ANÁLISIS DE TRÁFICO CURSADO Y EL ESTADO DE LOS EQUIPOS DE RED A PARTIR DE TÉCNICAS DE MINERÍA DE DATOS EN LA RED LAN DE LA UNIVERSIDAD PONTIFICIA BOLIVARIANA (UPB)

GUSTAVO ADOLFO SERNA LÓPEZ

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MEDELLÍN

2017

ANÁLISIS DE TRÁFICO CURSADO Y EL ESTADO DE LOS EQUIPOS DE RED A PARTIR DE TÉCNICAS DE MINERÍA DE DATOS EN LA RED LAN DE LA UNIVERSIDAD PONTIFICIA BOLIVARIANA (UPB)

GUSTAVO ADOLFO SERNA LÓPEZ

Trabajo de grado para optar al título de Magister en Tecnologías de la Información y la Comunicación

Asesor

CLAUDIA CARMONA RODRÍGUEZ

Magister en Ingeniería de Telecomunicaciones

UNIVERSIDAD PONTIFICIA BOLIVARIANA

ESCUELA INGENIERÍAS

FACULTAD DE INGENIERÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

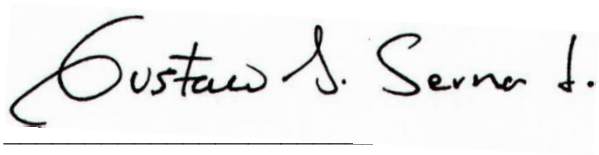
MAESTRÍA EN TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIÓN

MEDELLIN

2017

DECLARACIÓN ORIGINALIDAD

"Declaro que esta tesis (o trabajo de grado) no ha sido presentada para optar a un título, ya sea en igual forma o con variaciones, en esta o cualquier otra universidad". Art. 82 Régimen Discente de Formación Avanzada, Universidad Pontificia Bolivariana.

FIRMA AUTOR (ES) 

Medellín, 10 de Marzo de 2017.

AGRADECIMIENTOS

A Dios por haberme permitido llegar hasta este punto y haberme dado salud y constancia para lograr mis objetivos.

A mis padres Gustavo Serna Ramos y Eva López Fayad por el apoyo incondicional en todo momento, por sus consejos, valores y motivación a seguir siempre adelante y superar los obstáculos en el camino.

A mi directora Claudia Carmona por su apoyo antes y durante la realización del proyecto, por sus consejos, motivación y dedicación.

A los profesores: Roberto Hincapié y Ana Oviedo por el tiempo dedicado en las asesorías durante la realización del proyecto.

Al personal del CTIC por el apoyo brindado durante la realización del proyecto.

CONTENIDO

LISTA DE FIGURAS

LISTA DE ECUACIONES

LISTA DE TABLAS

GLOSARIO

Area ROC: *Receiver-Operating Characteristic* ó también Característica Operativa Relativa, comparación de dos características operativas (VPR= Razón de Verdaderos Positivos y FPR=Razón de falsos positivos) según se cambien el umbral para la decisión.

Dataset o vista minable: Conjunto de datos minables, los cuales se obtuvieron de la selección, limpieza, integración y formateo de una colección de datos inicial.

Weka 3.8.0: *Waikato Environment for Knowledge Analysis*, en español «entorno para análisis del conocimiento de la Universidad de Waikato»).

RESUMEN

En este trabajo se muestra el resultado de la aplicación de Inteligencia Operativa de Red utilizando una metodología de minería de datos en el área de redes de telecomunicaciones, donde a través de una prueba piloto y una serie de experimentos, se pretende realizar un análisis que permita identificar relaciones del desempeño de la red con el tráfico cursado y el estado de los equipos de red con el fin de determinar las causas que pueden llevar a fallas o a una degradación del desempeño de la red LAN en la UPB.

PALABRAS CLAVE: Técnicas de minería de datos; algoritmos supervisados; algoritmos no supervisados; estadísticas de red de telecomunicaciones; tráfico de red.

ABSTRACT

In this paper shows the result of the application of Network Operational Intelligence using a methodology of data mining in the telecommunication networks area, where through a pilot test and a series of experiments, we intend to conduct an analysis to let identify relationships between network performance with the traffic carried and the state of network equipment in order to determine the causes that can lead to failure or degradation of performance LAN in the UPB.

KEY WORDS: Data mining techniques; supervised algorithms; unsupervised algorithms; telecommunications network statistics; network traffic.

1 INTRODUCCIÓN

Las redes de telecomunicaciones soportan aplicaciones y servicios que son de carácter estratégico e indispensable para las organizaciones. (Grajales Bartolo, 2011) & (Neves, Leitao, & Almeida, 1995). De acuerdo al grado de complejidad de su estructura y distribución, los inconvenientes y fallas que suelen presentarse corresponden a incertidumbres ante situaciones irregulares conexas al tráfico y al rendimiento inadecuado de los dispositivos de red referente al consumo de recursos. (Grajales Bartolo, 2011).

Evaluar y optimizar el desempeño de una red de telecomunicaciones envuelve al estudio del tráfico de la red y es esencial en el proceso determinar puntos críticos, medir la calidad de la red a través de programas de monitoreo de servicios, evidenciar resultados a problemas de direccionamiento con los *routers*, problemas de conexión, problemas relacionados con bajos índices en los parámetros de: velocidad de conexión, tasa de transmisión, ancho de banda, tiempos de respuestas y en la disponibilidad del servicio (Rivero G, 2006), (António, Salvador, & Valadas, 2006), (Villadango & Magaña, 2001). Por ello, la planeación de la capacidad y control de una red de telecomunicaciones es una actividad primordial para las organizaciones actuales (Hernández Suarez, Martínez Sarmiento, & Escobar Díaz, Modelamiento y pronósticos de tráfico correlacionado, 2008).

Es posible complementar el proceso de Planeación de la Capacidad de Red (NCP) por medio de la Inteligencia Operativa en redes de telecomunicaciones a través de metodologías de minería de datos, realizando análisis del tráfico cursado e identificando variables correlacionadas que conlleven al descubrimiento automático de conocimiento (Gildardo, 2006).

En este trabajo se plantea el uso de Inteligencia Operativa en una Red para la descripción y aplicación de una metodología de minería de datos que permita identificar relaciones del desempeño de la red con el tráfico cursado y el estado de los equipos de red. Se llevó a cabo en la red LAN de la Universidad Pontificia Bolivariana (UPB) sede Medellín, tomando los datos del tráfico y estadísticas generadas por los nodos de red del bloque 22, el cual hace parte de las sedes al exterior de la Universidad.

2 PLANTEAMIENTO DEL PROBLEMA

2.1 PROBLEMA

Acorde al nivel de complejidad y distribución, las redes de telecomunicaciones generan grandes volúmenes de datos operacionales adicionales al tráfico como son: las estadísticas generadas de uso y los datos que son producidos por las señales de alarmas de los equipos de redes (Gildardo, 2006).

En el mercado se han desarrollado diversas aplicaciones para la captura de tráfico y el monitoreo del estado de la red basados en protocolos de gestión, los cuales generan grandes volúmenes de información dificultando su interpretación por parte del personal encargado para tal fin. Con crecimiento del volumen de datos almacenados resulta complicado realizar análisis y clasificaciones para anticiparse a problemas relacionados con el modelamiento del tráfico y también con la asignación del hardware para optimización de recursos con respecto a fallas.

A través de metodologías de minería de datos es posible obtener modelos y desplegar soluciones para resolver diversos requerimientos de la Ingeniería de tráfico como pueden ser: (Landa Laredo, 2009) (Neves & Leiato, 1995).

- Determinar el ancho de banda disponible.
- Detectar patrones de congestión de tráfico.
- Clasificación de tráfico de Internet.
- Predicción del uso de recursos como es el caso de los enlaces.
- Diagnósticos de fallas (clasificador).
- Mejoramiento de los algoritmos de enrutamiento.

2.2 JUSTIFICACIÓN

Es de interés para la Universidad Pontificia Bolivariana la toma de mejores decisiones con respecto a problemas relacionados con los nodos y el tráfico de red manera alterna a la metodología actual, dado que el reemplazo y adquisición de equipos sólo se realiza al momento de presentar fallas o deterioro total, sin análisis de situaciones concernientes al tráfico y a consumo de recursos.

Teniendo en cuenta que la importancia de aplicar metodologías de minería de datos está determinada por la capacidad de encontrar relaciones entre los datos que no son identificables fácilmente en grandes cantidades de información (Reyes Saldaña & García Flores, 2005), la Universidad a través del Centro de Tecnologías de Información y Comunicación (CTIC), busca concluir y proponer cambios que podrían envolver la estructura, configuración y reemplazo de equipos, siendo necesario realizar una clasificación de estos últimos por estado de acuerdo a ciertos factores. Para ello es fundamental determinar el comportamiento del tráfico y de los dispositivos de red, partiendo de identificar relaciones existentes entre el tráfico cursado y las estadísticas que generan los nodos por medio de la información recolectada a través de la gestión de red (Alvarez Menendez, 2008), (Reyes Saldaña & García Flores, 2005) && (Vicente Altamirano, 2003).

3 OBJETIVOS

3.1 OBJETIVO GENERAL.

Analizar relaciones del desempeño de la red con el tráfico cursado y el estado de los equipos de red a partir de técnicas de minería de datos para hallar problemas causantes de fallas o disminución de desempeño de la red LAN del bloque 22 en la UPB a través de una prueba piloto.

3.2 OBJETIVOS ESPECÍFICOS

- Describir los requerimientos del servicio de la red LAN al interior de la Universidad.
- Definir la preparación y dataset de las estadísticas de los dispositivos de red y las variables del tráfico cursado en la red LAN del bloque 22 de la Universidad Pontificia Bolivariana.
- Aplicar las técnicas de minería de datos a los dataset o conjunto de datos obtenidos mediante una prueba piloto en el bloque 22 de la UPB.
- Evaluar los resultados obtenidos en la prueba piloto determinando si son pertinentes para la integración de resultados a la toma de decisiones.

4 MARCO REFERENCIAL

4.1 MARCO CONTEXTUAL

El campus universitario de la Universidad Pontificia Bolivariana está conformado por bloques donde se agrupan usuarios de diferentes tipos y que se conectan a internet y a los diferentes servicios de tecnologías de información que la Universidad presta a través del CTIC (Centro de Tecnologías de Información y Comunicación). Actualmente la Universidad adicional a los bloques que se encuentran al interior del Campus Laureles posee algunas sedes fuera de éste y a las cuales debe ofrecerles los mismos servicios que se ofrecen al interior del Campus.

El proyecto propuesto se llevará a cabo en la red LAN de la Universidad Pontificia Bolivariana sede Medellín, tomando los datos del tráfico y estadísticas generadas por los nodos de red del bloque 22.

En el bloque 22 se encuentran ubicados docentes de diferentes escuelas de la Universidad y con una alta dedicación a labores de investigación lo que implica que someten a la red a procesos como altas tasas de descarga de contenido, servicios de videoconferencia y generación de *streaming* adicionales a los servicios de descarga de correo, consultas al sistema de información y bases de datos bibliográficas. En la actualidad durante calendario académico se tiene un consumo de ancho de banda de 1Gbps y se cuentan con aproximadamente 50 usuarios que acceden desde el bloque.

4.2 MARCO CONCEPTUAL

La inteligencia Analítica posibilita satisfacer la necesidad de modelar ciertos sistemas, tanto naturales como artificiales y llevarlos a un punto en donde se puedan aplicar procesos que permitan su descripción, optimización matemática, simularlos, entre otros. De acuerdo a (Gildardo, 2006) esta está conformada por las técnicas estadísticas como las series de tiempo y la correlación de *spearman*, y por la minería de datos.

Dentro de las técnicas estadísticas, que comprenden las medidas tomadas de una muestra de datos se encuentran las series de tiempo, que hacen parte de la estadística descriptiva aplicada a datos discretos. Se considera la cantidad de veces que un valor puede darse. Aplicada más que todo a estadísticas producidas por elementos o nodos de una red puesto que se pueden representar como series de tiempo puesto que son valores discretos de una variable (Gildardo, 2006). Permiten ilustrar cómo se comporta una variable aleatoria la cual varia con el tiempo para poder hacer pronósticos de la misma llevando a cabo modelos estadísticos correlacionados para tal fin. La metodología más utilizada es las de Box –Jenkins (Hernández Suarez, Salcedo Parra, & Pedraza Martínez, 2008). Los modelos de tráfico correlacionados usados en las series de tiempo son: Modelo AR (Auto regresivo), Modelo MA (promedios móviles), modelo ARMA (auto regresivo y promedio móvil) y modelo ARIMA (auto regresivo e integrado de promedio móvil) (Hernández Suarez, Salcedo Parra, & Pedraza Martínez, 2008) & (Torres Álvarez, Hernandez, & Predraza, 2011). Por su parte, la correlación de *spearman* es una buena técnica al momento de tener gran cantidad de variables a contemplar y se necesita disminuir el número a analizar a través de la identificación de las que no sean redundantes, midiendo así el nivel o grado de relación entre dos variables cuantitativas. Esto es muy significativo para el proceso minería de datos a la hora de conocer los datos.

La minería de datos puede definirse inicialmente, como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos (Pérez López, 2007). Es una tecnología que consiste en algoritmos que extraen conocimiento de grandes bases de datos que acumulan la historia de las actividades de las organizaciones. El conocimiento tiene como finalidad prevenir a los responsables de tomar decisiones sobre situaciones interesantes, anomalías, e incluso amenazas no detectadas con anticipación (Martinez Luna, 2011). Son herramientas que identifican patrones (tendencias, regularidades, correlaciones) existentes en las bases de datos. Sobresalen modelo descriptivo (indirecto): asociación, segmentación y modelo predictivo (directo): clasificación, estimación (Hernández Suarez, Salcedo Parra, & Pedraza Martínez, 2008) & (Torres Álvarez, Hernandez, & Predraza, 2011).

Como lo menciona (Gildardo, 2006), la minería de datos se podría aludir desde dos puntos de vista: minería de datos directa para dar esclarecimiento o categorizar algún campo en particular como entrada o respuesta, como es el caso en las redes de telecomunicaciones la predicción de la variable utilización de los enlaces en una red de datos, así pues que la entrada es un conjunto de variables dependientes y la salida la predicción de la utilización del enlace; y minería de datos indirecta para descubrir patrones o similitudes entre grupos de registros sin el uso de un campo en particular o colección de clases predefinidas. Algunas de las técnicas de minería de datos según (Montero Lorenzo, 2008) descriptivas (Indirectas): *clustering* (agrupación) y segmentación, asociación secuencial, asociación, redes neuronales; predictivas (directas: clasificar, estimar, predicción): árbol de decisión, series de tiempo, inducción de reglas, redes neuronales, máquinas de soporte vectorial.

El descubrimiento de conocimiento en datos busca identificar patrones que no son tan evidentes en un proceso de análisis normal. Se utilizan técnicas descriptivas como es la de asociación y agrupación, buscando identificar relaciones dentro del conjunto de datos y de esta manera obtener reglas que permitan relacionar a los diferentes atributos de ellas (Reyes Saldaña & García Flores, 2005).

En las técnicas de minería de datos descriptivas también llamadas clasificación no supervisada, no se tiene información previa sobre la existencia de clases entre los objetos. Se busca principalmente las posibles clases existentes en la matriz de datos partiendo de la agrupación de objetos con características y/o comportamientos similares. Entre las técnicas no supervisadas, más utilizadas son: clúster, dendogramas, reglas de asociación y redes neuronales (Henao Ríos, 2012).

El clúster**,** se considera de las metodologías mayormente empleadas en la clasificación no supervisada. El clúster obedece a una descripción cuantitativa de la similaridad entre pares de objetos o grupos en formación llamados dendogramas y de lo que representa esa medida para la conformación de los grupos (Henao Ríos, 2012).

A través de las reglas de asociación, se busca básicamente relaciones o afinidades entre conjunto de atributos. El soporte de una regla de asociación Se basa en dos conjuntos: premisa y conclusión (Reyes Saldaña & García Flores, 2005). El análisis de asociación usa el algoritmo apriori, el cual básicamente consiste en la búsqueda del descubrimiento de las reglas de asociación que muestran condiciones de pares <atributo – valor> que se presentan de manera repetida en un conjunto de detección automática de reglas de asociación (Cartegnova, 2005). (Naranjo Cuervo & Sierra Martínez, 2009).

Las reglas de asociación se miden de acuerdo a (Reyes Saldaña & García Flores, 2005):

- Soporte, es el número de ocurrencias del grupo de elementos en la base de datos de transacciones. Soporte de ($A \rightarrow C$) = Sop($A \cup C$).
- Confianza, es el porcentaje de transacciones que contienen a todos los elementos que componen la regla. Confianza de ($A \rightarrow C$)= Sop($A \cap C$)/Sop(A).

De las técnicas de minería de datos, las redes neuronales (RN) son aplicadas a predicción, pronóstico, clasificación y *clustering* en áreas de telecomunicaciones. Es importante tener en cuenta que entre menos variables se tienen, el modelo tendrá un mejor comportamiento. (Gildardo, 2006). Para el caso de algoritmos no supervisados hacen uso de los mapas auto-organizativos (SOM), que brindan la posibilidad de representar los datos de múltiples dimensiones en dimensiones más pequeñas. A éste algoritmo, su aprendizaje no supervisado le permite realizar la clasificación de los datos sin necesidad de ciertos controles externos (Naranjo Cuervo & Sierra Martínez, 2009).

Técnicas de minería de datos, como es el caso de la lógica difusa (*fuzzy logic*) son usadas como herramienta de descubrimiento de conocimiento basada en el lenguaje natural, se fundamenta en usar términos lingüísticos, pertenecientes al lenguaje natural, para expresar información y conocimiento oculto en colecciones de objetos potencialmente de gran tamaño. Su estructura es basada en reglas mezclando el modelo predictivo y el descriptivo. (Montesino Pouzols, 2009).

