

Establecimiento de patrones de comportamiento de los usuarios de la web

American Motos S.A

Presentado por:

Cristiam leonardo pineda tarazona

Estudiante ingeniería industrial

Director

Javier hernandez cáceres

Docente ingeniería industrial

Universidad Santo Tomás

División de ingenierías y arquitectura

Facultad de ingeniería industrial

2018

Tabla de contenido

1. Definición del problema.....	8
1.1 Descripción del problema.....	8
2. Pregunta de la investigación.....	19
3. Justificación.....	20
4. Objetivos	22
4.1 Objetivo General.....	22
4.2 Objetivos Específicos.....	22
5. Marco referencial.....	22
5.1 Marco teórico.....	22
5.1.1. Proceso de extracción del conocimiento (KDD)	22
5.1.2 Minería de datos Web	25
5.1.3 host.terra.es	28
5.1.4 Técnica de Minería de datos (Algoritmo de Reglas de Asociación).....	29
5.2 Marco Conceptual.....	31
5.2.1 Extracción de Conocimiento (KDD):.....	31
5.2.2 Minería de datos	31
5.2.3 Registro de Acceso	31
5.2.4 Web mining.....	32

5.2.5	Dirección IP:	32
5.2.6	Atributo:.....	32
5.2.7	Regla de asociación:	33
5.3	Marco Legal.....	33
5.3.1	Artículo 61 Constitución Política de Colombia	33
5.3.2	Artículo 56 del Reglamento estudiantil	33
5.3.3	GPL (<i>General Public License</i>) 3.0 for Weka > 3.7.5.....	33
5.4	Marco Histórico	34
5.4.1	Evolución de la Web.....	34
5.4.2	Evolución de la Minería de Datos	35
5.5	Estado del Arte.....	37
6.	Marco Metodológico.....	39
6.1	Fundamentos Epistemológicos.	39
6.2	Diseño de la investigación.....	39
6.3	Fases o etapas del proyecto.	40
6.3.1	Selección del log	40
6.3.2	Configuración software Analog	41
6.3.3	Organización y limpieza del log.	43
6.3.4.	Configuración de ejecutable en WEKA.....	45
6.3.5.	Análisis en ANALOG y WEKA	46

6.4 Estadísticas del servidor web de American Motos	46
6.4.1 Resumen General	47
6.4.2 Informe Mensual	47
6.4.3 Resumen Diario	48
6.4.4 Resumen Horario	48
6.4.5 Informe de códigos de estado	50
6.4.6 Informe del tamaño de los archivos	51
6.4.7 Informe de tipos de archivo	52
6.5 ANALISIS DE LOS DATOS MEDIANTE WEKA	53
6.5.1. Frecuencia de los datos	55
6.6 Datos generados por WEKA	58
7. Resultados	58
7.1 Graficas generadas por WEKA.....	59
7.1.1 Productos más visitados por los usuarios de la página web American Motos .	60
7.2 Reglas de Asociación en WEKA	66
7.3 Estrategias de mercadeo	68
7.3.1 Estrategias para el producto	69
7.3.2. Estrategias para el precio	69
7.3.3. Estrategias para plaza o distribución	70
7.3.4. Estrategias para la promoción o comunicación	70

8. Conclusiones.....72

9. Aportes y recomendaciones.....73

10. Referencias Bibliográficas75

Lista de figuras

Figura 1. American motos.....	41
Figura 2. Log inicial. elaboración propia	44
Figura 3. Tabulación del log. elaboración propia	44
Figura 4. Parámetros weka. elaboración propia.	46
Figura 5. Informe mensual analog. elaboración propia	47
Figura 6. Informe diario. elaboración propia	48
Figura 7. Informe horario. elaboración propia	49
Figura 8. Códigos de estado. elaboración propia	50
Figura 9. Peticiones de estado. elaboración propia	50
Figura 10. Gráfico de bites. elaboración propia	51
Figura 11. Tamaño de descarga. elaboración propia	51
Figura 12. Extensiones usadas. elaboración propia	52
Figura 13. Petición de extensión. elaboración propia.....	53
FIGURA 14. ARCHIVO ARFF. ELABORACIÓN PROPIA	54
FIGURA 15. PANTALLA INICIO WEKA. ELABORACIÓN PROPIA	55
Figura 16. Frecuencias ip. elaboración propia	56
Figura 17. Ip más frecuente. elaboración propia	56
Figura 18. Ip de mayores visitas extranjera. elaboración propia.....	57
Figura 19. Ip ecuatoriana. elaboración propia.....	57
Figura 20. Weka. elaboración propia.....	58
Figura 21. Graficas weka. elaboración propia.....	59
Figura 22. Url más visitadas. elaboración propia	60

