Enric Mateu. "Universidad Autónoma de Barcelona." TIPOS DE MUESTREO. Universitat Autònoma de Barcelona, n.d. Web. 26 Sept. 2011. <minnie.uab.es/~veteri/21216/TiposMuestreo1.pdf>.

[6]. Chatterjee, Samprit, and Ali S. Hadi. Regression analysis by example. 4th ed. Hoboken, N.J: Wiley-Interscience, 2006.

[7]. Chen, Ming, and Jong Soo Park. Efficient data mining for path traversal patterns. Yorktown Heights, N.Y.: IBM T.J. Watson Research Center, 1995. Print.

[8]. Cook, Diane J., and Lawrence B. Holder. Mining graph data. Hoboken, N.J.: Wiley-Interscience, 2007.

[9]. Dasarathy, Belur V.. Nearest neighbor pattern classification techniques. Los Alamitos, Calif.: IEEE Computer Society Press, 1990.

[10]. Dormido-Canto S, Dormido R, Duro N, Farias G, Martin J A, Sánchez J, Santos Peñas M, Vargas H ,Vega J, "Dynamic Clustering and Modeling Approaches for Fusion Plasma Signals",IEEE Transactions on Instrumentation and Measurement,58, 9, 2969-2978, Septiembre 2009

[11]. Dormido R ,Dormido-Canto S, Duro N, Farias G, Pajares G, Pereira A, Portas A, Sánchez J, Santos M, Sánchez E, Vega J, "Data mining technique for fast retrieval of similar waveforms in fusion massive databases",Fusion Engineering and Design,83, 132-139, 2008

[12]. Duro N, Dormido R, Dormido-Canto S, Farias G, Pajares G, Sánchez J, Santos M, Vega J, "Automated clustering procedure for TJ-II experimental signals",Fusion Engineering and Design,81, 1987-1991, 2006

[13]. Böhlen, Michael H.. Visual data mining. Berlin: Springer, 2008.

[14]. Farias G, López V, Santos Peñas M, "Making decisions on brain tumour diagnosis by soft computing techniques", Soft Computing, 14, 1287-1296, 2010

[15]. Fayyad, Usama M.. Advances in knowledge discovery and data mining. Menlo Park, Calif.: AAAI Press :, 1996.

[16]. Feigenbaum, E. A.. Expert Systems in the 1980s. Inglaterra: Pergamon-InfoTech, Maid Enhead, 1980.

[17]. Gallestey, Jorge, and José M Riestra. Arboles de regresión y otras opciones metodológicas aplicadas a la predicción del rendimiento académico. Habana: Centro de Investigaciones y Referencia de Aterosclerosis de La Habana, 2002.

[18]. Gilmer, B. von Haller, and Dionisio Pérez. Psicología general. 2. ed. México: HARLA, 1974.

[19]. Gurney, Kevin N.. An introduction to neural networks. London: UCL Press, 2002.

[20]. Jarke, Matthias, and Maurizio Lenzerini. Fundamentals of Dta Wrehouses. 2e éd. revue et augmentée. ed. Berlin: Springer, 2003.

[21]. Kimball, Ralph. The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. New York: Wiley, 1998.

[22]. Kitagawa, H.. Database systems for advanced applications 15th international conference, DASFAA 2010, Tsukuba, Japan, April 1-4, 2010 : proceedings. Berlin: Springer, 2010.

[23]. Lind, Douglas A., William G. Marchal, and Samuel Adam Wathen. Estadística aplicada a los negocios y la economía. 13a ed. México, D. F.: McGraw Hill, 2008.

[24]. López, César, and Daniel González. Minería de datos : técnicas y herramientas. 1a. ed. Madrid: Paraninfo Cengage Learning, 2008.

[25]. Macedo Cruz Antonia, Pajares Martinsanz, Santos Peñas Matilde. "Clasificación no supervisada con imágenes a color de cobertura terrestre.". ISSN (Versión impresa): 1405-3195. Colegio de Postgraduados, 6 Aug. 2012. Web. 6 Sept. 2012. <http://redalyc.uaemex.mx/pdf/302/30215554009.pdf>.

[26]. Morgan, J. N., and J. A. Snoquist. Problems in the analysis of survey data, and a proposal. Michigan: Journal of the American Statistical Association, 1963.

[27]. Orallo, José, and Ma. José Quintana. Introducción a la minería de datos. Madrid: Pearson Educación, 2004.

[28]. Pajares G. y Santos M., Ingeligencia Artificial e Ingeniería del conocimiento, ed. RA-MA, 2005.

[29]. Pasha, G.R, and Ali. Shah. An alternative approach to the Kalman's scheme of identification in the presence of multicollinearity. Pakistan.: Journal of Research (Science), Bahauddin Zakariya University, Multan, 1996.

[30]. Pliego, F. Javier. Introduccion a la estadistica economica y empresarial teoria y practica. Madrid, Espana: Paraninfo, 2004.

[31]. Quinlan, J. R.. C4.5: programs for machine learning. San Mateo, Calif.: Morgan Kaufmann Publishers, 1993.

[32]. Rossiter, D G . "Lecture Notes: Land Evaluation." ITC - Faculty of Geo-Information Science and Earth Observation . N.p., n.d. Web. 11 June 2012. <http://www.itc.nl/~rossiter/teach/le/s494toc.htm>.

[33]. Russell, Stuart J., Peter Norvig, and Ernest Davis. Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River: Prentice Hall, 2010.

[34]. Soukup, Tom, and Ian Davidson. Visual data mining techniques and tools for data visualization and mining. New York: Wiley, 2002.

[35]. Sreerama, K., and Murthy. Automatic construction of decision trees from data: a multi-disciplinary survey. Ontario: Murthy, 1998.

[36]. Vapnik, Vladimir Naumovich. The nature of statistical learning theory. 2nd ed. New York: Springer, 2000.

[37]. Viñuela, Pedro, and Inés M. León. Redes de neuronas artificiales: un enfoque práctico. Madrid: Prentice Hall, 2004.

[38]. Witten, Ian H., and Eibe Frank. Data mining: practical machine learning tools and techniques. 2. ed. San Francisco, CA [u.a.: Kaufmann [u.a.], 2005.

[39]. Zaki, Mohammed J.. Advances in knowledge discovery and data mining 14th Pacific-Asia conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010 : proceedings. Berlin: Springer, 2010.