De acuerdo a (Vieira Braga, Ortiz Valencia, & Ramirez Carvajal, 2009) el ciclo de vida de un proceso de minería de datos es:

1. Selección de la metodología
2. Definición del problema
3. Preparación de los datos
4. Modelado de los datos
5. Evaluación
6. Despliegue del modelo clasificador de conocimiento

Como lo mencionan los autores (Casilari et al., n.d.), (Moine, Haedo, & Gordillo, 2011), (Valencia Zapata, 2008) & (Britos, 2005), CRISP-DM (*Cross Industry Standard Process for Data Mining*) es la metodología más utilizada en minería de datos. Las metodologías SEMMA (*Sample, Explore, Modify, Model, Assess*) y KDD (*Knowledge Discovery in Databases*) únicamente suministran una orientación

universal del trabajo a realizar en cada etapa; CRISP-DM permite mayor exploración aplicado a las tareas y actividades que se llevan a cabo en cada una de las etapas.

En las organizaciones, es vital tener en cuenta el proceso de planeación de la capacidad de red (PCR). Muchas veces resulta bastante complicado de lograr si no se cuenta con la metodología y las herramientas necesarias, debido a la propiedad distribuida de las redes de telecomunicaciones y a la inmensa cantidad de variables necesarias a incluir en el proceso a estimación del tráfico esperado que la red tendría la capacidad de soportar (Gildardo, 2006).

La gestión de red se vale de muchas tareas tales como: monitoreo, diagnósticos, priorización de flujo y seguridad. Estas son esenciales para obtener la clasificación del tráfico a través de técnicas de minería de datos. (Shinde S & Abhang, 2012). El protocolo agente SNMP (*Simple Network Management Protocol*) es medio para adquirir la información de estadísticas de los elementos de una red de telecomunicaciones y al realizar este proceso contribuye a generar tráfico por lo cual hay que tener presente las distintas posibilidades de configuración proporcionadas por las herramientas de gestión de red (Escriche Fernández, 2011) & (García & Salcedo, 2010).

El tráfico de red es la porción de datos que fluye a través de una red de datos. Estos datos abarcan descripciones de las llamadas realizadas, las estadísticas de los dispositivos, de software y los servicios de red de telecomunicaciones. Dicho tráfico es producido por los dispositivos de red, la cantidad de usuarios a los cuales se les suministra servicios y la multiplicidad de servicios que provee la red (Gildardo, 2006) & (Alvarez Menendez, 2008). Los estudios referentes al tráfico en redes IP se fundamentan en la captura o registro de la información contenida en la trama (*frame*) o datagrama IP que se transfiere por un segmento red LAN por un enlace WAN (Rivero G, 2006). A través de aplicación de modelos matemáticos se puede comprender las relaciones presentes ente la capacidad de red, demanda de servicios por parte de los usuarios y el nivel de desempeño que una red tiene la posibilidad de alcanzar. En lo anterior se basa la teoría de tráfico. El propósito es el

desarrollo de modelos que permitan anticiparse al impacto de la carga atribuida por los diversos servicios y aplicaciones sobre los elementos de red, así pues, que se tenga capacidad de evaluar la calidad de servicio (QoS) ofrecida (Alzate & Peña).

La calidad de servicio (QoS) en una red de telecomunicaciones está determinado por el rendimiento de extremo a extremo de los servicios electrónicos de la misma manera que es percibido por el usuario final y tiene como medidas: el retardo, la variación del retardo y la pérdida de paquetes (Rivero G, 2006).

De acuerdo a la clase de aplicaciones y de servicios que se suministren en la red de telecomunicaciones, dependerá el tráfico que se produzca y es relevante identificar las variables que se relacionan con cada servicio, al mismo tiempo que se debe contemplar otras sin tener en cuenta la clase de aplicación (Gildardo, 2006). Una variable para determinar el nivel tráfico que fluye en una red de datos puede ser: nivel de utilización (medición) de los enlaces entre las redes de datos. Es el nivel de flujo de tráfico a través de la red (Gildardo, 2006).

Según (Gildardo, 2006) Las siguientes variables son relevantes a la hora de proporcionar información sobre el desempeño de los dispositivos de red:

- Porcentaje de utilización de la CPU
- Tráfico enviado y recibido (bits/sec)
- Tipo de procesamiento de las tareas
- Carga de trabajo (solicitudes/sec) y (sesiones/sec)

En las redes de telecomunicaciones, las alarmas son producidas por los dispositivos de la red a través de mensajes cada vez que se presente una inconsistencia o fallos, haciendo alusión a escenarios irregulares y que puede ser transparente al usuario (Hatonen, Klemettinen, Mannila, Ronkainen, & Toivonen, 1996). Cuando se generan alarmas producidas por los dispositivos de red, estas están determinadas por un grupo de variables las cuales detallan la irregularidad a la que hacen

referencia desde el tiempo o software generador y el contenido refiriéndose a los datos acerca del fallo (Gildardo, 2006).

En el caso de las estadísticas que son generadas por el uso de los dispositivos, cada uno de ellos individualmente, produce sus particulares estadísticas sobre cómo se comporta el tráfico (Gildardo, 2006).

4.3 MARCO LEGAL.

Actualmente no existe legislación para la gestión de redes. Aunque hay estándares internacionales para la gestión de infraestructura y de servicios como ITIL (*Information Technology Infraestructure Library*), eTOM (*Enhanced Telecom Operations Map*), COBIT (*Control Objectives Control Objectives for Information and related Technology*).

Por la naturaleza del proyecto y las capturas de información se debe este trabajo se fundamenta en garantizar la protección intelectual y de datos a través de las siguientes leyes:

Ley 1581 de 2012, Marco general de la Protección de datos en Colombia:" desarrollar el derecho constitucional que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos, y los demás derechos, libertades y garantías contusiónales a que se refiere el artículo 15 de la Constitución Política; así como el derecho a la información consagrado en el artículo 20 de la misma". (Molano Vega, 2013).

Leyes de Propiedad Intelectual, Constitución Política de 1991, articulo 61: "El Estado protegerá la propiedad intelectual por el tiempo y mediante las formalidades que establezca la ley". (Molano Vega, 2013).

4.4 ESTADO DEL ARTE

El auge y gran velocidad con las que evolucionan las redes de telecomunicaciones para poder ofrecer de manera satisfactoria la demanda de múltiples servicios de usuarios, implican la necesidad de mayor ancho de banda, mejor QoS, disminuir la latencia en la red, aumentar la calidad de transmisión ante aplicaciones con grades volúmenes de datos en transmisiones, multimedia, video conferencias y el envío de información en tiempo real. Pero con el crecimiento y el volumen de datos almacenados resulta complicado realizar análisis, clasificaciones o anticiparse a problemas relacionados con el tráfico, los niveles de utilización de los enlaces y así como también con el hardware de acuerdo a fallas y a la asignación y optimización de recursos. Este tipo de problemas que con técnicas convencionales no han sido tan exitosamente resueltos, pueden encontrar un alto grado de alivio en la minería de datos aplicando técnicas descriptivas y predictivas que permiten resolver muchas de las tareas de la Ingeniería de tráfico como pueden ser: (Landa Laredo, 2009) & (Neves et al., 1995).

- Determinar el ancho de banda disponible
- Detectar patrones de congestión de tráfico
- Clasificación de tráfico de Internet
- Predicción del uso de recursos como es el caso de los enlaces
- Diagnósticos de fallas (clasificador)
- Mejoramiento de los algoritmos de encaminamiento

De acuerdo a (Valencia Zapata, 2008), (Britos, 2005) & (Moine, Haedo, & Gordillo, 2011) la metodología más usada en el proceso de minería de datos CRISP-DM.

Para evaluar el desempeño de las redes de telecomunicaciones es necesario el estudio de tráfico de la red y es inherente al proceso: determinar puntos críticos, medir la calidad de la red a través de programas de monitoreo de servicios, evidenciar resultados a problemas de direccionamiento con los *routers*, problemas

de conexión, problemas de bajos índices en los parámetros de: velocidad de conexión, tasa de transmisión, ancho de banda, tiempos de respuestas y en la disponibilidad del servicio (Rivero G, 2006), (Nogueira, Salvador, & Valadas, 2006) & (Villadango & Magaña, 2001).

A través del uso de técnicas de inteligencia artificial descriptivas, en las redes de telecomunicaciones, se logra la obtención de modelos de los datos que permitan resolver problemas en diversos ámbitos. En el caso de seguridad en redes, se obtienen las características más importantes de las tramas involucradas para realizar una selección, procesamiento y una posterior clasificación, teniendo presente la aplicación de técnicas para reducción de la dimensionalidad a través del análisis de componentes principales (PCA) y clasificación basada en redes neuronales (ANN). Esto a través de herramientas de monitoreo como Tcpdump, Wireshark y herramientas de minería como Weka para comparar los resultados con distintos algoritmos utilizados (Jamuna & Vinodh Ewards, 2013).

Los modelos descriptivos de minería de datos como es las redes neuronales a través del algoritmo SOM (mapas auto-organizados) para entrenamiento de aprendizajes no supervisados en detección de intrusiones de red tienen como fin las clasificaciones del tráfico de la red en conexiones normales y de ataques, para ello se usó un dataset de Grupo de Tecnologías de Sistemas de Información (IST) del laboratorio Lincoln del Instituto Tecnológico de Massachusetts (LL-MIT), con la cooperación de la Agencia Avanzada de Defensa (DARPA ITO), técnicas de reducción de dimensionalidad como PCA y la herramienta de minería Weka. (De la Hoz, De la Hoz, Ortiz, & Ortega, 2012).

En la de detección de fraudes en telecomunicaciones móviles, la inteligencia analítica es útil para realizar análisis y perfilados de llamadas de los usuarios durante un periodo de tiempo con técnicas de minería de datos no supervisadas, aplicando redes neuronales con algoritmos de mapas auto-organizados (SOM) y larga memoria a corto plazo (LSTM) recurrentes con el fin de llevar a cabo una extracción de datos descriptivos sobre los patrones de llamadas de los usuarios.

Para ello se usó grabaciones durante 6 meses con 227,318 llamadas originadas de un conjunto de 500 suscriptores enmascarados (Adeniyi Abidogun, 2005).

Debido a que con el tiempo los servicios Web son cada vez más complejos y dinámicos, traen consigo también nuevas posibilidades a los atacantes e intentos de ataque que por lo general se registran en los registros coleccionados en un servidor Web Apache. Estos registros pueden ser analizados con la elaboración del análisis de clúster después de obtener la representación de pocas dimensiones que normalmente serial reducidas de acuerdo a técnicas de reducción de la dimensionalidad como mapas de difusión DM y PCA, y para minería algoritmos *K-means* (Juvonen & Sipola). En la aplicación conjunta de diversas técnicas de minería de datos especialmente para realizar *Web usage mining*, hay autores como (Cernuzzi & Molas) que realizan una propuesta para la descripción del comportamiento de los usuarios de un portal Web, tomadas de datos reales del portal de Rieder Internet. Esto a través de reglas de asociación con la intención de encontrar relaciones entre las reglas. Además, acompañado de otra técnica como es *clustering* para obtener conjuntos de datos con características o comportamientos similares de acuerdo a la cantidad de pruebas seleccionadas.

El uso de algoritmos no supervisados, permite descubrir automáticamente en una red de computadoras comprometidas infectadas con código malicioso que se pueden controlar de forma remota (también llamadas *botnets*), (Lu, Tavallaee, & Ghorbani) proponen un marco de trabajo en el que toman 100 millones de flujos capturados durante tres días consecutivos de un proveedor de red Wifi ISP. De esta manera, partiendo primero de la clasificación del tráfico de la red en diferentes grupos de aplicaciones mediante el uso de firmas de carga útil y un algoritmo de agrupamiento (*clustering*) y asociación-cruzada (*cross-association*) y la herramienta de minería Botminer, para después en cada grupo obtenido se analizan las características temporales frecuentes de los flujos que conducen a la diferenciación de canales maliciosos creados por los robots de tráfico normal a la generada por los seres humanos. El resultado de la propuesta se desarrolla exitosamente arrojando

dos tipos de flujos de aplicación *botnet*, con altas tasas de detección y bajas tasas de bajas alarmas (Lu, Tavallaee, & Ghorbani).

Por otra parte, el uso conjunto de algoritmos de inteligencia artificial supervisados y no supervisados en el análisis de tráfico TCP/IP consiste en modelar en forma efectiva el tráfico de una organización, además el reducir en forma esencial el porcentaje de falsos positivos mientras se conserva un nivel prudente de descubrimiento de anomalías. Este diseño es basado en un conjunto de "*Self-Organizing Maps*" (SOM) para el modelado del tráfico y en el uso de "*Linear Vector Quantization*" (LVQ) para la clasificación definitiva de los paquetes de tráfico. Entonces principalmente lo que se procede a realizar en orden es *clustering* de características (con algoritmos de: *clustering*, K-NN), luego se hace una agrupación y por ultimo una clasificación. Para ello se hace uso de la herramienta Snort para extracción y codificación de atributos y la herramienta SMO_PAK para el modelado de los datos, los cuales se tomaron de DARPA 1998 (Couche, Steine, San Vicente, & Ferreira). Cuando se quiere es una clasificación de tráfico basado en un enfoque semi-supervisado se hace uso del algoritmo *K-Means* de cluster llamado *Seeded-KMeans* para averiguar el valor apropiado de los clústeres y luego usar el proceso el diseño de clasificadores de tráfico de red basados en las etiquetas de los grupos conformados. Los autores proponen marcos de trabajo para los conjuntos de datos tomados de red IRIS, DARPA, flujos de datos exportados por enrutadores cisco a través de herramientas Cisco Netflow, tcpdump y uso de algoritmos de agrupación como es el caso de *K-Means* y las herramienta WEKA de minería (Shinde S & Abhang, 2012), (Gu, Zhang, Chen, & Du, 2011), (Erman, Arlitt, & Mahanti), (Munz, Li,, & Carle).

Metodologías como las técnicas de inteligencia analítica estadísticas basadas en series de tiempo, buscan "predecir la carga futura de enlaces individuales o rutas aisladas y así anticipar la aparición de situaciones críticas y establecer una mejor planificación del crecimiento de la infraestructura de red mediante los resultados predictivos obtenidos" (Escriche Fernández, 2011). Sin embargo, la inteligencia analítica tiene ramas como la minería de datos que permiten en un mayor grado

conseguir mejores resultados y más precisos. Además, también se apoya en herramientas estadísticas. Con la minería de datos en redes de telecomunicaciones se abarca el modelado predictivo de la carga de tráfico en enlaces basándose en series de tiempo y la lógica difusa. Eso incluye la generación de resúmenes de colecciones de flujos de red, la gestión activa de colas en *routers* y el control de congestión de extremo a extremo (Montesino Pouzols, 2009).

Teniendo en cuenta otros ámbitos como es el enrutamiento, las redes neuronales son de gran utilidad en cuanto a la optimización del enrutamiento de una red de telecomunicaciones. Se persigue encontrar el camino más corto entre una fuente y los nodos destino (Venkataram et al., 2002), teniendo presente los entornos del tráfico como son: flujo de tráfico de entrada, la ocupación *routers*, y capacidades de enlace impidiendo la pérdida de paquetes debido a la entrada de desbordamiento de buffer (Koji´c, Reljin, & Reljin, 2006). Para esto se hace uso de una red multicapa de telecomunicaciones usando el algoritmo de *backpropagation* (Ishrat & Kumar Sharma, 2012) A través del modelo de red neural también se hace uso del algoritmos como es el de enrutamiento óptimo llamado *flow deviation algorithm* o algoritmo de desviación de flujo(Venkataram et al., 2002).

Para la recolección de la información necesaria a analizar, existe gran variedad de herramientas que permiten monitorear y obtener datos acerca del estado de la red y las estadísticas de los dispositivos de red. Algunas de estas herramientas son:

- **Cacti.**

Está escrita en PHP, se vale de RRDTools para captura y representar la información gráfica de monitoreo de estadísticas de dispositivos conectados a una red. Funciona bajo entorno Apache+MySQL. Posee una cómoda su interfaz Web. Se puede implementar en Linux, Solaris, BSD y Windows. Tiene la ventaja de uso de *plugins*, a través de los cuales es posible tener módulos para acceder a las características proporcionadas por otras herramientas de gestión de redes (Rosemberg Diaz, 2007).

Cacti se vale del protocolo SNMP para gestionar la información que necesita obtener de los dispositivos. (Rosemberg Diaz, 2007).

Cacti brinda información de:

- Trafico de Red
- Uso de la CPU
- Uso de la memoria
- Latencia

- **Solarwinds**

Software de gestión de red de licencia paga que permite la administración del desempeño vía web fundamentada en la disponibilidad, ancho de banda y fallas, permitiendo a los administradores tener acceso a las estadísticas en tiempo real. Provee acceso a información del tráfico y las estadísticas de los dispositivos de red (IT, 2015).

Solarwinds hace uso del protocolo SNMP y brinda información de monitoreo de (IT, 2015):

- Trafico de red.
- Uso de CPU
- Temperatura
- Uso de memoria
- Latencia

- **Ntop** es un *sniffer* de red. Permite visualizar el uso de la red diferenciando tanto protocolos, como puertos y también aplicaciones. Es conocido también como TCPDump. Para la captura se basa en librería de paquetes “pcap”. (Guedez Maldonado, 2005).

- **Wireshark**. Considerado un *sniffer* poderoso que permite el análisis de protocolos de redes. Usa librerías para captura basadas en librerías de paquetes “pacap”. Puede ser usado tanto en Windows como en Linux. Soporta poco más de 300 protocolos. Perite la captura de los datos directamente de desde una red o a partir de una captura ya guardada, permitiendo hacerlo hasta en 20 tipos de formatos distintos. (Guedez Maldonado, 2005).

- **Tshark**. Herramienta de análisis de tráfico.

Para el caso de software para minería de datos, a través del uso de **Weka** como herramienta se tiene la posibilidad de pre-procesamiento de los datos. Además de hacer uso de los algoritmos de la minería como son clasificación, regresiones, *clustering*, reglas de asociación y visualización. Weka usa archivos con formato de extensión .ARFF (*Attribute-Relation File Format)* (Lapeña Parreño, 2014).