Figura 23. Llanta más visitada	61
Figura 24. Llanta h923.....	62
Figura 25. Llanta z8.....	63
Figura 26. Trafico ip-hora. elaboración propia	64
Figura 27. Matriz hora-día. elaboración propia.....	65
Figura 28. Matriz tráfico por mes. elaboración propia	66
Figura 29. Reglas asociación. elaboración propia	67
Figura 30. Vista american motos web	73
Figura 31. Pantalla llantas.....	74

1. Definición del problema

1.1 Descripción del problema

El manejo de la información en las empresas en la actualidad surge a partir de todos los niveles de la organización, permitiendo la generación de conocimiento. Para esto es importante “la capacitación del personal para lograr que la información se convierta en datos útiles con los cuales se puedan tomar decisiones, incentive al desarrollo del conocimiento” [1].

El manejo de la información ha evolucionado en las últimas décadas, lo cual ha fomentado el desarrollo continuo de las organizaciones:

En la década de 1970 se produjo la llamada *revolución telemática*¹. Aparecieron los primeros sistemas de información que utilizaban la informática y las telecomunicaciones de forma conjunta y coordinada: las bases de datos en línea y el videotexto fueron los primeros exponentes de la era de la información digital. Es a partir de este preciso momento cuando, además de un crecimiento cuantitativo de las fuentes de información, nos encontramos también frente a una disparidad cualitativa en relación al soporte material en que se registran, y cuando el papel empezará a perder la primacía como soporte portador de cultura y se producirá una notable diversificación de soportes documentales. [2]

Entonces nos adentramos en una nueva sociedad de la información, en la cual se pueden almacenar grandes cantidades de datos en documentos digitales, que a su vez ha generado una gran necesidad como se cita:

¹ La telemática es la disciplina científica y tecnológica que analiza e implementa servicios y aplicaciones que usan tanto los sistemas informáticos como las telecomunicaciones, como resultado de la unión de ambas disciplinas.

Las sociedades de la información se caracterizan por basarse en el conocimiento y en los esfuerzos por convertir la información en conocimiento. Cuanto mayor es la cantidad de información generada por una sociedad, mayor es la necesidad de convertirla en conocimiento. Otra dimensión de tales sociedades es la velocidad con que tal información se genera, transmite y procesa. En la actualidad, la información puede obtenerse de manera prácticamente instantánea y, muchas veces, a partir de la misma fuente que la produce, sin distinción de lugar. Finalmente, las actividades ligadas a la información no son tan dependientes del transporte y de la existencia de concentraciones humanas como las actividades industriales. Esto permite un reacondicionamiento espacial caracterizado por la descentralización y la dispersión de las poblaciones y servicios. [3]

Dentro del marco de la competitividad nace un concepto que describe la gestión del conocimiento a partir de fuentes de información conocido como, el sistema nervioso digital de una empresa, que en otras palabras es la extracción de conocimiento a partir de información tratada mediante un análisis metodológico. El famoso empresario Bill Gates menciona:

El sistema nervioso digital no garantiza las respuestas correctas a las preguntas. Pero sí elimina toneladas de papel, junto con los viejos procesos, librando tiempo para que sea posible reflexionar sobre aquéllas. Proporcionará datos que darán el pistoletazo de salida a esa reflexión y los presentará de manera que permitirá distinguir las tendencias que vienen. Además, el sistema nervioso digital hará posible que los datos y las ideas broten desde los escalones inferiores de la organización, donde está la gente que tiene la información sobre esas preguntas... y hasta es posible que muchas de las respuestas. Y lo más importante, permitirá que todo eso se haga con celeridad. El punto crítico es que los directivos dispongan de la información que les permita saber cuál es su ventaja competitiva, y cuál

podría ser su próximo gran mercado.(Gates & Bravo, 1999) . Surge entonces la posibilidad para las organizaciones de detectar posible mercado potenciales o encontrar inconsistencias en su modelo de negocio.

La información se puede extraer de diferentes fuentes, estas se clasifican en: “fuentes internas (registros de clientes, productos, ventas, contactos, promociones, visitas, proveedores, empleados) y fuentes externas (estudio de mercados, información de sectores, investigación y benchmarking)” [4].