Dentro de los archivos .arff, se distinguen etiquetas siempre antecedidas por el símbolo “@” tales como son (Lapeña Parreño, 2014):

- *Relation*: Nombre de la relación de datos.
- *Attribute:* Se refiere a los atributos y tipo de datos contenido en los mismos.
- *Data:* Compuesto por los datos que se van a utilizar, donde cada instancia ocupa un registro/fila en el archivo representada por los valores para cada atributos y separados por “,”.

Para la gestión de red eficiente es necesario buscar la optimización del desempeño de la red de telecomunicaciones. Se debe anticipar a escenarios críticos a través de monitoreo, diagnósticos y descripción del comportamiento de la red. Esto comprende el modelamiento del tráfico para mejorar la planificación de la capacidad de red. Se propone el uso de técnicas de inteligencia analítica como métodos

alternos para resolver problemas y tareas que pertenecen a la ingeniería de tráfico (Rivero G, 2006). Algunos autores plantean metodologías para soportar lo anterior.

El uso minería de datos se basan en su capacidad de descubrir conocimiento difícil de percibir o que está oculto en grandes cantidades de información, sin dejar por fuera que soporta muchas de sus algoritmos en las técnicas estadísticas (Martinez Luna, 2011).

La combinación de técnicas de minería descriptivas y predictivas es un gran avance para lograr el modelado de los datos y llegar a una óptima solución al problema que se desea resolver, de tal forma que se puedan hacer procesos de los datos, agrupaciones para posteriores clasificaciones (Couche, Steine, San Vicente, & Ferreira).

Entre las técnicas no supervisadas, más utilizadas son:

El clúster (Henao Ríos, 2012), el dendograma (Henao Ríos, 2012), las redes neuronales artificiales a través de los SOM (Naranjo Cuervo & Sierra Martínez, 2009), y llas reglas de asociación (Reyes Saldaña & García Flores, 2005).

Las herramientas en las que más se hizo énfasis son:

- **Monitoreo, captura y medición:** *Wireshark, Tcpdump, Cisco Netflow* (Jamuna & Vinodh Ewards, 2013).

- **Software de Minería:** *Weka* (Jamuna & Vinodh Ewards, 2013) (De la Hoz, De la Hoz, Ortiz, & Ortega, 2012) (Adeniyi Abidogun, 2005) (Couche, Steine, San Vicente, & Ferreira) (Shinde S & Abhang, 2012), (Gu, Zhang, Chen, & Du, 2011), (Erman, Arlitt, & Mahanti), (Munz, Li,, & Carle), SMO PAK (Couche, Steine, San Vicente, & Ferreira), *Botminer* (Lu, Tavallaee, & Ghorbani).

- **Herramienta Bases de datos de redes:** *Red IRIS*. (Escriche Fernández, 2011), *RDDTools* (Escriche Fernández, 2011), DARPA (De la Hoz, De la Hoz, Ortiz, & Ortega, 2012). (Shinde S & Abhang, 2012), (Gu, Zhang, Chen,

& Du, 2011), (Erman, Arlitt, & Mahanti), (Munz, Li,, & Carle), Cisco (Couche, Steine, San Vicente, & Ferreira).

- **Algoritmos Predictivos:** *algoritmo back-propagation, feedForward Red Neuronal Tipo Hopfield*. (Ishrat & Kumar Sharma, 2012) (Koji´c, Reljin, & Reljin, 2006).

- **Algoritmos Descriptivos:** K-means (Juvonen & Sipola). (Cernuzzi & Molas), Apriori, K-NN (Couche, Steine, San Vicente, & Ferreira) SOM (De la Hoz, De la Hoz, Ortiz, & Ortega, 2012).

5 METODOLOGÍA

Para el desarrollo del proyecto se usó la metodología de minería de datos CRISP-DM (*Cross Industry Standard Process for Data Mining*). Esta se basa principalmente en: comprender el negocio, comprender los datos, preparar los datos, realizar el modelado de los datos, evaluar el modelo y por ultimo hacer un despliegue de la solución.

En este trabajo se iniciaron entrevistas con el personal del CTIC (Centro de Tecnologías de Información y Comunicación) encargado del área de administración de la red e infraestructura con el fin de analizar el contexto técnico (umbrales y topologías) y no técnico (los requerimientos del negocio).

Para la selección de las herramientas de gestión de red se procedió a la instalación de algunas de licencia libre (Cacti, MTRG, Nagios, Smokeping) y luego compararlas con la herramienta Solarwinds por la cual la UBP paga licencia. Como resultado se comprobó que ésta última proporcionaba las estadísticas necesarias en los enlaces principales. Para la captura de tráfico se utilizó el *sniffer* de red Wireshark a través de la extensión Tshark.

El CTIC proporcionó un usuario al proyecto para acceder al monitoreo y captura de los datos. Una vez obtenidos y procesados los datos a partir de las capturas de tráfico y desempeño de los equipos de la red LAN del bloque 22 de la UPB, se realizó una descripción para entender el contexto. Seguido se solicitó el esquema de direcciones MAC de dispositivos finales conectados a cada *switch*, tomando las direcciones dinámicas y las estáticas.

El siguiente paso fue explorar los datos y verificar la calidad de los mismos para luego seleccionar los necesarios y aplicar las técnicas de minería de acuerdo a los siguientes experimentos:

Primero: Definir estados a partir de todas las estadísticas y transacciones de los *switch* (tomando todos los datos en conjunto).
Segundo: Análisis de Factores.

- Correlación entre las variables y la variable estado
- Componentes principales (PCA)
- Árbol de decisiones
- Regresión logística

Una vez obtenidos los resultados de las técnicas anteriores, se escogieron las variables que más influyeron para que los *switchs* incurrieran en los estados definidos.

Por último, se realizó la evaluación y despliegue del modelo de datos obtenido para determinar si realmente los resultados son de utilidad a los requerimientos de la problemática de la red LAN al interior de la UPB y que permita integrarlos en las tareas de toma de decisiones.

6 PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

6.1 DESCRIPCIÓN DE LOS REQUERIMIENTOS Y ENTENDIMIENTO DEL NEGOCIO.

El campus universitario de la Universidad Pontificia Bolivariana está conformado por bloques. En ellos se agrupan diversos tipos de usuarios, los cuales tienen acceso a internet y a los diferentes servicios de Tecnologías de Información (TI) que la Universidad presta a través del CTIC (Centro de Tecnologías de Información y Comunicación).

Adicional a los bloques que se encuentran al interior del Campus Laureles, la Universidad posee algunas sedes fuera de éste y a las cuales debe ofrecer los mismos servicios que se ofrecen al interior del Campus, como es el caso del bloque 22.

En el bloque 22 se encuentran ubicados docentes de diferentes escuelas de la Universidad y con una alta dedicación a labores de investigación. Esto implica someter la red a procesos como altas tasas de descarga de contenido, servicios de videoconferencia, y generación de *streaming* que son adicionales a los servicios de descarga de correo, consultas al sistema de información y bases de datos bibliográficas. En la actualidad durante calendario académico se tiene un consumo de ancho de banda de 1Gbps y se cuentan con aproximadamente 50 usuarios que acceden desde el bloque.

6.1.1 Servicios y aplicaciones ofertadas por CTIC.

El CTIC proporciona el acceso a los servicios de TIC garantizando la continuidad y disponibilidad de la operación de la infraestructura de TI. Los servicios y aplicaciones ofertados tienen como origen y ubicación al interior del Campus Laureles de la UPB y se pueden mencionar los siguientes:

- Conectividad: Red cableada e inalámbrica, telefonía IP (Servicio de telefonía IP convencional sobre el mismo medio. No hay QoS configurado).

- Aplicaciones: Correo electrónico, Sistema de Información (SIGAA), Internet, Moodle.
- Tipos y perfiles de usuarios: Usuarios Administrativos, Docentes e Investigadores y/o estudiantes de doctorado.

- Disponibilidad: Se pacta una disponibilidad de los servicios de 7x24.

- Ancho de banda: Conexiones de red cableada de 100 Mbps en unos puntos y 1 Gbps en otros; la conexión hacia el sitio es por fibra óptica con un ancho de banda de 1 Gbps.
  Hacia el interior del 22A es a 400Mbps.
  Hacia el instituto de termodinámica es a 1Gbps.

- Directorio Activo: Implica autenticación de cuentas de usuario y de impresión desde la UPB.

- Acceso remoto: Usuarios externos hacen uso de servicios igual de planos que otros usuarios que están allí.

6.1.2 Dispositivos de red Intermedios.

Concerniente a los tipos y cantidad de dispositivos que se encuentran implementados, no se cuenta con enrutadores, sino con diez (10) *switchs* en total entre la casa 22, 22a y 22b. Estos son equipos capa 2, marca Cisco modelo *Catalyst* 2960 con puertos de 100 Mbps y 1 Gbps. Así mismo hay un grupo de investigación (bioingeniería) con un servidor propio el cual no sale hacia la LAN de la UPB, pero genera tráfico interno y que además esa porción del tráfico surge hacia la subred configurada.

6.1.3 Problemas con respecto a la infraestructura de red.

De acuerdo a las entrevistas realizadas con el CTIC, los problemas que se presentan con respecto a la infraestructura de red en el bloque 22 son los asociados a la congestión de tráfico. Está en los equipos capa dos, debido al alto flujo de tráfico en ciertos momentos ocasionado por las actividades de los diferentes grupos de investigación que trabajan en el sitio, teniendo en cuenta que no hay segmentación de ancho de banda ni prioridad. Por ello por parte de los grupos de investigación que en el bloque residen se presenta lo siguiente:

- Eventos de inyección de tráfico no cotidianos que hacen que equipos se saturen y haya congestión en la red.
- Habilitación de manera inconsciente de servicios o procesos no debidos o que coadyuvan al deterioro o poca optimización del servicio.
- Habilitación de tráfico Ipv6.
- Eventos de conexiones y desconexiones de medios.
- Traslados de dispositivos, invertir cables sin tener conocimiento de tipos de conexiones y topologías, así como la existencia de VLAN´s.
- Grupos de investigación que utilizan servidores (al interior del bloque) que generan tráfico interno a través de procesos y aplicaciones sin muchas veces dimensionar cual es el equipo (o configuración) que realmente se necesitan ocasionando picos de tráfico en la red.

- Dificultad al momento de identificar si las causas de problemas de congestión radican en la necesidad de verificar o hacer cambio definitivo de un dispositivo o nodo de red, debido a la necesidad de optimización de recursos.

6.1.4 Herramientas, requisitos y restricciones.

6.1.4.1. Herramientas/Requisitos.

El CTIC proporcionó un computador de escritorio que reposó en las instalaciones del *rack* del bloque 22. Se utilizó con el propósito de capturar el tráfico cursado desde y hacia el bloque 22 y el campus universitario de la UPB. La conexión se realizó entre el computador y el *switch* principal, al cual se le configuró un puerto espejo (*mirroring)* por el que se hizo fluir todo el tráfico de la red del bloque. Una vez obtenidas las capturas, se clasificó el tráfico correspondiente por cada *switch*. El computador cumplía con las siguientes características: procesador Intel core i3, 4 GB de RAM, disco duro de 500GB, dos tarjetas de red, sistema operativo Linux Ubuntu.

Un computador portátil propio que cumplía con las siguientes características: procesador Intel core i5 de sexta generación, 6 GB de RAM, disco duro de 1TB, una tarjeta de red, sistema operativo Linux Ubuntu. Además, un disco duro externo propio de 1TB para almacenamiento extra.

Para el monitoreo de nodos, se obtuvo acceso al software utilizado por la Universidad llamado Solarwinds en los enlaces principales a través del protocolo simple de administración de red (SNMP).

Para capturar información acerca del tipo de tráfico que cursó en la red, se contó con el *sniffer* Wireshark/Tshark.

En cuanto al análisis descriptivo estadístico de los datos se contó con *Dq Analyzer* y *Fril*. Por último, para el proceso de minería el software libre *Weka*.

6.1.4.2. Restricciones/Limitaciones

Para el acceso a la información, se firmó un contrato de confidencialidad.

Para la prueba inicial, el CTIC proporcionó acceso al monitoreo del nodo principal cuya medición era cada 12 horas. Debido a la necesidad de realizarlo cada hora e incluir al resto de nodos, fue necesario realizar nuevamente las capturas.

No fue posible realizar las capturas en todos los *switchs* al mismo tiempo. Basándose en eso, se tomó la decisión de realizar las capturas en el nodo principal para contrarrestar esa limitación.

6.2 PREPARACIÓN DE LOS DATOS O DATASET A PARTIR DE LAS ESTADÍSTICAS DE LOS DISPOSITIVOS DE RED Y LAS VARIABLES DEL TRÁFICO CURSADO.

Con el propósito de efectuar una evaluación inicial del modelo de minería, descartar variables y datos que fuesen irrelevantes, y además ahorrar espacio de almacenamiento en datos no aportantes al proceso, sólo se tomó información acerca del protocolo de red, por lo cual se capturaron los primeros 120 bytes de cada paquete, es decir el encabezado, descartando información como el *payload*.

Una vez finalizado el monitoreo y las capturas, y obtenidas las variables, se hizo la petición al CTIC de los mapas de direcciones IP y MAC del bloque 22 y también el acceso a monitoreo del resto de s*witchs*. Luego se realizó la evaluación del nuevo modelo incorporando las variables relevantes e identificando el tráfico correspondiente a cada *switch. Esto* llevó a realizar monitoreo y capturas nuevamente durante un mes adicional.

Se obtuvo el archivo de direcciones MAC por cada *switch*, pero no una topología que permitiera tener certeza de las conexiones entre *switchs* y una jerarquía de distribución de los mismos.

6.2.1 Entendimiento de los datos.

Para entender los datos se realizó la descripción del proceso de las capturas y del monitoreo.

6.2.1.1. Proceso en Wireshark/Tshark.

Se realizó en Linux Ubuntu a través de un *script* y el administrador de ejecución de tareas *crontab* con el cual se iniciaba cada hora la captura de paquetes. Para cada archivo generado, se utilizó como nombre “fecha_hora” de la captura.

La captura se llevó a cabo por niveles en una trama Ethernet, dedicandose al análisis de estadisticas de Capa 4 y ARP e ICMP. Por ello, las estadisticas de jerarquia o niveles de protocolos TCP y UDP se tomaron de manera conceptual y a partir de esto se clasificó como variables a la cantidad de trafico de los protocolos de nivel de capas superiores o subprotocolos. En la **Figura 1** se puede observar el proceso para generar estadisticas de protocolos por jerarquía aplicado a cada archivo resultante de la captura. El resultado de los valores despues fué comparado con el mismo archivo de captura en Wireshark en las opciones de jerarquía de protocolos con el fín de cersiorarse de que fuesen iguales, como se puede observar en la **Figura 2**.

```
G:\Capturas3>tshark -r captura20151127_10.01.pcap -nqzio,stat,3600," tcp and ip
and eth and frame, data and tcp and ip and eth and frame, ssl and tcp and ip and
 eth and frame, http and tcp and ip and eth and frame, nbss and tcp and ip and e
th and frame, rtsp and tcp and ip and eth and frame, dcerpc and tcp and ip and e
th and frame, ldap and tcp and ip and eth and frame, kerberos and tcp and ip and
 eth and frame, stun and tcp and ip and eth and frame,ospf,icmp,arp, udp and ip
and eth and frame, data and udp and ip and eth and frame, snmp and udp and ip an
d eth and frame, dns and udp and ip and eth and frame, bootp and udp and ip and
eth and frame, db-lsp-disc and udp and ip and eth and frame, sip and udp and ip
and eth and frame, ntp and udp and ip and eth and frame, http and udp and ip and
 eth and frame, quic and udp and ip and eth and frame, ajns and udp and ip and e
th and frame, classicstun and udp and ip and eth and frame, mndp and udp and ip
and eth and frame, nbdgm and udp and ip and eth and frame, cldap and udp and ip
and eth and frame ">>captura20151127_10.01.csv
```

Figura 1 Comando para generar estadística de protocolos por jerarquía

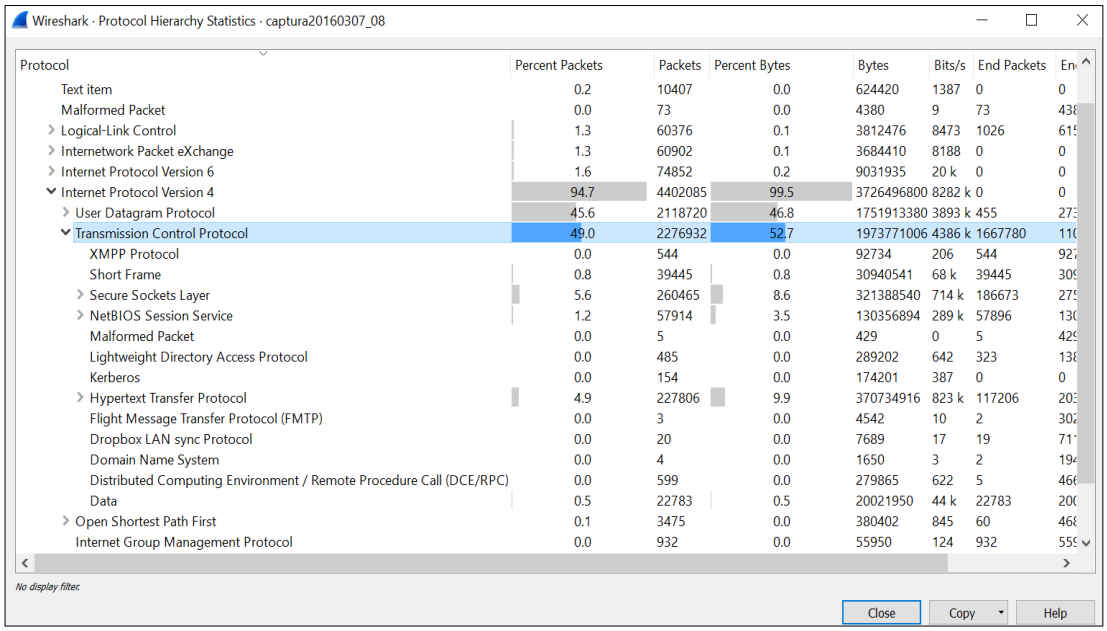


Figura 2 Estadísticas de jerarquía de protocolos-Wireshark

La **Figura 2**, hace alución a uno de los archivos de captura abierto en W*ireshark* en opción de jerarquía de protocolos. Se puede observar IPv4 (*internet protoclo v4*), seguido en la jerarquía UDP y TCP, en donde los valores totales de bytes de TCP y UDP se distribuyen entre los protocolos contenidos en la jerarquía de niveles más altos.

Por cada hora se generó un archivo con extensión *".pcap"* para un total de 309 archivos. El proceso con *tshark* fue el siguiente:

Primero, para clasificar el tráfico por *switch,* se hizo filtro por direcciones MAC a cada archivo de captura.

Segundo, a cada archivo resultante se le aplicó análisis de estadísticas de jerarquía de protocolos, ver **Figura 1**. Así se obtuvieron las estadísticas necesarias por cada protocolo, con un total de 3.090 archivos.

En la **Tabla 1**, se puede observar el tipo de trafico obtenido por protocolo. Se obtuvieron dos tipos de valores: *frames y bytes* como se observa en la **Figura 3**. *Luego s*e eliminó la columna de *frames*, pues se trabajó en cantidad de *bytes* por cada protocolo.

Tipo de tráfico generado por protocolo		
Tcp, data, ssl, http, nbss, rtsp, dcerpc, ldap, Kerberos, stun	**Udp**, data_udp, snmp, dns, bootp, db-lsp-disc2, sip, ntp, http_udp, quic, ajns, classicstun, mndp, nbdgm, cldap.	**Ospf, icmp, arp**

Tabla 1 Tipo de tráfico generado por protocolo a través de Tshark

		TCP	TCP	DATA	DATA	SSL	SSL	HTTP	HTTP	NBSS	NBSS	RTSP	RTSP	DCERPC	D
6	TIME	Frames	Bytes	Frames	Bytes	Frames	Bytes	Frames	Bytes	Frames	Bytes	Frames	Bytes	Frames	B
7	captura20160301_16	9308861	7275931258	719311	1063446468	1152543	1383580895	559129	836818134	321	526682	138	24101	11378	
8	captura20160301_17	505800	382594140	30596	42218197	72131	89258721	19814	30123646	18	17854	12	2106	1334	
9	captura20160301_18	4143089	3717334941	128368	209765276	699465	1000804429	120525	201425133	91	152543	140	24424	940	
10	captura20160301_19	1793707	1522188240	278281	469106762	78341	78793594	280073	470164851	104	192110	136	23824	901	
11	captura20160301_20	881332	691840555	12539	8481935	58170	52481781	15160	9188069	85	159839	148	26447	918	
12	captura20160301_21	753802	595413862	13671	9846113	35302	19878024	15142	9246606	55	92944	140	24474	789	
13	captura20160301_22	1232216	977964896	39204	48556006	114391	116513218	46900	53216156	95	172620	146	26146	941	
14	captura20160301_23	1402582	1268827975	132771	195499206	60852	58233968	144757	201723526	74	119704	140	24430	830	
15	captura20160302_00	1146502	953414873	23862	23721411	117925	137842182	14708	8852695	102	192737	138	24213	1167	
16	captura20160302_01	998182	856323701	11485	6464826	104195	123092266	12551	6030366	80	120136	148	26407	824	
17	captura20160302_02	827565	727441636	12077	10869857	55142	58246839	14207	8617981	83	159680	140	24514	1003	
18	captura20160302_03	680784	552421308	9609	5480240	25341	11846604	13426	7057897	121	212039	144	25755	956	
19	captura20160302_04	766853	630971935	19273	21172426	33993	22879553	24283	23582094	78	126994	140	24464	851	
20	captura20160302_05	912984	708855116	33324	43000117	30877	14438887	38317	45390149	97	181270	146	26113	968	
21	captura20160302_06	1879950	1681886732	70878	94054312	148043	201888856	72009	88708871	960	2309060	138	23943	1570	
22	captura20160302_07	7329075	7108658134	529606	956569915	1054063	1583749384	541213	939620191	37424	79700699	140	24414	9681	
23	captura20160302_08	16624152	16137602623	840507	1366048911	3280609	4808368067	741153	1218458013	35850	71126661	140	24883	12501	

Hoja1 Hoja2

Figura 3 Estadísticas de jerarquía de protocolos generadas en *Tshark* por hora.

6.2.1.2. Proceso en Solarwinds

En la **Tabla 2**, se pueden observar la selección inicial de las variables generadas por el software de monitoreo Solarwinds.

Variables iniciales generadas por el monitoreo.	
Por dispositivos:	**Por interfaces:**
Latency	bps_In
CPU Load	bps_Out
Memory	Troughput
ResponseTime	Utilization
Packet Lost	Utilization_Rec
	Utilization_Transm

Tabla 2 Variables generadas por el software de monitoreo-Solarwinds.

6.2.1.3. Variables a utilizar en el proceso de minería.

De acuerdo al monitoreo y al tipo de tráfico obtenido por protocolo, se obtuvieron en total 40 variables para cada uno de los *switchs*, observar **Tabla 3.**

No	Variable	Descripción	Medida	Tipo
1	Fecha	Fecha y hora de la captura de datos por hora.	Tiempo	Fecha
2	TCP	Cantidad de bytes transmitidos por hora con el Protocolo de Control de Transmisión.	Bytes/h	Numérico
3	DATA	Cantidad de bytes transmitidos por hora con el Identificador del cuerpo del mensaje para TCP.	Bytes/h	Numérico
4	SSL	Cantidad de bytes transmitidos por hora con el protocolo de Seguridad de la Capa de Transporte.	Bytes/h	Numérico
5	HTTP	Cantidad de bytes transmitidos por hora con el protocolo de Protocolo de Transferencia de Hipertexto.	Bytes/h	Numérico
6	NBSS	Cantidad de bytes transmitidos por hora con el protocolo de Servicio de sesión NetBIOS.	Bytes/h	Numérico
7	RTSP	Cantidad de bytes transmitidos por hora con el Protocolo de Transmisión en Tiempo Real	Bytes/h	Numérico
8	DCERPC	Cantidad de bytes transmitidos por hora con el protocolo de DCE Llamada a procedimiento remoto. Útil en directorio activo.	Bytes/h	Numérico
9	LDAP	Cantidad de bytes transmitidos por hora con el Protocolo Ligero de Acceso de Directorios.	Bytes/h	Numérico
10	KERBEROS	Cantidad de bytes transmitidos por hora con el Protocolo de autenticación de redes de ordenador	Bytes/h	Numérico
11	STUN	Cantidad de bytes transmitidos por hora con el protocolo de Session Traversal Utilities for NAT	Bytes/h	Numérico
12	OSPF	Cantidad de bytes transmitidos por hora con el Protocolo de Enrutamiento Open source	Bytes/h	Numérico
13	ICMP	Cantidad de bytes transmitidos por hora con el Protocolo de Mensajes de Control de Internet.	Bytes/h	Numérico

14	ARP	Cantidad de bytes transmitidos por hora con el Protocolo de Resolución de Direcciones	Bytes/h	Numérico
15	UDP	Cantidad de bytes transmitidos por hora con el Protocolo de Datagrama de Usuario	Bytes/h	Numérico
16	DATA_UDP	Cantidad de bytes transmitidos por hora con el protocolo que Identifica el cuerpo del mensaje para UDP	Bytes/h	Numérico
17	SNMP	Cantidad de bytes transmitidos por hora con el Protocolo Simple de Administración de Red	Bytes/h	Numérico
18	DNS	Cantidad de bytes transmitidos por hora con el protocolo de Sistema de Nombres de Dominio	Bytes/h	Numérico
19	BOOTP	Cantidad de bytes transmitidos por hora con el Protocolo de secuencia de arranque. Utilizado por clientes UDP para obtener su dirección IP.	Bytes/h	Numérico
20	DB-LSP-DISC2	Cantidad de bytes transmitidos por hora con el protocolo Dropbox LAN Sync Discovey Protocol	Bytes/h	Numérico
21	SIP	Cantidad de bytes transmitidos por hora con el Protocolo de Inicio de Sesiones.	Bytes/h	Numérico
22	NTP	Cantidad de bytes transmitidos por hora con el Protocolo de tiempo de red.	Bytes/h	Numérico
23	HTTP_UDP	Cantidad de bytes transmitidos por hora con el Protocolo de Transferencia de Hipertexto.	Bytes/h	Numérico
24	QUIC	Cantidad de bytes transmitidos por hora con el protocolo Quick UPD Internet Connections, protocolos para optimizar conexiones de baja velocidad y alta latencia.	Bytes/h	Numérico
25	AJNS	Cantidad de bytes transmitidos por hora con el protocolo AllJoyn Name Service Protocol	Bytes/h	Numérico
26	CLASSICSTUN	Cantidad de bytes transmitidos por hora con el protocolo Simple Traversal of UDP Through NAT. Útil para Telefonía VoIP, SIP.	Bytes/h	Numérico
27	MNDP	Cantidad de bytes transmitidos por hora con el protocolo The MikroTik Neighbor Discovery Protocol	Bytes/h	Numérico
28	NBDGM	Cantidad de bytes transmitidos por hora con el protocolo Servicios de Datagrama de NetBios	Bytes/h	Numérico
29	CLDAP	Cantidad de bytes transmitidos por hora con el Protocolo Ligero de Acceso a Directorios no Orientado a Conexión.	Bytes/h	Numérico
30	Network_Latency	Sumatoria por hora de los retardos temporales dentro de una red de datos.	Milisegundos	Numérico
31	CPU_Load	Porcentaje por hora de Carga de la Unidad central de procesamiento	Porcentual	Numérico
32	Memory	Porcentaje por hora de Memoria de acceso aleatorio	Porcentual	Numérico
33	ResponseTime	Sumatoria por hora de los retardos temporales dentro de una red de datos.	Milisegundos	Numérico
34	Packet_Loss	Porcentaje por hora de Paquetes perdidos	Porcentual	Numérico
35	bps_In	Tasa de transferencia cada hora de entrada de datos en la red.	Bits por segundo	Numérico
36	bps_out	Tasa de transferencia cada hora de salida de datos en la red	Bits por segundo	Numérico
37	Utilization_Rec	Porcentaje de utilización por hora de la interfaz, al recibir	Porcentual	Numérico
38	Utilization_Transm	Porcentaje de utilización por hora de la interfaz, al enviar	Porcentual	Numérico
39	Utilization	Porcentaje total de utilización por hora dela interfaz	Porcentual	Numérico
40	Troughput bps	Total de entrada/salida cada hora de transferencia en la red. Bps_in+Bps_out	Bits por segundo	Numérico

Tabla 3 Variables a utilizar para el proceso de minería
Tomado de https://wireshark.org/docs/dfref

6.2.2 Preparación de los Datos.

6.2.2.1. Integración de los datos.

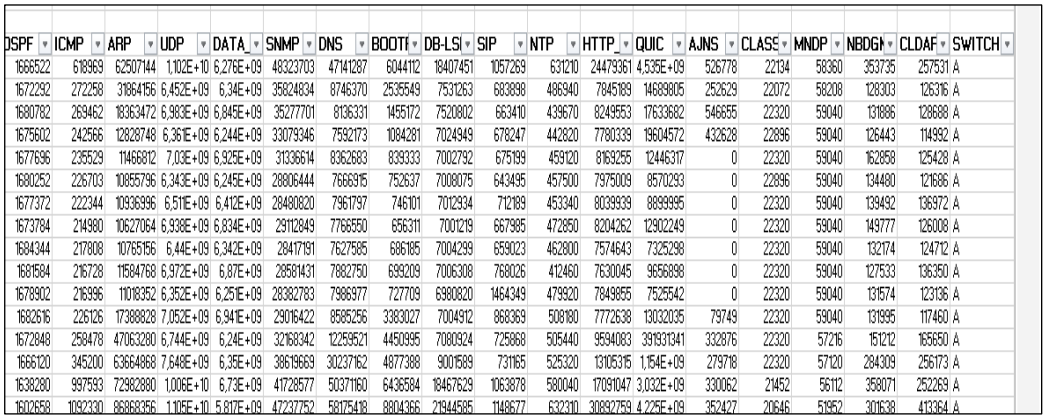
Se procedió a realizar la integración de los datos de las capturas de tráfico y de monitoreo por cada *switch*.

El monitoreo de estadísticas de los dispositivos al igual que el de capturas de tipo de tráfico se realizó por horas, pero en la última, se omitieron domingos y algunas horas del día por motivos ajenos al proceso: como fluido eléctrico y otros factores. De acuerdo a lo anterior, se desecharon las filas de registros que no coincidían con ambos archivos de tal forma que se ajustaran exactamente las capturas y el monitoreo.

De los 10 *switch (de la **A** a la **J**)*, sólo usaron nueve (09), pues del *switch* H no se obtuvo información acerca de las interfaces monitoreadas.

Una vez integrados los datos de las estadísticas y el tipo de tráfico, se procedió a conformar un solo archivo con todos los datos y agregar una columna para

identificar a que dispositivo pertenecía cada transacción, observar la **Figura 4**. El total de registros fue de 2.782 incluyendo el encabezado.



OSPF	ICMP	ARP	UDP	DATA_	SNMP	DNS	BOOTI	DB-LS(	SIP	NTP	HTTP_	QUIC	AJNS	CLASS	MNDP	NBDG	CLDAF	SWITCH
1666522	618969	62507144	1,102E+10	6,276E+09	48323703	47141287	6044112	18407451	1057269	631210	24479361	4,535E+09	526778	22134	58360	353735	257531	A
1672292	272258	31864156	6,452E+09	6,34E+09	35824834	8746370	2535549	7531263	683898	486940	7845189	14689805	252629	22072	58208	128303	126316	A
1680782	269462	18363472	6,983E+09	6,845E+09	35277701	8136331	1455172	7520802	663410	439670	8249553	17633682	546655	22320	59040	131886	128688	A
1675602	242566	12828748	6,361E+09	6,244E+09	33079346	7592173	1084281	7024949	678247	442820	7780339	19604572	432628	22896	59040	126443	114992	A
1677696	235529	11466812	7,03E+09	6,925E+09	31336614	8362683	839333	7002792	675199	459120	8169255	12446317	0	22320	59040	162858	125428	A
1680252	226703	10855796	6,343E+09	6,245E+09	28806444	7666915	752637	7008075	643495	457500	7975009	8570293	0	22896	59040	134480	121686	A
1677372	222344	10936996	6,511E+09	6,412E+09	28480820	7961797	746101	7012934	712189	453340	8039939	8899995	0	22320	59040	139492	136972	A
1673784	214980	10627064	6,938E+09	6,834E+09	29112849	7766550	656311	7001219	667985	472850	8204262	12902249	0	22320	59040	149777	126008	A
1684344	217808	10765156	6,44E+09	6,342E+09	28417191	7627585	686185	7004299	659023	462800	7574643	7325298	0	22320	59040	132174	124712	A
1681584	216728	11584768	6,972E+09	6,87E+09	28581431	7882750	699209	7006308	768026	412460	7630045	9656898	0	22320	59040	127533	136350	A
1678902	216396	11018352	6,352E+09	6,251E+09	28382783	7986977	727709	6980820	1464343	479920	7849855	7525542	0	22320	59040	131574	123136	A
1682616	226126	17388828	7,052E+09	6,941E+09	29016422	8585256	3383027	7004912	868369	508180	7772638	13032035	79749	22320	59040	131995	117460	A
1672848	258478	47063280	6,744E+09	6,24E+09	32168342	12259521	4450995	7080924	725868	505440	9594083	391931341	332876	22320	57216	151212	165650	A
1666120	345200	63664868	7,648E+09	6,35E+09	38619669	30237162	4877388	9001589	731165	525320	13105315	1,154E+09	279718	22320	57120	284309	256173	A
1638280	997593	72982880	1,006E+10	6,73E+09	41728577	50371160	6436584	18467629	1063878	580040	17091047	3,032E+09	330062	21452	56112	358071	252269	A
1602658	1092330	86868356	1,105E+10	5,817E+09	47237752	58175418	8804366	21944585	1148677	632310	30892759	4,225E+09	352427	20646	51952	301638	413364	A

Figura 4 Integración de los datos

6.2.2.2. Descripción Estadística de los datos

Se hizo uso de *DQ Analyzer* para realizar el análisis estadístico de las variables seleccionadas en la integración, obteniendo mininos, máximos, media y desviación estándar. En la **Tabla 4,** se puede observar el resultado del análisis estadístico de cada variable.