Una de las fuentes de información más importantes es conocida como la base de datos (*database*) de un servidor web. Las empresas suelen contratar los servicios de un host para el registro de información y la publicación de dominios (páginas web).

Los desarrolladores, diseñadores y encargados de su mantenimiento tienen la formación para medir la eficiencia de su base de datos, para lograrlo se debe cuestionar ¿cómo están los clientes usando la base de datos? La última pregunta requiere el análisis de la estructura de los *hyperlinks* así como de los documentos que contiene la base de datos. El producto del análisis debe incluir:

- La frecuencia de visitas por documento
- Las visitas más recientes por documento
- Quién está visitando qué documento
- La frecuencia de uso de cada *hyperlink*

Toda esta información es el punto de partida para un análisis más sofisticado y el desarrollo de nuevas herramientas que cumplan la misma función. [5]

La información que podemos recolectar a partir de bases de datos de un servidor web recibe un tratamiento diferente a la información que podemos adquirir con la realización de encuestas, inspección, etc..., debido a que:

Internet hoy en día afecta al comportamiento del consumidor, puesto que modifica el lugar, el momento y la forma en que se producen las compras. También es un elemento que cambia la forma en que se desenvuelven los vendedores, ya sean estos intermediarios o fabricantes. Los fabricantes encuentran en Internet un sistema para acercarse al cliente, y en ocasiones tal acercamiento elimina la necesidad de toda entidad que ejerza la labor de intermediario. También la Red tiene sus repercusiones para con las instituciones que regulan el mercado, mostrándose las principales en la regulación de un nuevo sistema de venta, notablemente diferente al actual y con una evolución arrolladora. Finalmente, el entorno que nos rodea sufre también consecuencias derivadas de la aparición y desarrollo de Internet. [6].

Actualmente existen varios problemas con los que nos topamos los usuarios al usar la web son: “Encontrar información relevante, crear nuevo conocimiento, presentación de la información. La enorme cantidad de información disponible hace de la web un área fértil para la minería de datos, cuyas técnicas puede solucionar las problemáticas ya mencionadas” [7, p. 546]

El mercado actual ha facilitado el acceso del consumidor al producto, ofreciendo además la posibilidad de comparar características y precios desde la comodidad de su ordenador, este acercamiento directo con el consumidor ha producido un gran estallido en el comercio electrónico. El famoso empresario Bill Gates mencionó en su libro “Los negocios en la era digital” (resumido por Jos Antonio Bravo) lo siguiente:

Mientras se produce la explosión del comercio electrónico, el intermediario capaz de hallar maneras creativas de utilizar la Red no será el único que fortalezca sus relaciones con los clientes. Despuntará el comerciante capaz de abordar el e-commerce como algo más que una especie de caja registradora digital. La venta siempre es el objetivo último, naturalmente, pero en sí misma no será más que una parte de la nueva experiencia de clientela on-line. Algunas empresas utilizarán la Red para relacionarse con sus clientes de maneras que antes no eran posibles; la venta pasará a formar parte de una secuencia de servicios al cliente, para lo cual Internet ofrece posibilidades inéditas. [8]

La visión de Bill Gates tuvo un gran acercamiento a la realidad del comercio electrónico en los últimos años. El comercio electrónico ha evolucionado en cuanto a la relación con el cliente, esto se debe a que las empresas invierten cada vez más en técnicas de mercadeo electrónico que les permiten tener un acercamiento más sofisticado con el cliente, además se facilita el registro de estos eventos, lo cual se traduce finalmente en información, que puede aprovecharse para generar conocimiento.

La era digital ha evolucionado a una era electrónica en la que el flujo de información es mucho más extenso, variado y sistemático. Prácticamente no existe hoy en día una faceta de la realidad de la cual no se disponga información de manera electrónica, ya sea estructurada, en forma de bases de datos, o no estructurada, en forma textual o hipertextual. Desgraciadamente, gran parte de esta información se genera con un fin concreto y posteriormente no se analiza ni se integra con el resto de información o conocimiento del dominio de actuación. Afortunadamente, la situación ha cambiado de manera significativa respecto a unos años atrás, donde el análisis de datos se realizaba exclusivamente en las grandes corporaciones gobiernos y entidades bancarias, por departamentos especializados con nombres diversos: planificación y prospectiva, estadística,

logística, investigación operativa, etc. Tanto la tecnología informática actual, la madurez de las técnicas de aprendizaje automático y las nuevas herramientas de minería de datos de sencillo manejo, permiten a una pequeña o mediana organización tratar grandes cantidades de datos almacenados en las bases de datos (propias de la organización, externas o de la web). [7, p. XIV]