Variable	Type	Nulls	Distinct	Min	Max	Median	Average	Standard deviation
Fecha	STRING	0	309	1/03/2016 16:00	9/03/2016 9:00	NA	NA	NA
Latency	INTEGER	0	31	0	109	4	3,8	4,2
CPU_Load	INTEGER	0	17	2	37	6	8,37	9,11
Memory	FLOAT	0	2.737	20,21	54,14	35,06	39,44	10,48
Response_Time	INTEGER	0	31	0	109	4	3,8	4,2
Packet_Lose	INTEGER	0	1	0	0	0	0	0
bps_In	FLOAT	0	2.781	13.091,94	138.349.360	207.319,48	5.691.117,29	15.619.847,86
bps_out	FLOAT	0	2.781	830,09	99.088.360	437.412,94	1.452.844,99	7.815.952,87
Throughput_bps	FLOAT	0	2.781	15.520,62	232.608.200	766.074,05	7.143.962,28	21.252.928,51
Utilization_Rec	FLOAT	0	2.781	0	20,15	0,05	0,61	1,46
Utilization_Transm	FLOAT	0	2.781	0	9,83	0,05	0,29	0,73
TCP	LONG	0	2.679	153.058	41.943.343.359	583.141.763	2.206.484.068,20	4.023.433.304,82
DATA	LONG	0	2.677	11.165	6.018.493.507	17.597.805	246.834.366,24	510.606.218,57
SSL	LONG	0	2.245	0	8.370.066.622	32.883.231	461.224.970,10	843.507.725,65
HTTP	LONG	0	2.642	0	5.892.885.406	17.753.855	231.811.627,77	495.805.352,97
NBSS	LONG	0	2.108	1.547	31.312.684.440	152.032	147.403.541,18	1.836.570.580,62
RTSP	INTEGER	0	267	0	80.081	0	2.723,44	7.846,47
DCERPC	INTEGER	0	2.089	0	37.374.981	184.634	1.694.426,33	3.677.839,14
LDAP	INTEGER	0	2.123	0	3.266.493	334.447	521.191,84	457.341,80
KERBEROS	INTEGER	0	1.983	0	3.480.403	135.778	289.102,18	349.650,33
STUN	INTEGER	0	15	0	8.616	140	128,54	372,79
OSPF	INTEGER	0	627	152.216	1.701.858	623.454	721.400,62	333.185,91
ICMP	INTEGER	0	2.620	13.837	2.042.923	152.962	254.780,05	253.178,78
ARP	INTEGER	0	2.678	500.236	128.232.044	4.367.352	11.657.197,40	18.854.158,96
UDP	LONG	0	2.683	3.079.441	15.219.399.918	109.947.875	1.586.199.308,52	2.723.286.941,63
DATA_UDP	LONG	0	2.481	375.447	7.656.974.000	5.901.258	723.838.951,58	2.013.842.220,67
SNMP	INTEGER	0	2.683	1.104.451	62.715.262	14.580.984	17.413.987,49	11.368.926,91
DNS	INTEGER	0	2.546	106.336	79.814.627	4.329.890	10.080.076,14	12.083.168,35
BOOTP	INTEGER	0	2.555	3.425	28.614.786	332.933	1.138.889,23	2.242.338,01
DB_LSP_DISC	INTEGER	0	1.611	0	30.538.414	4.651.385	6.212.910,79	5.252.013,69
SIP	INTEGER	0	2.669	42.939	1.652.914	357.888	417.166,60	220.692,39
NTP	INTEGER	0	2.457	26.260	781.550	131.580	176.274,63	143.951,15
HTTP_UDP	INTEGER	0	2.168	6.300	55.912.531	4.885.046	7.337.688,43	6.811.214,37
QUIC	LONG	0	1.859	0	8.527.531.160	13.371.522	774.122.402,16	1.297.983.739,12
AJNS	INTEGER	0	526	0	709.034	0	53.280,16	108.113,05
CLASSICSTUN	INTEGER	0	85	0	44.838	0	6.405,28	9.961,23
MNDP	INTEGER	0	113	0	73.488	0	12.484,43	23.775,90
NBDGM	INTEGER	0	1.006	0	713.305	51.616	73.989,65	86.488,69
CLDAP	INTEGER	0	2.073	0	494.614	55.482	93.766,75	83.530,33

Tabla 4 Descripción estadística de las variables

6.2.2.3. Limpieza de los datos

Simultáneamente a la descripción estadística, se obtuvo información acerca de los atributos:

- Valores nulos: como se puede observar la **Tabla 4,** no se presentaron valores nulos.

- Completitud: como se puede observar la **Tabla 4,** los datos están completos y cumplen con un Dominio, con lo que se puede determinar que no hubo variables incompletas.
- Valores Atípicos: como se puede observar la **Tabla 4,** de acuerdo al resultado de la descripción estadística, los valores arrojados por el monitoreo y las capturas de datos son normales dentro del área de redes de telecomunicaciones.
- Duplicidad: con la herramienta FRIL se puede resolver este tipo problemas y de completitud. De acuerdo a nuestro tipo de datos y a lo arrojado por DQ Analyzer no fue necesario, puesto que son variables numéricas y es normal contar con registros duplicados. **Ver Tabla 4.**

6.2.2.4. Selección de Variables para el proceso de minería.

- Variables irrelevantes: Sólo se contó con una variable irrelevante a simple vista: *Packet_loss*, puesto que los valores de este atributo para todos los *switchs* fue de cero.

- Análisis de correlaciones: Es necesario identificar variables con contenidos redundantes o correlacionados. Por lo tanto, en los datos tener TCP y UDP junto con protocolos de nivel más alto sería tener la misma información puesto que los superiores estan encapsulados en estos. Para el análisis se usó la fórmula de Karl Pearson, con la cual se obtiene un rango de valores [-1 a 1]. Ver **Ecuación 1.**

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

Ecuación 1 Fórmula de Karl Pearson para el análisis de correlaciones

Si los valores altos de una variable tienden a estar asociados con los valores altos de otra, sería una correlación positiva; si los valores bajos de una variable tienden a estar asociados con los valores bajos de otra, sería una correlación negativa; y si los valores de ambas tienden a no estar relacionados tienden a cero.

Dado que la información está empaquetada en diferentes capas del modelo TCP/IP, para éste análisis se utilizó un umbral máximo moderado de 0,8 en las correlaciones, teniendo en cuenta que el tráfico no tiene el mismo comportamiento en todos los *switchs,* por lo cual algunas variables pueden tener mucha incidencia en un nodo durante el tiempo de captura, mientras que, en otros cero, pero eso no implica que en otro lapso no sea aportante a la congestión de tráfico y se busca que el modelo pueda ser aplicado a otros dispositivos, por ende, se corrió el riesgo de dejar algunas variables irrelevantes, dado que el tráfico estaba correlacionado con otro, pero no se quería tener sesgo en la información. En la **Tabla 5** se puede observar la matriz de correlaciones obtenida, en la cual los círculos rojos rellenos en amarillo tuvieron un valor >0,8.

Matriz de Correlaciones

Columna1	Latency	CPU_Load	Memory	Response_Tim	bps_In	bps_out	Throughput	Utilization_	Utilization_	Utilization	TCP	DATA	SSL	HTTP	NBSS	RTSP	DCERPC	LDAP	KERBEROS	STUN	OSPF	ICMP	ARP	UDP	DATA_UDP	SNMP	DNS	BOOTP	DB-LSP-DISC	SIP	NTP	HTTP_UDP	QUIC	AJNS	CLASSICSTUN	MNDP	NBDGM	CLDAP
Latency	1																																				
CPU_Load	-0,025	1																																			
Memory	0,135	-0,717	1																																		
Response_	1,000	-0,025	0,135	1																																	
bps_In	0,048	0,030	-0,153	0,048	1																																
bps_out	-0,027	0,001	-0,063	-0,027	0,600	1																															
Throughput	0,025	0,023	-0,136	0,025	0,956	0,809	1																														
Utilization_	0,043	-0,020	-0,072	0,043	0,793	0,489	0,763	1																													
Utilization_	-0,164	-0,096	0,057	-0,164	0,404	0,755	0,574	0,539	1																												
Utilization	-0,029	-0,051	-0,032	-0,029	0,744	0,648	0,785	0,949	0,777	1																											
TCP	-0,039	0,054	-0,114	-0,039	0,618	0,670	0,701	0,616	0,631	0,697	1																										
DATA	-0,013	0,101	-0,142	-0,013	0,386	0,180	0,350	0,388	0,145	0,344	0,699	1																									
SSL	-0,015	0,087	-0,109	-0,015	0,366	0,142	0,321	0,394	0,148	0,350	0,719	0,673	1																								
HTTP	-0,013	0,097	-0,140	-0,013	0,369	0,153	0,327	0,369	0,117	0,319	0,663	0,938	0,649	1																							
NBSS	-0,002	-0,005	-0,032	-0,002	0,486	0,878	0,680	0,448	0,794	0,632	0,629	0,055	0,006	0,025	1																						
RTSP	0,037	-0,043	-0,145	0,037	0,368	0,189	0,340	0,260	0,028	0,205	0,328	0,277	0,311	0,273	0,074	1																					
DCERPC	-0,023	0,073	-0,015	-0,023	0,153	0,070	0,138	0,229	0,119	0,215	0,430	0,488	0,518	0,465	0,010	0,087	1																				
LDAP	-0,041	0,153	-0,113	-0,041	0,250	0,102	0,221	0,314	0,135	0,285	0,610	0,654	0,763	0,635	-0,015	0,285	0,627	1																			
KERBEROS	-0,008	0,138	-0,092	-0,008	0,227	0,100	0,203	0,281	0,123	0,256	0,539	0,566	0,683	0,547	0,003	0,224	0,631	0,915	1																		
STUN	-0,019	0,042	-0,006	-0,019	0,049	0,007	0,039	0,066	0,020	0,057	0,136	0,125	0,204	0,118	-0,016	0,077	0,083	0,220	0,170	1																	
OSPF	0,023	-0,043	-0,148	0,023	0,321	0,189	0,306	0,209	0,026	0,166	0,325	0,263	0,301	0,259	0,082	0,972	0,089	0,231	0,225	0,084	1																
ICMP	-0,046	0,090	-0,092	-0,046	0,278	0,106	0,244	0,308	0,119	0,275	0,523	0,556	0,650	0,537	-0,012	0,389	0,396	0,670	0,504	0,235	0,384	1															
ARP	-0,093	-0,008	-0,187	-0,093	0,351	0,195	0,329	0,459	0,304	0,457	0,569	0,558	0,600	0,549	0,063	0,519	0,290	0,556	0,454	0,173	0,510	0,622	1														
UDP	0,014	0,017	-0,176	0,014	0,451	0,202	0,406	0,396	0,092	0,330	0,572	0,589	0,637	0,578	0,042	0,854	0,333	0,536	0,480	0,150	0,841	0,660	0,684	1													
DATA_UDP	0,036	-0,042	-0,148	0,036	0,362	0,171	0,329	0,257	0,013	0,197	0,322	0,283	0,322	0,279	0,052	0,390	0,095	0,294	0,232	0,082	0,980	0,395	0,523	0,866	1												
SNMP	-0,083	0,272	-0,549	-0,083	0,336	0,128	0,294	0,286	0,046	0,231	0,522	0,594	0,597	0,585	-0,018	0,647	0,314	0,611	0,487	0,146	0,664	0,592	0,684	0,812	0,658	1											
DNS	-0,026	0,138	-0,121	-0,026	0,340	0,132	0,299	0,384	0,138	0,338	0,697	0,774	0,851	0,754	0,000	0,322	0,620	0,913	0,826	0,210	0,312	0,709	0,624	0,690	0,328	0,665	1										
BOOTP	-0,084	0,001	-0,181	-0,084	0,271	0,125	0,245	0,376	0,230	0,367	0,456	0,476	0,512	0,471	0,005	0,446	0,244	0,470	0,396	0,148	0,439	0,577	0,866	0,599	0,451	0,564	0,530	1									
DB-LSP-DI	-0,116	0,126	-0,194	-0,116	0,252	0,085	0,217	0,394	0,202	0,370	0,613	0,718	0,724	0,704	-0,040	0,296	0,502	0,798	0,643	0,204	0,300	0,736	0,750	0,652	0,305	0,729	0,854	0,665	1								
SIP	-0,103	0,265	-0,331	-0,103	0,309	0,166	0,288	0,310	0,106	0,271	0,530	0,559	0,612	0,547	0,013	0,661	0,374	0,666	0,563	0,184	0,687	0,603	0,688	0,801	0,673	0,780	0,691	0,611	0,711	1							
NTP	-0,047	0,034	-0,223	-0,047	0,395	0,201	0,364	0,322	0,089	0,274	0,471	0,440	0,491	0,433	0,060	0,922	0,220	0,493	0,400	0,142	0,927	0,524	0,671	0,910	0,932	0,775	0,531	0,577	0,527	0,821	1						
HTTP_UDP	-0,043	0,081	-0,087	-0,043	0,324	0,121	0,282	0,339	0,103	0,292	0,620	0,677	0,753	0,660	-0,018	0,471	0,508	0,779	0,642	0,186	0,471	0,763	0,647	0,753	0,481	0,689	0,834	0,555	0,823	0,739	0,646	1					
QUIC	-0,022	0,089	-0,119	-0,022	0,340	0,140	0,301	0,390	0,161	0,352	0,631	0,725	0,755	0,708	0,006	0,209	0,513	0,727	0,590	0,171	0,196	0,704	0,548	0,673	0,218	0,611	0,855	0,487	0,816	0,564	0,396	0,755	1				
AJNS	0,038	-0,095	-0,188	0,038	0,267	0,122	0,241	0,151	-0,026	0,103	0,254	0,296	0,272	0,295	0,023	0,376	0,092	0,166	0,125	0,044	0,367	0,278	0,358	0,424	0,373	0,455	0,258	0,287	0,247	0,262	0,380	0,310	0,270	1			
CLASSICST	-0,281	-0,116	-0,239	-0,281	0,221	0,110	0,203	0,242	0,179	0,248	0,263	0,237	0,238	0,238	0,008	0,544	0,099	0,283	0,209	0,067	0,552	0,287	0,573	0,523	0,552	0,538	0,271	0,516	0,420	0,563	0,673	0,355	0,207	0,241	1		
MNDP	-0,169	-0,086	-0,237	-0,169	0,206	0,111	0,192	0,279	0,235	0,296	0,254	0,219	0,248	0,218	0,015	0,658	0,082	0,274	0,202	0,076	0,661	0,332	0,687	0,600	0,666	0,597	0,264	0,614	0,430	0,631	0,730	0,372	0,186	0,203	0,821	1	
NBDGM	-0,122	0,142	-0,427	-0,122	0,280	0,099	0,242	0,356	0,162	0,327	0,487	0,583	0,555	0,577	-0,019	0,415	0,323	0,541	0,465	0,134	0,418	0,624	0,803	0,627	0,425	0,745	0,610	0,810	0,747	0,664	0,572	0,578	0,588	0,342	0,570	0,614	1
CLDAP	-0,039	0,122	-0,143	-0,039	0,299	0,120	0,264	0,342	0,122	0,302	0,628	0,678	0,747	0,661	-0,017	0,462	0,548	0,905	0,778	0,199	0,467	0,752	0,632	0,749	0,474	0,738	0,880	0,533	0,844	0,750	0,649	0,846	0,762	0,263	0,370	0,385	0,608	1

Tabla 5 Matriz de correlaciones

En la **Tabla 6,** se presenta la selección de las variables irrelevantes/redundantes para el análisis de la minería de datos de acuerdo al resultado de la matriz de correlaciones.

Item	***Atributo***	***Variables Correlacionadas***	***Valor***	***Justificación***
1	Packet_Loss	N/A	0	Todos los registros están en cero.
2	Response_Time	Latency	1	Es la misma información.
3	Troughput	Bps_in	0,95	Bps_in+Bps_out= Troughput. Directamente relacionadas. Es relevante identificar cantidad de tráfico entrante/saliente.
		Bps_out	0,81	
4	Utilization	Utilization_Rec	0,94	Utilization=Utilization_rec+Utilization_trans. Directamente relacionadas.
5	TCP	SSL	0,71	SSL por ser de mayor nivel está encapsulado en TCP, por lo que sería la misma información.
6	Data	HTTP	0,99	Por especificación de tipo de tráfico es más relevante Http.
7	NBSS	Bps_out	0,87	Protocolo de uso interno y presenta un porcentaje relativamente bajo en comparación y es más relevante Bps_out para el proceso.
8	KERBEROS	LDAP	0,91	Por especificación de tipo de tráfico es más relevante LDAP.
9	OSPF	RTSP	0,97	Es más relevante conocer uso de aplicaciones en tiempo real y QoS, y no información acerca de protocolo sobre el cual se enrutan los servicios.
10	UDP	RTSP	0,85	UDP por ser de menor nivel encapsula a los de mayor, por lo que sería la misma información.
		SNMP	0,81	
		SIP	0,80	
		OSPF	0,84	
11	DATA_UDP	RTSP	0,98	Por especificación de tipo de tráfico es más relevante NTP.
		LDAP	0,86	Por especificación de tipo de tráfico es más relevante LDAP.
12	DNS	SSL	0,85	Es relevante saber las aplicaciones que hacen uso de DNS y cuales afectan el proceso.
		LDAP	0,91	
		DB-LSP-DISC	0,85	
		KERBEROS	0,82	
13	ARP	BOOTP	0,86	Por especificación de tipo de tráfico es más relevante BOOTP.
14	NTP	RTSP	0,92	Por especificación de tipo de tráfico es más relevante RTSP, NTP es básicamente sincronización de la hora.
		OSPF	0,92	Ya eliminada en el Item 9.
		UDP	0,91	Ya eliminada en el Item 10.
		SIP	0,82	Por especificación de tipo de tráfico es más relevante SIP.
15	HTTP_UDP	DNS	0,83	Ya eliminada en el Item 12.
		DROP BOX	0,82	Por especificación de tipo de tráfico es más relevante DROP BOX.
16	MNDP	CLASSICSTUN	0,82	Por especificación de tipo de tráfico es más relevante CLASSICSTUN.
17	NBDGM	ARP	0,80	Ya eliminada
		BOOTP	0,80	Por especificación de tipo de tráfico es más relevante BOOTP.
18	CLDAP	LDAP	0,90	Por especificación de tipo de tráfico es más relevante LDAP y DROP BOX.
		DNS	0,88	
		DROP BOX	0,84	
		HTTP_UDP	0,84	

Tabla 6 Variables irrelevantes/redundantes para el proceso de minería

6.3 EXPERIMENTOS Y EVALUACIÓN DE RESULTADOS

A través de una prueba piloto se buscó la implementación de un modelo que fuese aplicable no sólo a los dispositivos del experimento, sino que también pueda ser aplicado a otros *switchs* en diversas áreas.

Los datos analizados se componen de 22 variables seleccionadas. En la **Tabla 7,** se pueden observar las variables relevantes para el proceso de minería de acuerdo al resultado de la matriz de correlaciones. Se resalta que para evitar sesgos en la información se corrió el riesgo de dejar algunas variables irrelevantes, debido a que ese tráfico estaba correlacionado con otro. Luego a través de los siguientes experimentos se buscó determinar sólo las variables más relevantes.

Variables	
1. Fecha	12. DCERPC
2. Latency	13. LDAP
3. CPU_load	14. STUN
4. Memory	15. ICMP
5. Bps_in	16. SNMP
6. Bps_out	17. BOOTP
7. Utilization_trans	18. DB-LSP-DISC
8. Utilization_rec	19. SIP
9. SSL	20. QUIC
10. HTTP	21. AJNS
11. RTSP	22. CLASSICSTUN

Tabla 7 Variables seleccionadas para el proceso de minería de datos.

6.3.1 Definir Estados a partir de todos los datos de tráfico recolectados de los *switchs*.

El objetivo de este experimento fue definir ciertos estados para establecer en cuáles incurrió cada *switch* durante el tiempo de monitoreo y capturas. Se realizó a partir del conjunto total de las estadísticas y transacciones (intercambio de datos en el protocolo TCP/IP) de todos los *switchs*. La variable resultante se llamó Estados.

- Se inició con el algoritmo *Simple EM (expectation maximisation) class*. Este genera una cantidad automática de grupos con características similares y así poder obtener una idea del número ideal de *clusters* a utilizar. El resultado obtenido fue de 5 estados a partir de las estadísticas y transacciones con un "Log likelihood: -243.23734" de los cuales a simple vista algunos resultados estaban solapados.

- Se utilizó el algoritmo *SimpleKmeans* para generar clústeres de 3, 4 y 5 grupos diferentes de estados y obtener los valores de los centroides y las desviaciones estándar de cada uno, lo que proporcionó información más ampliada para los rangos de valores de las variables. Se obtuvieron los siguientes resultados de evaluación:

    - Para 5 grupos de estados, Within cluster sum of squared errors: 585.43.
    - Para 4 grupos de estados, Within cluster sum of squared errors: 668.78.
    - Para 3 grupos de estados, Within cluster sum of squared errors: 721.04.

A simple vista era viable tomar los resultados con 5 grupos por tener errores más bajos, pero se decidió tomar 4 grupos, pues con 3 grupos se obtuvo información menos detallada y con 5 grupos se obtuvo el resultado de datos muy solapados.

Se definieron 4 estados, cuya distribución de los datos quedó conformada como se muestra en la **Figura 5** y gráficamente como se puede observar el histograma de la **Figura 6**.

Name: Cluster — Type: Nominal
Missing: 0 (0%) — Distinct: 4 — Unique: 0 (0%)

No.	Label	Count	Weight
1	cluster0	865	865.0
2	cluster1	333	333.0
3	cluster2	343	343.0
4	cluster3	1240	1240.0

Figura 5 Resultado de la distribución de los datos algoritmo SimpleKmeans para los 4 Estados en Weka

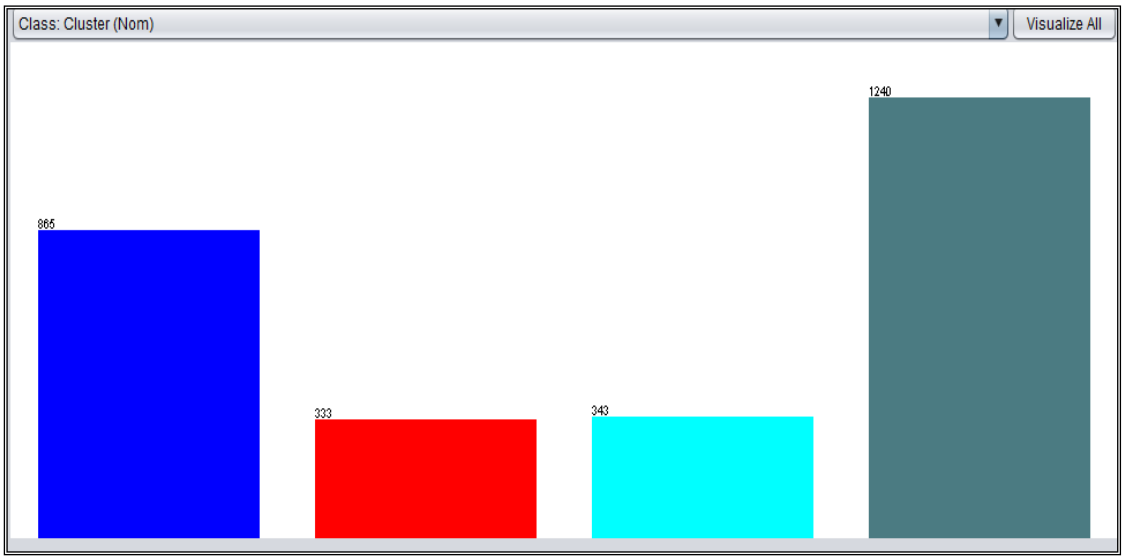


Figura 6 Histograma de los datos algoritmo SimpleKmeans para los 4 Estados en Weka

Para la realizar la descripción, los rangos de valores referenciados se tomaron de los datos obtenidos en los cuatro estados y se tuvieron en cuenta los datos de los servicios ofertados por el CTIC referente a conexiones de red cableada.

Para nombrar las categorías de la variable Estados de forma clara, evitando nombres extensos, que permitieran ser de fácil recordación, se utilizó una letra que representó el valor para el estado de cada variable en: Altos **(A)**, Medios **(M)** y Bajos **(B).** Las variables fueron: Memoria RAM, Porcentaje de uso de CPU, Latencia y Tráfico externo. Observar **Tabla 8.**

Variables	Valores Bajos (B)	Valores Medios (M)	Valores Altos (A)
Memoria RAM	Entre 0% y 29%	Entre 30 y 50%	De 51% a 100%
Uso de CPU	Entre 0% y 19%	Entre 20 y 50%	De 51% a 100%
Latencia	Entre 0ms y 1ms	Entre 3ms y 9ms	De 10ms en adelante.
Tráfico Externo	Entre 0Mbps y 300Mbps	Entre 301Mbps y 700 Mbps	De 701Mbps a 1Gbps.

Tabla 8 Referencias para rango de valores de las variables

En la **Tabla 9**, se pueden observar los estados definidos a partir del conjunto total de las todas las estadísticas y transacciones de los *switchs*. De acuerdo al resultado**,** es posible apreciar que los valores para el tráfico externo no llegaron a ser altos durante el tiempo de monitoreo y capturas, por lo que se supondría que el consumo de recursos fue en mayor parte por tráfico interno. Esto teniendo en cuenta que los puertos monitoreados tenían ancho de banda de 1Gbps.

Variable Estados	***Descripción***
M/M/B/B	En este estado se agruparon las estadísticas y transacciones con valores para: • Memoria RAM: Medios (entre 24% y 35%). • Consumo de CPU: Medios (entre 20% y 30%). • Latencia: Bajos. • Interfaz Monitoreada: Bajos - recepción/transmisión. • Tráfico Externo: Bajos - entre 15Mbps y 20Mbps para recepción - entre 7,5Mbps y 14,5Mbps para transmisión. Distribución de valores para consumo tráfico externo: Moderados: LDAP, HTTP, SSL, VoIP y DB-LSP-DISC2. Medio-altos: SNMP y QUIC.
A/B/A/B	En este estado se agruparon las estadísticas y transacciones con valores para: • Memoria RAM: Altos (entre 51,6% y 58%). • Consumo de CPU: Bajos (entre 3,5% y 6%). • Latencia: Altos. • Interfaz Monitoreada: Bajos - recepción/transmisión. • Tráfico Externo: Bajos - entre 6,3Mbps y 10,3Mbps para recepción - entre

	0Mbps y 2,8Mbps para transmisión. Distribución de valores para consumo tráfico externo: Altos: LDAP, HTTP, SSL y QUIC. Medios: DB-LSP-DISC2 y SNMP. Bajos: VoIP.
M/B/M/B	En este estado se agruparon las estadísticas y transacciones con valores para: • Memoria RAM: Medios (entre 34% y 36%). • Consumo de CPU: Bajos (entre 7% y 9%). • Latencia: Medios. • Interfaz Monitoreada: Bajos - recepción/transmisión • Tráfico Externo: Medios - entre 38,4Mbps y 99Mbps para recepción - entre 20,5Mbps y 25,88Mbps para transmisión. Distribución de valores para consumo de tráfico externo: Altos: LDAP, HTTP SSL, ICMP, NSMP, BOOTP, QUIC, VoIP, y DB-LSP-DISC2.
A/B/M/B	En este estado se agruparon las estadísticas y transacciones con valores para: • Memoria RAM: Altos (entre 36% y 52,26%). • Consumo de CPU: Bajos (entre 3,5% y 6%). • Latencia: Medios. • Interfaz Monitoreada: Bajos - recepción/transmisión. • Tráfico Externo: Bajos - entre 19,4Mbps y 60,2Mbps para recepción - entre 4Mbps y 12,4Mbps para transmisión. Distribución de valores para consumo de tráfico externo: Medios: ICMP y SNMP. Bajos: LDAP, HTTP, SSL, DB-LSP-DISC2, VoIP y BOOTP.

Tabla 9 Descripción de los estados definidos a través del conjunto de datos.

6.3.2 Análisis de Factores

El objetivo de este experimento fue determinar las variables que más influyeron para que los *switchs* pasen de un estado según lo definido en la **Tabla 9.** Se realizó a través de la aplicación individual de distintas técnicas de minería de datos (correlaciones, componentes principales, árbol de decisiones y regresión logística) para luego analizar en conjunto los resultados obtenidos y formalizar el despliegue del modelo.

Para implementar algunas técnicas de minería, es necesario realizar balanceo de los datos cuando el valor de la variable a utilizar tiene una distribución lo suficientemente desequilibrada como para causar problemas a la mayoría de los algoritmos que generan un modelo de predicción. En las **Figuras 7 y 8**, se puede apreciar la distribución normal de los datos de los grupos y el balanceo de los datos respectivamente.

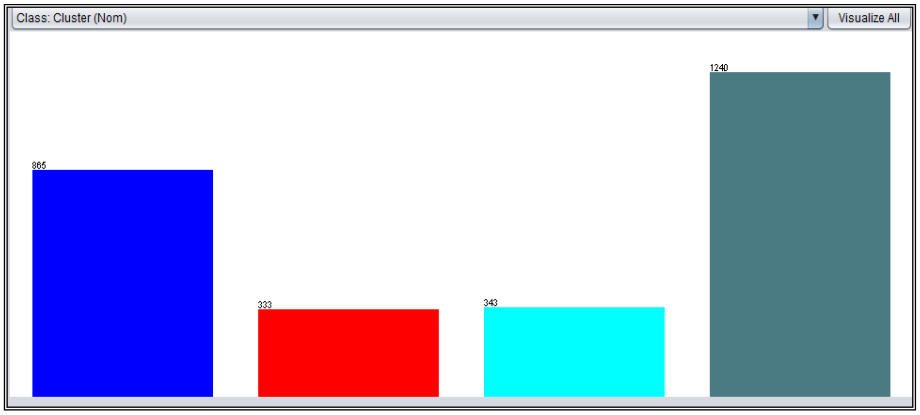


Figura 7 Distribución normal de los datos en los estados

Figura 8 Datos balanceados de los estados definidos a través de filtro SMOTE en Weka

En la **Tabla 10,** se pueden observar las técnicas de minería utilizadas, discriminando cuáles requieren para su implementación aplicar balanceo de los datos. El balanceo es útil cuando se obtiene una distribución desequilibrada de los datos, como es el caso la variable resultante llamada Estados.

Técnica Implementada	Balanceo de datos	Justificación
Correlación entre las variables y la variable resultante Estados	No requiere	La distribución desequilibrada de los datos no afecta el resultado.
Componentes principales PCA	No requiere	La distribución desequilibrada de los datos no afecta el resultado.
Árbol de decisiones	Requiere	La distribución desequilibrada de los datos podría afectar el resultado.
Regresión logística	Requiere	La distribución desequilibrada de los datos podría afectar el resultado.

Tabla 10 Técnicas de minería implementadas

6.3.2.1. Correlación entre las variables y la variable *Estados.*

La técnica de Correlaciones mide el grado de relación entre las variables y la variable dependiente, la cual fue Estados. Para esta técnica el coeficiente de relación va desde -1 a 1, siendo 1 el valor máximo de relación.

Ranked attributes
0.4737 21 CLASSICSTUN
0.4634 3 Memory
0.4501 15 SNMP
0.4233 18 SIP
0.3912 12 LDAP
0.3748 2 CPU_Load
0.3583 17 DB-LSP-DISC
0.3404 19 QUIC
0.3172 8 SSL
0.3068 14 ICMP

Tabla 11 Análisis de correlación entre las variables y la variable estado

Los resultados arrojados por el análisis de correlaciones en la **Tabla 11**, permitieron determinar la relación entre esas variables y la variable Estados, es decir el grado de influencia. Se seleccionaron las 10 primeras variables por orden de relevancia o mayor coeficiente de relación, pues el coeficiente de relación de las restantes fue demasiado bajo.

6.3.2.2. Componentes Principales (PCA)

La técnica de Componentes Principales (PCA) es utilizada para reducir la dimensionalidad (número de variables) de un conjunto de datos, en donde los nuevos componentes principales serán una combinación lineal de las variables originales.

Ranked attributes
1 21 CLASSICSTUN
1 7 Utilization_Transm
1 8 SSL
1 10 RTSP
1 6 Utilization_Rec
1 5 bps_out
1 4 bps_In
1 3 Memory
1 2 CPU_Load
1 9 HTTP

Tabla 12 Análisis de Componentes Principales PCA

Los resultados arrojados por el análisis de Componentes Principales (PCA) en la **Tabla 12**, determinaron la influencia de esas variables sobre la variable Estados.

Para esta técnica se seleccionaron las 10 variables más relevantes de acuerdo al orden de los resultados.

6.3.2.3. Árbol de decisión

Se usó el algoritmo (J48), validación cruzada con 20 hojas y los datos balanceados.

```
=== Classifier model ===
J48 pruned tree
------------------
Memory <= 35.292377
|   SIP <= 332140
|   |   Memory <= 33.584114: cluster0 (20.0/5.0)
|   |   Memory > 33.584114
|   |   |   DB-LSP-DISC <= 6608761: cluster3 (292.0/10.0)
|   |   |   DB-LSP-DISC > 6608761: cluster0 (23.0/11.0)
|   SIP > 332140
|   |   DB-LSP-DISC <= 11038983.010077
|   |   |   bps_In <= 12281825.486189
|   |   |   |   AJNS <= 318927.645858
|   |   |   |   |   SSL <= 1116794674.626502: cluster0 (893.0/8.0)
|   |   |   |   |   SSL > 1116794674.626502
|   |   |   |   |   |   CPU_Load <= 8.822187: cluster2 (31.0/13.0)
|   |   |   |   |   |   CPU_Load > 8.822187: cluster0 (21.0)
|   |   |   |   AJNS > 318927.645858: cluster2 (25.0/11.0)
|   |   |   bps_In > 12281825.486189
|   |   |   |   ICMP <= 284326: cluster0 (21.0/5.0)
|   |   |   |   ICMP > 284326: cluster2 (72.0/13.0)
|   |   DB-LSP-DISC > 11038983.010077
|   |   |   CPU_Load <= 12: cluster2 (1030.0/18.0)
|   |   |   CPU_Load > 12: cluster0 (84.0)
Memory > 35.292377
|   LDAP <= 392880.269909: cluster3 (791.0/3.0)
|   LDAP > 392880.269909
|   |   DCERPC <= 381646: cluster3 (30.0/8.0)
|   |   DCERPC > 381646: cluster1 (1100.0/3.0)

Number of Leaves  :             14
Size of the tree :      27

=== Evaluation result ===
Correctly Classified Instances        4786               97.1777 %
Incorrectly Classified Instances       139                2.8223 %
Kappa statistic                          0.9624
Mean absolute error                      0.0205
Root mean squared error                  0.11
Relative absolute error                  5.4709 %
Root relative squared error             25.3961 %
Total Number of Instances             4925

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,954    0,016    0,951      0,954    0,952      0,937    0,992     0,976     cluster0
               0,988    0,002    0,995      0,988    0,991      0,989    0,999     0,993     cluster1
               0,970    0,010    0,970      0,970    0,970      0,960    0,994     0,982     cluster2
               0,975    0,010    0,971      0,975    0,973      0,964    0,995     0,986     cluster3
Weighted Avg.  0,972    0,009    0,972      0,972    0,972      0,962    0,995     0,984

=== Confusion Matrix ===
   a    b    c    d   <-- classified as
 1163    0   37   19 |    a = cluster0
   0 1217    0   15 |    b = cluster1
  34    1 1197    2 |    c = cluster2
  26    5    0 1209 |    d = cluster3
```

Tabla 13 Resultado del algoritmo del árbol de decisiones J48

De acuerdo al resultado de la evaluación del modelo, ver **Tabla 13**, se obtuvo un 97% de los datos clasificados correctamente. El valor que resultó para el área ROC fue de 0.9, lo que demostró que el resultado tuvo un alto porcentaje de confiabilidad. Además, sobre la diagonal principal de la matriz de confusión se obtuvieron la mayoría de los datos, lo que concibió una deducción cercana a la realidad.

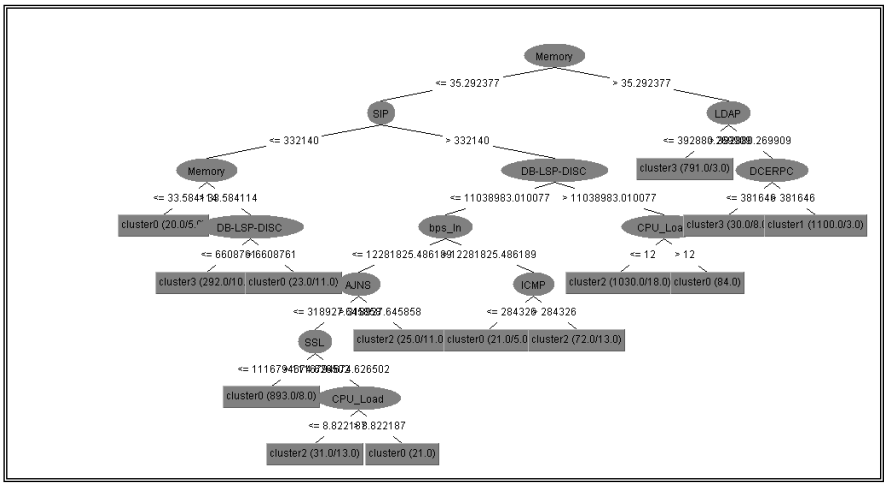


Figura 9 Árbol de decisiones J48

Conforme a los resultados del árbol de decisión, este arrojó que la variable que más influyó en orden de jerarquía para los Estados fue el porcentaje de uso de la *Memoria RAM*, como se puede observar en la **Figura 9.** Por ello es importante hacer chequeo y monitoreo de la memoria RAM, pues podría ser relevante al momento de actualización o reemplazo de equipos, punto importante en la planeación de red.