En cuanto al sector colombiano, según el DANE y el DPN “el censo nacional de empresas en el país presenta una estadística para el año 2005, la cual indica que el porcentaje de Microempresas en Colombia representan el 96,1% del total de los establecimientos y tan solo el 3.2% pertenecen a pequeñas empresas”. [9]. Tanto las micro como pequeñas empresas que utilizan una página web para promover sus productos o servicios cuentan en su mayoría con registros electrónicos de acceso y registros acerca del movimiento de su actividad económica y sus clientes, sin embargo, este registro de información sólo se usa como soporte legal del desarrollo de dichas operaciones comerciales. De esta manera es notorio que se está desaprovechando una gran oportunidad en la amplia mayoría de las empresas del país, ya que en su mayoría son micro y pequeñas empresas y el análisis de los registros de información que estas guardan puede aprovecharse para tomar mejores decisiones y mejorar su modelo de negocio. Por otro lado, esto también fomentaría a nivel general el desarrollo de las micro y pequeñas empresas, que con el tiempo pueden crecer a partir de la toma de mejores decisiones, las cuales evitan en muchos casos el estancamiento o la quiebra.

Uno de los problemas que evita que el manejo y análisis de la información en las empresas de menor tamaño es el desconocimiento de la materia y su falta de interés por experimentar nuevas alternativas para mejorar su modelo de negocio, ya que en muchos casos se maneja una actitud conformista, lo cual provoca un estancamiento en las organizaciones de menor tamaño. Otra razón es también la falta de motivación y la difícil situación económica que experimentan este tipo de

empresas. Sin embargo, se puede notar que últimamente la evolución de las redes sociales y nuevas herramientas como las aplicaciones móviles han permitido que cualquier tipo de organización pueda aprovechar los beneficios que estas herramientas ofrecen, fortaleciendo así la relación con los clientes y facilitando el procedimiento de compra, lo cual se traduce a mayor satisfacción del cliente y por lo tanto mejor acogida en el mercado.

Por otro lado, no hay que dejar de lado el importante papel de las páginas web comerciales en las empresas ya que son la imagen virtual de la empresa y representan gran parte de la estructura comercial, además es la fuente de información más importante y manejable a la hora de acceder a la información y tratarla mediante herramientas de análisis de patrones y tendencias.

Como uno de los propósitos que tengo con este proyecto es que la metodología utilizada para el análisis de los registros, además de la información acerca del panorama actual de la materia, sirva como un soporte para futuras investigaciones, y de esta forma motivar a los empresarios e investigadores a profundizar en el tema del análisis de registros de información en las empresas comerciales con el fin de obtener respuestas e implementar mejoras en el modelo de negocio.

El proyecto se apoya en las siguientes teorías:

En primer lugar, dentro del contexto empresarial, el comercio electrónico ha permitido el análisis de tendencias y patrones de consumo de los clientes con mayor rapidez, cantidad y asertividad.

La gestión del conocimiento a partir de fuentes de información abarca un alto grado de complejidad y depende tanto de la destreza del empresario para tomar decisiones, como la aplicación de un modelo de minería de datos (MD). La minería de datos constituye una de las fases del “proceso de descubrimiento de conocimientos en bases de datos” (*Knowledge Discovery in*

Databases, KDD. Hernández Orallo y colaboradores definen el KDD como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos". Este proceso no solo incluye la obtención de modelos o conjunto de patrones, también se refiere a la evaluación y posible interpretación de estos modelos. [10]

En segundo lugar, complementando el proceso KDD, está la aplicación de un modelo de minería de datos; la naturaleza de la información de los registros de acceso permite aplicar la minería de datos de manera descriptiva.

Las técnicas de minería web difieren significativamente de las demás técnicas de minería de datos, ya que la web es un repositorio de gran tamaño donde los documentos contienen datos de diferentes tipos (texto, imágenes, audio, etc). Además, los documentos son hipertexto o hipermedia, al hacer referencias a otros documentos a través de hipervínculos que pueden ser recorridos por distintos usuarios, esta diversidad permite minar la web basándose en tres conceptos: el contenido, estructura y uso. [7, p. 545].

“La minería web puede descomponerse en las siguientes subtareas: descubrimiento de fuentes, selección y pre-procesado de la información, generalización y análisis” [7, p. 548]. Para la ejecución del proceso de minería web: “como herramienta no comercial para el análisis de archivos log de servidores web, se cita Analog (<https://www.analog.cx>)” [7, p. 568].

En tercer lugar, como técnica principal de análisis de la información está la aplicación de reglas de asociación y agrupamiento. “Primero genera grupos de las sesiones basándose en patrones de co-ocurrencia de referencias URL o bien basándose en cuán a menudo las referencias

URL ocurren juntas en las sesiones de los usuarios (usando la técnica ARPH, Association Rule Hypergraph Partitioning)”. [7, p. 568].

En contraposición a la minería de datos aparece la primera teoría conocida como el manejo de la información “Big Data”.

Big Data es un término para un conjunto de datos de gran tamaño. Conjuntos de grandes datos son aquellos que superan el simple tipo de arquitecturas de bases de datos y manejo de datos que se utilizaron en los primeros tiempos, cuando el procesamiento de los grandes datos era más caro y menos factible. En general, la minería de datos se refiere a las operaciones que implican búsquedas relativamente sofisticadas que devuelven resultados específicos y concretos, a diferencia hace referencia a grandes cantidades de datos que superan la capacidad de procesamiento habitual del software informático existente. [11]

En segundo lugar, respecto a la técnica a utilizar (reglas de asociación), existen otras técnicas que tienen el potencial de obtener conocimiento útil de la información, entre ellas está la técnica de agrupación de Clúster.

Las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente (entre los miembros de la misma clase) y a la vez diferente entre los miembros de las diferentes clases. [12]

En tercer lugar, está la teoría del modelo de redes neuronales de minería de datos, que difiere directamente del sistema nervioso digital que menciona Bill Gates, ya que el anterior se enfoca en el fortalecimiento de la comunicación de las partes y el tratamiento de la información

adquirida a partir de esta interacción con metodologías de agrupación y reglas de asociación. Por el contrario, el modelo de redes de información tiene “un foco en matemáticas y estadística. Se basan en una idea sencilla: dados unos parámetros hay una forma de combinarlos para predecir un cierto resultado” [13].

Dentro de la experiencia de otros proyectos ya realizados respecto a la temática de Minería de datos está proyecto titulado “Práctica de Minería de Datos, Introducción al Weka” desarrollado por José Hernández Orallo y Cesar Ferri Ramirez en la Universidad Politécnica de Valencia. Este documento ofrece una introducción al manejo de la herramienta de minería de datos conocida como Weka, con ayuda de ejemplos respecto a diferentes problemas relacionados con las metodologías de clasificación, agrupación, reglas de asociación y dependencias. Además, como Libro guía se tendrá el Libro “Introducción a la Minería de Datos” de los autores anteriores incluyendo José Ramirez Quintana.

El trabajo que más se acerca a la metodología que se usará es titulado “ ESTUDIO WEBMINING Y DATAMINING DEL LOG HUAWEI LTDA” realizado por Javier Ulises Arias Rivera y Cesar E. Polo Castro en la Universidad de Magdalena en el año 2012. Este estudio muestra la aplicación de una metodología de Minería de datos para establecer reglas de asociación y definir patrones y proporciones respecto a los registros de acceso (Logs) de los visitantes de la página Web de la empresa Huawei LTDA.

El artículo científico publicado en la revista IEE “*Association Rule Mining for Web Usage Data to improve Websites*” realizado por Avadh Kishor, Ajeet Kumar, Ashish K. Maurya en la Universidad, UP, India. Es un estudio de aplicación de minería de datos con la técnica de reglas de asociación para mejorar una web a partir del análisis de datos de uso de la misma. Los datos son extraídos automáticamente de un web log, además el modelo posee un script de pre-

procesamiento de la información para su respectiva clasificación y limpieza. La técnica de reglas de asociación utiliza el algoritmo a priori propuesto como (PNAR_1MLMS).

Dentro del marco legal este proyecto se atiene al artículo 61 de la Constitución Política de Colombia. “El Estado protegerá la propiedad intelectual por el tiempo y mediante las formalidades que establezca la ley”. [14]

El artículo 56 del Reglamento estudiantil de la Universidad Santo Tomás, en su párrafo señala: “para los trabajos escritos, los estudiantes deberán emplear debidamente las referencias bibliográficas y documentales...” [15].