Como se presenta en la **Tabla 14,** los resultados arrojados por el analisis de la **Tabla13**, seccion Arbol podado "*J48 pruned tree*", permitieron determinar las variables con más influencia para cada uno de los *Estados (tomados como clusteres del 0 al 3)*.

	VARIABLES CON MAYOR INFLUENCIA									
ESTADOS (clúster)	**CPU**	**Memory**	**Bps_in**	**SSL**	**DCERP**	**LDAP**	**ICMP**	**DropBox**	**SIP**	**AJNS**
M/M/B/B	X	X	X	X			X	X	X	X
A/B/A/B		X			X	X				
M/B/M/B	X	X	X	X			X	X	X	X
A/B/M/B		X			X	X		X	X	

Tabla 14 Variables con mayor influencia para cada estado de acuerdo al Árbol de decisiones J48

6.3.2.4. Regresión logística

Se usó el algoritmo de Validación cruzada, utilizando10 iteraciones (*10-fold cross-validation*) y balanceo de los datos. Esta técnica permite estimar la relación existente entre una variable dependiente y un conjunto de variables independientes. Busca modelar como influyen las variables regresoras en la probabilidad de ocurrencia de un suceso particular.

Como se puede ver en la **Tabla 15**, hay variables no representativas en algunos clústeres por la decisión que se tomó del numeral "6.2.2.4 Selección de variables para el proceso de minería/Análisis de correlaciones".

SimpleLogistic:

Class 0 : 80.16 + [Latency] * 0.09 + [CPU_Load] * 0.57 + [Memory] * -2.43 + [Utilization_Transm] * 0.21 + [RTSP] * 0 + [LDAP] * -0 + [STUN] * -0 + [AJNS] * -0 +
[CLASSICSTUN] * 0

Class 1 : -74.87 + [Memory] * 1.35 + [Utilization_Rec] * 0.22 + [Utilization_Transm] * -0.71 + [LDAP] * 0 +
[ICMP] * 0 + [AJNS] * -0 + [CLASSICSTUN] * -0

Class 2 : -35.89 + [LDAP] * 0 + [ICMP] * 0 + [DB-LSP-DISC] * 0 + [SIP] * 0 + [AJNS] * 0 +
[CLASSICSTUN] * 0

Class 3 : 22.49 + [CPU_Load] * -0.16 + [Memory] * -0.14 + [Utilization_Rec] * -0.23 + [LDAP] * -0 + [STUN] * -0 +
[ICMP] * -0 + [SIP] * -0 + [CLASSICSTUN] * -0

=== Evaluation result ===

Correctly Classified Instances	4897	99.4315 %
Incorrectly Classified Instances	28	0.5685 %
Kappa statistic	0.9924	
Mean absolute error	0.0064	
Root mean squared error	0.0504	
Relative absolute error	1.7082 %	
Root relative squared error	11.6407 %	
Total Number of Instances	4925	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,991	0,003	0,991	0,991	0,991	0,988	1,000	1,000	cluster0
	0,998	0,001	0,998	0,998	0,998	0,997	1,000	1,000	cluster1
	0,996	0,002	0,994	0,996	0,995	0,993	1,000	1,000	cluster2
	0,993	0,002	0,995	0,993	0,994	0,992	1,000	1,000	cluster3
Weighted Avg.	0,994	0,002	0,994	0,994	0,994	0,992	1,000	1,000	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
1208	0	8	3	a = cluster0
0	1229	0	3	b = cluster1
4	1	1229	0	c = cluster2
7	2	0	1231	d = cluster3

Tabla 15 Resultados del algoritmo de regresión logística SimpleLogistic

De acuerdo a los resultados de la evaluación del modelo presentado en la **Tabla 15**, se obtuvo un 99% de los datos clasificados correctamente. El valor que resultó para el área ROC fue de 1, lo que demostró que el resultado tuvo un alto porcentaje de confiabilidad. Además, sobre la diagonal principal de la matriz de confusión se obtuvieron la mayoría de los datos, lo que concibió una deducción cercana a la realidad.

Como se presenta en la **Tabla16**, los resultados arrojados por el análisis de la Regresión Logística de la **Tabla 15,** sección *SimpleLogistic* permitieron determinar las variables con más influencia para cada uno de los Estados (tomados como class desde 0 a 3).

	VARIABLES CON MAYOR INFLUENCIA				
ESTADOS (class)	**Latency**	**CPU_Load**	**Memory**	**Utiliza_rec**	**Utiliza_tra**
M/M/B/B	**X**	**X**	**X**		**X**
A/B/A/B			**X**	**X**	**X**
M/B/M/B					
A/B/M/B		**X**	**X**	**X**	

Tabla 16 Variables con mayor influencia por cada estado de acuerdo al algoritmo de regresión logística

6.4 **DESPLIEGUE DEL MODELO DE MINERÍA**

Se analizaron en conjunto los resultados obtenidos en los experimentos, lo que permitió determinar las variables que más influyeron para que los *switchs* incurrieran en los estados. Observar **Tabla 17.**

Variable	Método 1 Correlaciones	Método 2 PCA	Método 3 Árbol	Método 4 Regresión	Cantidad de Veces
Fecha					
Latency				**X**	**1**
CPU_Load	**X**	**X**	**X**	**X**	**4**
Memory	**X**	**X**	**X**	**X**	**4**
bps_In		**X**	**X**		**2**
bps_out		**X**			**1**
Utilization_Rec		**X**		**X**	**2**
Utilization_Transm		**X**		**X**	**2**
SSL	**X**	**X**	**X**		**3**
HTTP		**X**			**1**
RTSP		**X**			**1**
DCERPC			**X**		**1**
LDAP	**X**		**X**	**X**	**3**
STUN					**0**
ICMP	**X**		**X**		**2**
SNMP	**X**				**1**
BOOTP					**0**
DB-LSP-DISC			**X**		**1**
SIP	**X**		**X**		**2**
QUIC	**X**				**1**
AJNS			**X**		**1**
CLASSICSTUN	**X**	**X**			**2**

Tabla 17 Resultados finales de los experimentos

A partir del resultado de la **Tabla 17,** las variables que más tuvieron influencias de acuerdo a las técnicas utilizadas a través de los experimentos se pueden apreciar

en la **Tabla 18**. Es importante resaltar que se pueden ver cuáles son los protocolos más utilizados por los usuarios en el bloque. Si bien el tráfico se genera por protocolos de capas superiores, éste va empaquetado en Ethernet que es un protocolo de capa 2, por lo que genera carga de tráfico en los *switchs.*

Variable Influyente	Descripción
Memoria RAM	Porcentaje de uso Memoria de Acceso Aleatorio. En ella se carga el sistema operativo, los programas y la mayor parte del software.
Carga de la CPU	Porcentaje de uso de Unidad Central de Procesamiento. Tiene como función principal el procesamiento de todas las funciones.
LDAP	Cantidad de tráfico generado por hora por el protocolo Compacto de Acceso a Directorios. Permite el acceso a bases de información de objetos (usuarios, grupos, impresoras etc.) de una red mediante protocolos TCP/IP. La infraestructura de directorios del Bloque 22 está implementada en LDAP.
SSL	Cantidad de tráfico generado por hora por el Protocolo de Seguridad de la Capa de Transporte. Protocolos que aportan comunicaciones seguras por red.
Utilization_Rec	Porcentaje por hora de utilización de la interfaz, al recibir**.**
Utilization_Transm	Porcentaje por hora de utilización de la interfaz, al enviar**.**
bps_In	Tasa de transferencia por hora de entrada de datos en la red**.**
ICMP	Cantidad de tráfico generado por hora por el Protocolo de Mensajes de Control de Internet.
SIP	Cantidad de tráfico generado por hora por el Protocolo de Inicio de Sesiones.
CLASSICSTUN	Cantidad de tráfico generado por hora por el Protocolo Simple Traversal of UDP Through NAT. Útil para Telefonía VoIP, SIP.

Tabla 18 Variables que tuvieron mayor influencia

En la **Tabla 19**, se puede apreciar el porcentaje de tiempo en el que los *switchs* estuvieron en cada estado durante el tiempo total capturado.

	ESTADOS			
EQUIPOS	***M/M/B/B***	***A/B/A/B***	***M/B/M/B***	***A/B/M/B***
A	56%	0	42%	2%
B	62%	0	32%	6%
C	100%	0	0	0
D	14%	0	17%	69%
E	0	36%	0	64%
F	45%	0	22%	33%
G	0	0	0	100%
I	0	36%	0	64%
J	0	36%	0	64%

Tabla 19 Porcentaje de tiempo en que incurrió cada *switch en los estados*

6.4.1 Análisis por cada switch.

Antes de realizar el análisis por *switch*, es importante recordar las condiciones de la captura:

- El trafico analizado sólo se refiere al tráfico principal que se cursó desde y hacia el bloque 22 y el campus universitario de la UPB.

- Se desconocía la existencia de servidores o generadores de tráfico al interior del bloque 22.

- No fue posible hacer capturas en cada uno de los *switchs* para monitoreo de tráfico interno. La información de cada *switch* se refiere solamente al tráfico del puerto principal.

- No había información de las versiones de IOS, por ende, dado que, de acuerdo a los experimentos, la variable porcentaje de uso de memoria RAM es relevante, se consideró mencionar que en ciertas versiones de IOS como la 15.0(2) SE2 para el caso de los *switchs* Cisco Catalyst 2960 gama 2K y 3K, se presentan situaciones de alto consumo de memoria RAM, que tienden a generar un error "*%% Low on memory; try again later*", lo que significa que el dispositivo se queda sin memoria. Esto no permite acceso remoto (vía ssh o telnet) por lo que se debe reiniciar. El fallo es producido por una fuga de memoria conocida como *memory leak* debido a la funcionalidad *Auto SmartPorts* que trae activada por defecto y aplica macros sobre interfaces cuando se presenta un evento (Cisco Systems, 2016). Actualmente no se ha resuelto el bug, pero con el workaround explicado en (Cisco Systems, 2016) se puede evitar la fuga de memoria aplicando y configurando "no macro auto monitor".

Se realizó una descripción por *switch,* en la cual se determinaron los estados y horarios de la semana en los que incurrió durante el tiempo de monitoreo y captura de datos. El tiempo se presentó en formato 24 horas.

Estados	*M/M/B/B*	*A/B/A/B*	*M/B/M/B*	*A/B/M/B*
Porcentaje	**56%**	***0%***	***42%***	***2%***
Horario	Lunes a viernes de las 19:00 hasta las 06:00 horas. Sábados todo el día.		Lunes a viernes de las 07:00 a las 18:00 horas.	Sólo dos sábados continuos, entre las 17:00, 19:00 y 20:00 horas.

Tabla 20 Switch A - Cisco Catalyst 2960-24TC

Para el *switch* A, de acuerdo al resultado de la **Tabla 20,** el funcionamiento fue estándar durante el tiempo de monitoreo y captura de datos. El nivel de consumo de recursos fue medio, sólo aumentó en días y horas laborales, pero no llegó a ser alto. Este es el *switch* principal.

Estado	*M/M/B/B*	*A/B/A/B*	*M/B/M/B*	*A/B/M/B*
Porcentaje	**62%**	***0%***	***32%***	***6%***
Horario	Lunes a viernes de las 18:00 hasta las 06:00 horas. Sábados todo el día.		Lunes a viernes de 07:00 a 17:00 horas.	Sólo dos viernes y un jueves, entre las 17:00 y 21:00 horas.

Tabla 21 Switch B - Cisco Catalyst 2960-24TC

Para el *switch* B, de acuerdo al resultado de la **Tabla 21,** el funcionamiento fue estándar durante el tiempo del monitoreo y captura de datos. El porcentaje para consumo de recursos fue medio, sólo aumentó en días y horas laborales, pero no llegó a ser alto. Se podría suponer que por su comportamiento representa un nivel alto o medio en la jerarquía de la distribución de la topología.

Estado	*M/M/B/B*	*A/B/A/B*	*M/B/M/B*	*A/B/M/B*
Porcentaje	**100%**	***0%***	***0%***	***0%***
Horario	Todo el tiempo del monitoreo y captura permaneció en este estado.			

Tabla 22 Switch C - Cisco

Para el *switch* C, de acuerdo al resultado de la **Tabla 22,** el funcionamiento fue estándar durante todo el tiempo del monitoreo y captura de datos. El nivel de consumo de recursos fue medio y siempre se mantuvo en ese estado.

Estado	*M/M/B/B*	*A/B/A/B*	*M/B/M/B*	*A/B/M/B*
Porcentaje	**14%**	***0%***	***17%***	***69%***
Horario	Lunes a viernes 8:00,12:00,13:00 y 17:00 horas.		Lunes a viernes 09:00,10:00,11:00,14:00,15:00,16:00.	Lunes a viernes de las 18:00 hasta las 07:00 horas. Sábados todo el día.

Tabla 23 Switch D - Cisco Catalyst 2960-24TC

Para el *switch* D, de acuerdo al resultado de la **Tabla 23,** es recomendable realizar análisis más completos, ya que durante esta primera fase del proyecto sólo se tuvo información del enlace principal. Es necesario realizar capturas en cada uno de los puertos del *switch,* que permitan determinar el por qué permaneció en el estado **A/B/M/B** durante días y horas no laborales en más de la mitad del tiempo del monitoreo y de las capturas. Por ello podría suponerse que fue debido al alto consumo de tráfico interno.

Estado	*M/M/B/B*	*A/B/A/B*	*M/B/M/B*	*A/B/M/B*
Porcentaje	**0%**	***36%***	***0%***	***64%***
Horario		Lunes a viernes de las 07:00 hasta las 17:00 horas.		Lunes de las 18:00 horas hasta las 06:00 horas. Sábados todo el día.

Tabla 24 Switch E - Cisco Catalyst 2960-24S

Para el *switch* E, de acuerdo al resultado de la **Tabla 24,** es recomendable realizar análisis completos, ya que durante esta primera fase del proyecto sólo se tuvo información del enlace principal. Es necesario realizar capturas en cada uno de los puertos del *switch,* que permitan determinar el por qué permaneció en los estados **A/B/A/B** durante días y horas laborales y **A/B/M/B** el resto del tiempo del monitoreo y de las capturas. Por ello podría suponer que fue debido alto consumo de tráfico interno.

Estado	*M/M/B/B*	*A/B/A/B*	*M/B/M/B*	*A/B/M/B*
Porcentaje	**45%**	***0%***	***22%***	***33%***
Horario	Lunes a viernes de las 18:00 hasta las 07:00 horas. Pocas horas sábados, alrededor de 8 veces.		Lunes a viernes de las 08:00 hasta las 17:00 horas.	Lunes a sábado durante el día algunas horas intermedias.

Tabla 25 Switch F - Cisco Catalyst 2960-24TC

Para el *switch* F, de acuerdo al resultado de la **Tabla 25,** el funcionamiento fue estándar durante el tiempo del monitoreo y captura de datos. El porcentaje para consumo de recursos fue medio, sólo aumentó en días y horas laborales, pero no llegó a ser alto. Dado que durante esta primera fase del proyecto sólo se tuvo

información del enlace principal, es necesario realizar capturas en cada uno de los puertos del *switch,* que permitan determinar el por qué permaneció en el estado **A/B/M/B,** y analizar el comportamiento en ciertas horas del día implicando picos en consumo de recursos, o si está siendo afectado por otro dispositivo conectado a él.

Estado	***M/M/B/B***	***A/B/A/B***	***M/B/M/B***	***A/B/M/B***
Porcentaje	**0%**	***0%***	***0%***	***100%***
Horario				Todo el tiempo del monitoreo y captura permaneció en este estado.

Tabla 26 Switch G - Cisco Catalyst 2960T 24

Para el *switch* G, de acuerdo al resultado de la **Tabla 26,** teniendo en cuenta que durante esta primera fase del proyecto sólo se tuvo información del enlace principal, es necesario realizar capturas en cada uno de los puertos del *switch,* que permitan determinar el por qué permaneció en el estado **A/B/M/B** durante todo el tiempo de monitoreo de y capturas.

Estado	***M/M/B/B***	***A/B/A/B***	***M/B/M/B***	***A/B/M/B***
Porcentaje	**0%**	***0%***	***0%***	***0%***
Horario				.

Tabla 27 Switch H - Cisco Catalyst 2960-24TC-S

No se obtuvo acceso a información de monitoreo del *switch* H.

Estado	***M/M/B/B***	***A/B/A/B***	***M/B/M/B***	***A/B/M/B***
Porcentaje	**0%**	***36%***	***0%***	***64%***
Horario		Lunes a viernes desde las 07 horas hasta las 17 horas.		Lunes desde las 18 horas hasta las 06 horas y sábados todo el día.

Tabla 28 Switch I - Cisco Catalyst 2960-24G

Para el *switch* I, de acuerdo al resultado de la **Tabla 28,** teniendo en cuenta que durante esta primera fase del proyecto sólo se tuvo información del enlace principal, es necesario realizar capturas en cada uno de los puertos del *switch* que permitan determinar el por qué permaneció en los estados **A/B/A/B** durante días y horas laborales y **A/B/M/B** el resto del tiempo del monitoreo y de las capturas.