El software que se utilizará para la creación del modelo de minería de datos es WEKA, un software libre desarrollado por la universidad de Waikato, disponible en: <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

WEKA está licenciado bajo la licencia GNU General Public (GPL 2.0 para Weka 3.6 y GPL 3.0 para Weka > 3.7.5). Cualquier trabajo derivado obtenido bajo esta licencia debe ser licenciado bajo la GPL si este trabajo derivado es distribuido a un tercero. Para proyectos comerciales que requieren la capacidad de distribuir código WEKA como parte de un programa que no se puede distribuir bajo la GPL, puede ser posible adquirir una licencia apropiada de los propietarios de derechos de autor enumerados en las clases Java correspondientes. El copyright de la mayoría del código WEKA es propiedad de la Universidad de Waikato. [16].

Para el desarrollo de este proyecto la fuente de información es un Log (Registro de eventos llevados a cabo durante un periodo de tiempo) de una página web comercial. Debido a que estos registros contienen información acerca de la fecha, hora, acción y ruta de acceso proporcionada por los usuarios, nace en consecuencia la oportunidad de extraer información y conocimiento de

estos registros. “El problema de acceder a un Log gratuito es que estos contienen información confidencial y sensible de las organizaciones, por lo tanto, estas prefieren ocultarlos para el público en la mayoría de los casos”. [17]. Existen dos formas de adquirir un Log, la primera es solicitarlo a una empresa bajo un contrato de confidencialidad y la segunda es adquirir un Log gratuito del servidor de una página web.

Como aún no se ha adquirido el Log apropiado para el proyecto, lo primero que se debe hacer es adquirir un Log fiable con información reciente, ya que en el caso de los Logs gratuitos, muchos están desactualizados, por esta razón encontrar un Log pertinente al proyecto requiere de una consulta previa, una vez adquirido el log se procede con la instalación de la herramienta de extracción de información, para finalmente aplicar todo el proceso de descubrimiento de conocimientos en bases de datos (Knowledge Discovery in Databases, KDD).

Como resultado obtendremos conocimiento acerca del estado, tendencias y comportamiento de compra de los consumidores que acceden a la página web de la empresa implicada. Lo cual ofrece la oportunidad de darle un uso más adecuado a la información, con la extracción de conocimiento mediante reglas de asociación y el establecimiento de patrones de comportamiento como actividad central. El conocimiento adquirido puede ser utilizado como un recurso para el análisis de la situación de la página web de la empresa en cuestión y además para a propuesta de acciones de mejora.

2. Pregunta de la investigación

¿Qué patrones de comportamiento representados por reglas de asociación se pueden destacar del registro de acceso (*access log*) de la compañía American Motos, analizado respecto a la actividad de los usuarios?

3. Justificación

En la actualidad el acceso a la información digital ha tomado las riendas de la generación de conocimiento, la exposición de ideas y el desarrollo de nuevas herramientas para el estudio de tendencias que impacta el desarrollo del mercado y la economía en general. Se puede denominar la sociedad actual como una “sociedad de información” debido a:

Los desarrollos tecnológicos que se consolidaron en la última década del siglo: la televisión, el almacenamiento de información, la propagación de video, sonido y textos, han podido comprimirse en soportes de almacenamiento como los discos compactos o a través de señales que no podrían conducir todos esos datos si no hubieran sido traducidos a formatos digitales. La digitalización de la información es el sustento de la nueva revolución informática. Su expresión hasta ahora más compleja, aunque sin duda seguirá desarrollándose para quizá asumir nuevos formatos en el mediano plazo, es la Internet. [18]

El manejo de la información digital se ha convertido en una gran oportunidad para las empresas, en cuanto a su capacidad para tomar decisiones con mayor rapidez, y en la implementación de mejoras en la cadena de valor, debido a que: “el principal uso de que se da a los sistemas de información es el de optimizar el desarrollo de las actividades de una organización con el fin de ser más productivos y obtener ventajas competitivas” (Laudon & Laudon, 1996)

“Es indudable que las actividades online son fuentes inagotables de datos, con una importancia muy superior a otras fuentes de origen en lo que a cantidad y variedad de datos se refiere (y, por supuesto, en lo que concierne a la velocidad con la que se generan). “ [19]. La información que se puede adquirir de la Web es variada, sin embargo, para este caso el manejo de los registros de información que contienen fuentes documentales e hipervínculos es la pauta para el desarrollo del proyecto.

El tratamiento de la información ofrecida por los registros mediante reglas de asociación corresponde a unas de las técnicas pertinentes para el caso. La implementación de esta técnica ofrecerá resultados contundentes acerca de los patrones de comportamiento de los usuarios que acceden a la página web de una empresa comercial, es decir que se podrán especificar las tendencias de estos comportamientos, concluir y proponer mejoras.

Una página web de buena calidad debe cumplir con requisitos como: funcionalidad, fiabilidad, eficiencia, capacidad de mantenimiento y portabilidad. Se quiere demostrar que: “con la aplicación de minería web se puede crear o mejorar una página web de buena calidad al encontrar factores ocultos que puedan afectar aquellos sitios web de actividad comercial.” [20]

Para empezar la metodología se adquirirá un registro de acceso (*Access log*) de la compañía American Motos, sustentando su uso como una actividad netamente académica y sin ánimo de lucro, al adquirir el *log* se procederá en primer lugar a realizar una limpieza de la información con la herramienta Analog, seguido de la organización de la información limpia con ayuda de Microsoft Excel, después, se preparará la información en un documento de texto ejecutable en el software de uso libre de minería de datos “WEKA” con la extensión “arff”. A continuación, se aplicará la metodología, con su análisis respectivo.

El desarrollo de este proyecto tiene un fin netamente académico, es decir que la información tratada será de libre acceso para futuros estudiantes o profesionales más experimentados, fomentando así el flujo de la información y el interés de todas las partes académicas y empresariales en el tema de la minería de datos web. Es decir que se espera generar un impacto positivo en la recurrencia de documentos relacionados con esta línea de investigación, beneficiando a las partes que accedan a esta información y a mí personalmente como carta de

presentación en mi formación como investigador. Además, se espera optar por el título de Ingeniero Industrial con el desarrollo y sustentación de este proyecto.

4. Objetivos

4.1 Objetivo General

Establecimiento de patrones de comportamiento de usuarios de la empresa American Motos por medio de web minning.

4.2 Objetivos Específicos

- Identificar tendencias de los clientes de la empresa American Motos por medio de reglas de asociación.
- Caracterizar los clientes de la empresa American Motos utilizando el log del servidor.
- Analizar la información generada por el log del servidor de la empresa American Motos para generar estrategias de mercadeo de productos y aumentar la efectividad de sus campañas promocionales.

.

5. Marco referencial

5.1 Marco teórico

5.1.1. Proceso de extracción del conocimiento (KDD)

El desarrollo del proyecto parte de un proceso fundamental conocido como KDD (*Knowledge Discovery in Databases*, Descubrimiento de Conocimiento en Bases de datos), también llamado “proceso de extracción de conocimiento”. Hernández Orallo y colaboradores describe al KDD como “un proceso complejo que incluye no sólo la obtención de los modelos o

patrones, sino también la evaluación y posible interpretación de los mismos. Así, los sistemas de KDD permiten la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento”. [7, pp. 13, 14]. Por otro parte Jiawi Han y colaboradores describen el mismo proceso como: “un proceso que extrae información de calidad que puede usarse para dibujar conclusiones basadas en relaciones o modelos dentro de los datos; las etapas del proceso KDD se dividen en 5 fases: selección de datos, preprocesamiento, transformación, minería de datos, interpretación y evaluación”. [21]. Se puede describir entonces un proceso KDD como un proceso de extracción de conocimiento que parte de la recolección y selección de información de calidad, que permita proyectar y analizar la información mediante técnicas de minería de datos, con el fin de extraer patrones y modelos de información que entreguen conocimiento útil para su uso.

Fases del proceso de descubrimiento de conocimiento en bases de datos, KDD. Extraído del libro “introducción a la minería de datos”, por Orallo, Quintana, & Ramirez, 2004.

- Fase de integración y recopilación: En esta fase se pretende realizar una selección de las fuentes de información, una vez seleccionadas se procede a la recolección de la información que será objeto del tratamiento. En cuanto a la integración, si el volumen de la información no es muy grande, se puede trabajar con los datos originales en formatos heterogéneos, si el volumen de la información es muy grande y procede de diferentes fuentes y formatos, se suele acceder un servicio de *data warehousing*, que consiste en una tecnología de almacenes de datos, es decir la integración de múltiples fuentes de datos.
- Fase de selección, limpieza y transformación: En esta etapa se selecciona la información relevante respecto al problema a tratar, teniendo en cuentas aspectos

como el formato, la estructura y su naturaleza, de igual forma se desecha la información que puede perjudicar la calidad y veracidad de los resultados. Además, se pueden crear modelos usando una muestra de datos, además de la construcción de atributos y la transformación de atributos en variables numéricas y viceversa.