Estado	*M/M/B/B*	*A/B/A/B*	*M/B/M/B*	*A/B/M/B*
Porcentaje	**0%**	***36%***	***0%***	***64%***
Horario		Lunes a viernes desde las 07 horas hasta las 17 horas.		Lunes desde las 18 horas hasta las 06 horas y sábados todo el día.

Tabla 29 Switch J - Cisco Catalyst 2960-24G

Para el *switch* J, de acuerdo al resultado de la **Tabla 29,** teniendo en cuenta que durante esta primera fase del proyecto sólo se tuvo información del enlace principal, es necesario realizar capturas en cada uno de los puertos del *switch* que permitan determinar el por qué permaneció en los estados **A/B/A/B** durante días y horas laborales y **A/B/M/B** el resto del tiempo del monitoreo y de las capturas.

7 CONCLUSIONES

La aplicación de metodologías de minería de datos en redes de telecomunicaciones conlleva a impactos positivos en diversos procesos. Teniendo en cuenta técnicas de análisis descriptivos y predictivos se podría mencionar: planeación de la capacidad de red de acuerdo a la cantidad de variables necesarias a incluir dentro del proceso, la identificación de variables que contribuyen con la variabilidad de los niveles de uso de los dispositivos y el descubrimiento automático de conocimiento.

Interpretando los resultados de los diferentes experimentos en conjunto, se determinaron las variables que contribuyeron para que los *switchs* incurrieran en los Estados definidos. Tener en cuenta esas variables (**Tabla 18**), es un punto de inicio para que el CTIC implemente de manera eficaz planeación a través de modelos construidos por medio de herramientas de minería de datos. Estos modelos podrían aportar a encontrar soluciones óptimas o identificar de manera oportuna inconvenientes, y a futuro la creación de nuevas metodologías que contribuyan con una política del CTIC.

Fue de mucha utilidad el perfilamiento de las variables compuestas por las estadísticas y el número de transacciones por hora, lo que permitió determinar los estados en los que podría incurrir cada uno de los *switchs,* pues se pudo obtener información de las variables que influyeron y sus valores de acuerdo a los centroides de cada estado definido. Esto posibilita aportar información para mejores tomas de decisiones respecto a un determinado dispositivo de red o el tráfico que fluye por él.

El modelo resultante puede ser desplegado en la UPB, puesto que permite perfilar los datos para determinar estados en los que incurrirían con más relatividad cada

uno de los nodos de red monitoreados adicionando detalles como rangos de tiempo y las variables que contribuyen para que suceda. Para ellos es necesario obtener mayor flexibilidad en cuanto a acceso a información y recursos oportunamente provistos por parte del CTIC, y por supuesto mayor tiempo de monitoreo y captura de datos y luego el modelo resultante también podría ser aplicado en otras áreas de la Universidad y a futuro en todo el CAMPUS.

El consuno de memoria RAM en los *switchs* varía en los requerimientos mínimos según: la versión del Cisco IOS, la categoría de la capa de distribución (si es Core o de distribución), el número de máquinas conectadas y aplicaciones internas que se ejecuten. Entre los procesos que más consumen memoria se tienen: las *trunking*, VLAN´s, agotamiento de memoria debido a descargas por lista de acceso del usuario o ACL´s, además de la cantidad de memoria utilizada por procesos normales y anormales.

El proceso de minera además de identificar las variables más importantes y sus comportamientos en el análisis, puso en evidencia la importancia de mantener actualizada versión del IOS con el fin de no dar falsas soluciones o análisis incorrectos, teniendo en cuenta que el CTIC no entregó versiones de IOS y tenía desconocimiento de que va conectado a cada punto de su red. Por ello, dado que el análisis del enlace principal proporcionó mucha información para la toma de decisiones, es indispensable monitorear el trafico cursado por cada uno de los puertos y dar respuesta a los estados definidos en la **Tabla 9,** para considerar y/o confirmar algunos supuestos: estudios más completos por cambios de consumos, suposiciones en la rotación de usuarios por conexiones a los AP y ataques de DoS. Además, tener en cuenta que muchas veces si se conoce el estado habitual de la red, se podría detectar anomalías.

El proceso de minería como ejercicio fue importante pues ayudó en el aprendizaje a: evaluar cómo capturar los datos, clasificar el tráfico, el procesamiento de datos antes del modelo, cómo se hacen los experimentos, que tipos de experimentos son importantes y las variables más relevantes. El despliegue le puede proporcionar

información más eficiente al CTIC sobre qué hacer y qué decisiones tomar sobre la red.

El caso de estudio del presente trabajo permitió identificar *switchs* con un comportamiento atípico, en donde se presenta falta de correspondencia entre el tráfico cursado y la permanencia de usuarios en el bloque. Con el fin de hacer un análisis más profundo se sugiere realizar monitoreo de todos los puertos de los *switch*s con este comportamiento.

7. TRABAJOS FUTUROS

Puede ampliarse el análisis basándose no solo en Capa de Transporte, sino también en otras capas. Además implementar series de tiempo de tal forma que sea posible predecir la tendencia o comportamiento de las variables influyentes en periodos de tiempo futuros y de esa manera mejorar la toma de decisiones con respecto a la información obtenida.

Usando modelos implementados a través de minería de datos, crear grupos o estados definidos a través del perfilamiento de los datos para determinar los estados en los que podrían incurrir no sólo los nodos de un bloque, sino de todo el Campus y otras sucursales determinando las variables que más influyen para cada *estado*.

Realizar clasificación de tráfico interno y externo en el Campus e identificar con exactitud lo que sucede con dicho tráfico cursado, así poder determinar problemas de consumo y/o también de saturación u otros, incluso delimitando la zona del problema.

Realizar análisis predictivos a través series de tiempo con datos históricos para obtener información acerca consumos futuros en un canal (principal o extremo a extremo) para toma de mejores decisiones en cuanto a planificación y así evitar escenarios complicados en un determinado momento a futuro.

8. REFERENCIAS

Adeniyi Abidogun, O. (2005). Data Mining, Fraud Detection and Mobile Telecommunications: Call Pattern Analysis with Unsupervised Neural Networks. *University of the Western Cape*.

Alvarez Menendez, J. (2008). Minería de Datos: Aplicaciones en el sector de las telecomunicaciones. *Technical report, Universidad Carlos III*.

Alzate, M., & Peña, N. (s.f.). Modelos de Tráfico en Análisis y Control de Redes de Comunicaciones. *Universidad Distrital, Universidad de los Andes*.

António, N., Salvador, P., & Valadas, R. (2006). Predicting the quality of service of wireless LANs using neural networks. *MSWiM '06 Proceedings of the 9th ACM international symposium on Modeling analysis and simulation of wireless and mobile systems, ISBN:1-59593-477-4*, 52-60.

Britos, P. (2005). Objetivos de Negocio y Procesos de Minería de Datos Basados en Sistemas Inteligentes. *Reportes Técnicos en Ingeniería del Software*.

Cartegnova, S. G. (2005). Detección Automática de Reglas de Asociación. *Instituto Tecnológico de Buenos Aires*.

Casilari, E., Alfaro, A., Reyes, A., Sandoval, F., Electrónica, D. T., & Telecomunicación, E. T. S. I.(n.d.). Modelado neuronal de tráfico ethernet sobre redes mta, (95)

Cernuzzi, L., & Molas, M. L. (s.f.). Integrando diferentes técnicas de Data Mining en procesos de Web Usage Mining. *Universidad Católica "Nuestra Señora de la Asunción"*.

Cisco Systems. (16 de 7 de 2016). *Tools & Resources Bug Search*. Obtenido de https://bst.cloudapps.cisco.com/bugsearch/bug/CSCtz06177

Couche, J., Steine, M., San Vicente, R., & Ferreira, E. (s.f.). Una Aproximación Efectiva a la Detección de Anomalías en el Tráfico TCP/IP Usando Técnicas de Inteligencia Artificial. *Universidad Católica del Uruguay*.

De la Hoz, E., De la Hoz, E. M., Ortiz, A., & Ortega, J. (2012). Modelo de detección de intrusiones en sistemas de red, realizando selección de características con FDR y entrenamiento y clasificación con SOM. *Revista INGE CUC*, Volumen 8, Número 1. pp. 85-116.

Erman, J., Arlitt, M., & Mahanti, A. (s.f.). Traffic Classification Using Clustering Algorithms. *University of Calgary,University Drive NW,Calgary, AB, Canada*.

Escriche Fernández, S. (2011). Predicción de tráfico en redes IP. *Universidad Poliyecnica de Catalunya. Escola Politecnica Superior de Castelldefels*.

García, G. A., & Salcedo, O. (2010). Predicción de Fallos en Redes IP empleando Redes Neuronales Artificiales. *Red, 1(X2), X3.*

Garzón Rodriguez, Y., & Wanumen Silva, L. F. (s.f.). Tratamiento de los datos para el pronóstico de tráfico en redes Wi-Fi, mediante series de tiempo. *Universidad Distrital Francisco José de Caldas Bogotá, Colombia.*

Gildardo, M. A. (2006). Predicción de Tráfico en Redes de Telecomunicaciones basado en Técnicas de Inteligencia Analıtica. *Instituto Politécnico Nacional, México D.F.*

Grajales Bartolo, M. (2011). Análisis de tráfico para la red de datos de las instituciones educativas del núcleo 5 de la ciudad de Pereira. *Universidad Tecnológica de Pereira Colombia*.

Gu, C., Zhang, S., Chen, X., & Du, A. (2011). Realtime Traffic Classification Based on Semi-supervised Learning. *Journal of Computational Information Systems*.

Guedez Maldonado, J. (2005). Propuesta de un Modelo de Predicción de Tráfico para el Acceso a Redes de Banda Ancha. *Universidad Fermín Toro- Barquisimeto*.

Hatonen, K., Klemettinen, M., Mannila, H., Ronkainen, P., & Toivonen, H. (1996). Knowledge discovery from telecommunication network alarm databases. *Proceedings of the Twelfth International Conference on Data Engineering*, (March). doi:10.1109/ICDE.1996.492095.

Henao Ríos, J. L. (2012). Definición De Un Modelo De Seguridad En Redes De Cómputo, Mediante El Uso De Técnicas De Inteligencia Artificial. *Universidad Nacional de Colombia*.

Hernández Suarez, C. A., Martínez Sarmiento, F. H., & Escobar Díaz, A. (2008). Modelamiento y pronósticos de tráfico correlacionado. *Revista Tecnura, 11(22)*, 124-133.

Hernández Suarez, C. A., Salcedo Parra, O. J., & Pedraza Martínez, L. F. (2008). Modelo de trafico Wimax basado en series de tiempo para pronosticar valores futuros de trafico. *Journal of Information Systems and Technology Management*, 5(3), 505-525.

Introducción al Business Intelligence. (2010). *Universidad de Cantabria*.

Ishrat, Z., & Kumar Sharma, S. (2012). *Study of Path Optimization in Packet Switching Network using Neural Network.* Meerut., India: IJCA Special Issue on Issues and Challenges in Networking, Intelligence and Computing Technologies ICNICT(1):23-25.

IT, T. P. (2015). *SolarWinds*. Obtenido de http://www.solarwinds.com/es/

Jamuna, A., & Vinodh Ewards, S. (2013). Efficient Flow based Network Traffic Classification using Machine Learning. *International Journal of Engineering Research and Applications (IJERA)*, Vol. 3, Issue 2, pp.1324-1328.

Juvonen, A., & Sipola, T. (s.f.). Adaptive Framework for Network Traffic Classification Using Dimensionality Reduction and Clustering. *University of Jyvaskyla*.

Koji´c, N., Reljin, I., & Reljin, B. (2006). *Neural network for optimization of routing in communication networks.* Facta universitatis-series: Electronics and Energetics, 19(2), 317-329.

Landa Laredo, M. G. (2009). Aplicación de Redes Neuronales Artificiales en la Ingeniería de Tráfico de Internet. *Revista de Información, Tecnología y Sociedad*, 27.

Lapeña Parreño, J. (2014). Uso de técnicas de aprendizaje automático para reducir las colisiones en redes Wi-Fi. *Universidad de Castilla-La Mancha*.

Lu, W., Tavallaee, M., & Ghorbani, A. (s.f.). Automatic Discovery of Botnet Communities on Large-Scale Communication Networks. *University of New Brunswick*.

Martinez Luna, G. L. (2011). Minería de Datos: Como Hallar una Aguja en un Pajar. *Revista Ciencia, Vol 62*.

Moine, J. M., Haedo, A. S., & Gordillo, S. (2011). Estudio comparativo de metodologías para Minería de Datos. *XIII Workshop de Investigadores en Ciencias de la Computación*.

Molano Vega, D. (27 de Junio de 2013). *MINTIC*. Obtenido de http://www.mintic.gov.co/portal/604/articles-4274_documento.pdf

Montero Lorenzo, J. M. (2008). *Minería de Datos - Técnicas y Herramientas.* Madrid: Thomson Ediciones Paraninfo S.A.

Montesino Pouzols, F. (2009). Mining and control of network traffic by computational intelligence/minería de datos y control de tráfico de red mediante inteligencia computacional. *(Doctoral dissertation, Universidad de Sevilla (US))*.

Munz, G., Li,, S., & Carle, G. (s.f.). Traffic Anomaly Detection Using K-MeansClustering. *University of Tuebingen, Germany*.

Naranjo Cuervo, C. R., & Sierra Martínez, L. M. (2009). Herramienta software para el análisis de canasta de mercado sin selección de candidatos. *Revista Ingeniería e Investigación* , VOL. 29 No. 1.

Neves, J. E., & Leiato, M. J. (1995). Neural networks in B-ISDN flow control: ATM traffic prediction or network modeling? *IEEE COMMUNICATIONS MAGAZINE*.
doi:10.1109/35.466219

Nogueira, A., Salvador, P., & Valadas, R. (2006). Predicting the quality of service of wireless LANs using neural networks. *Proceedings of the 9th ACM International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems - MSWiM '06*, 52. doi:10.1145/1164717.1164728.

Pérez López, C. (2007). Minería de Datos: Tecnicas y Herramientas. *Thomson Editores*.

Reyes Saldaña, J. F., & García Flores, R. (2005). El proceso de descubrimiento de conocimiento en bases de datos. *Ingenierías en Línea*, Vol. VIII, No.26.

Rivero G, Y. C. (2006). Análisis de tráfico de la red del servicio de la administración aduanera del estado zulia. *Télématique: Revista Electrónica de Estudios Telemáticos*, 5(2), 140-160.

Rosemberg Diaz, A. (2007). Diseño e Implementación del centro de operación y gestión de la red académica Peruana en Software libre. *Universidad Católica del Perú*.

Shinde S, S., & Abhang, S. (2012). A Network Traffic Classification Technique using Clustering on Semi-Supervised Data. *Special Issue of International Journal of electronics, Communication & Soft Computing Science & Engineering*.

Torres Álvarez, N. S., Hernandez, C., & Predraza, L. F. (2011). Redes neuronales y predicción de tráfico. *Tecnura, 15(29)*, 90-97.

Urdaneta Montiel, A. J. (2006). Análisis de tráfico en una red LAN aplicando la tecnología de redes neuronales. *Télématique*.

Valencia Zapata, G. A. (2008). *La Minería de Datos Como Herramienta para la Toma de Decisiones Estratégicas.* Recuperado el 27 de Agosto de 2014, de http://www.gustavovalencia.com/app/webroot/img/Documents/DM/Articulo%20DM.pdf

Venkataram, P., Ghosal, S., & Kumar, B. P. V. (2002). Neural network based optimal routing algorithm for communication networks. *Neural Networks : The Official Journal of the International Neural Network Society*, *15*, 1289–1298.

Vicente Altamirano, C. A. (2003). Un modelo funcional para la administración de redes. *Escuela Industrial N.4. Centro de Operación RedUNAM (NOC-UNAM)*.

Vieira Braga, L. P., Ortiz Valencia, L. I., & Ramirez Carvajal, S. S. (2009). *Introducción a la Minería de Datos.* Brasil: E-papers Servicios Editorials Ltda.

Villadango, J., & Magaña, E. (2001). Garantía de calidad de servicio basada en la predicción del ancho de banda. *Universidad Pública de Navarra*.

9. ANEXOS

Anexo A - Ficha Técnica de *Switch* Cisco Catalyst 2960-24PC-L

Descripción del producto	Cisco Catalyst 2960-24PC-L - conmutador - 24 puertos
Tipo de dispositivo	Conmutador - Managed
Factor de forma	Montable en bastidor - 1U
Dimensiones (Ancho x Profundidad x Altura)	44.5 cm x 33.2 cm x 4.4 cm
Peso	5.4 kg
Memoria RAM	64 MB
Memoria Flash	32 MB
Cantidad de puertos	24 x Ethernet 10Base-T, Ethernet 100Base-TX
Velocidad de transferencia de datos	100 Mbps
Protocolo de interconexión de datos	Ethernet, Fast Ethernet
Puertos auxiliares de red	2x10/100/1000Base-T/SFP (mini-GBIC)(señal ascendente)
Protocolo de gestión remota	SNMP 1, RMON 1, RMON 2, RMON 3, RMON 9, Telnet, SNMP 3, SNMP 2c, HTTP, HTTPS, SSH-2
Modo comunicación	Semidúplex, dúplex pleno
Características	Conmutación Layer 2, auto-sensor por dispositivo, soporte de DHCP, alimentación mediante Ethernet (PoE), negociación automática, soporte VLAN, señal ascendente automática (MDI/MDI-X automático), snooping IGMP, snooping DHCP, Quality of Service (QoS)
Cumplimiento de normas	IEEE 802.3, IEEE 802.3u, IEEE 802.3z, IEEE 802.1D, IEEE 802.1Q, IEEE 802.3ab, IEEE 802.1p, IEEE 802.3af, IEEE 802.3x, IEEE 802.3ad (LACP), IEEE 802.1w, IEEE 802.1x, IEEE 802.1s, IEEE 802.3ah, IEEE 802.1ab (LLDP)
Alimentación por Ethernet (PoE)	Sí
Alimentación	CA 120/230 V (50/60 Hz)

Anexo B - Weka KnowledgeFlow – Análisis de Factores