- Fase de minería de datos: consiste en la construcción de un modelo basado en los datos recopilados. El modelo es una descripción de los patrones y relaciones entre los datos que puede usarse para hacer predicciones. Para la construcción de un modelo de minería de datos se deben tomar las siguientes decisiones antes de empezar el proceso:
 1. Determinar qué tipo de tarea de minería es la más apropiada.
 2. Elegir el tipo de modelo a aplicar a dicha tarea.
 3. Elegir un algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo que estamos buscando.
- Fase de Evaluación e interpretación: En esta fase se debe medir la calidad de los patrones descubiertos por un algoritmo de minería de datos. Idealmente los patrones descubiertos deben tener tres cualidades: ser precisos, comprensibles e interesantes
- Fase de difusión, uso y monitorización: Una vez construido y validado el modelo puede usar con dos finalidades: para que un analista recomiende acciones basándose en el modelo y en sus resultados, o bien para aplicar el modelo a diferentes conjuntos de datos. Tanto en el caso de una aplicación manual o automática, es necesaria su difusión. [7, pp. 21-39]

5.1.2 Minería de datos Web

La minería de datos web es una adaptación del concepto y función de la minería de datos enfocado al mejoramiento de la estructura de un sitio web y la extracción de conocimiento potencialmente útil para las organizaciones.

Etelligent define la minería de datos web como “una serie de técnicas encaminadas a obtener inteligencia a partir de datos procedentes de la web. Aunque las técnicas utilizadas tienen su raíz en las técnicas de data mining o minería de datos, presentan características propias debido a las particularidades que presentan las páginas webs”. [22].

Gyves Camacho FM menciona tres clases fundamentales de minería web, a partir de su artículo “web mining: fundamentos básicos” .

- **Minería Web de Contenido:** En la Web existen variados documentos, desde hipertexto, imágenes, vídeos, audio, símbolos, datos, meta-datos, link, textos, hasta pdf, lo cual dificulta de su clasificación, la minería de datos web extrae conocimiento del contenido de la página web y propone una mejor organización de este contenido para facilitar el acceso y la recuperación de la información
- **Minería Web de Estructura:** Da a conocer el esqueleto del sitio web y cómo se está navegando por él, con el fin de detectar falencias en la estructura y las rutas de acceso a las diferentes secciones de información.
- **Minería de uso Web.** Tiene como principal objetivo extraer patrones de uso del Web por parte de los usuarios. Para esto, se utilizan los archivos Log (registros de sucesos/eventos) de los servidores Web.

Este tipo de minería tiene dos objetivos principales:

1. Extraer patrones generales de uso de un sitio Web de manera que pueda reestructurarse para que sea más fácil de utilizar y mejore el acceso por parte de los usuarios.
2. Obtener perfiles de los distintos tipos de usuarios a partir de su comportamiento y navegación, para ofrecer una atención más personalizada.

El procesamiento de Log que se genera automáticamente en los servidores produce información de alto valor. Los datos almacenados en los Log siguen un formato estándar y se almacenan en un archivo de texto, separado cada campo por comas (",") y cada acceso es un renglón distinto.

Entre los datos que registran los llamados Log se encuentran:

- Dirección IP del usuario.
- Fecha y hora de acceso.
- URL de la página accedida.
- Protocolo utilizado para la transmisión de los datos.
- Código de error.
- Número de bytes transmitidos.

[23]

Este proyecto se desarrolla bajo el marco de una metodología de minería de uso web, es decir que se requiere que la fuente de información provenga directamente desde un registro de acceso (*access log*).

Cuando se instala el servidor de una página, ya sea Apache, AOLserver o un “Servlet Engine” como Tomcat, estos generan un registro sobre cada solicitud de página del sitio en cuestión, estos registros (“Logs”) contienen información útil no sólo para fines técnicos, sino también para fines de mercadotecnia, ya que contienen información como:

- Primeramente, el nodo IP del usuario (“Navegador”) que solicitó la información del sitio, esto permite identificar el origen de nuestros visitantes, sea país o empresa.
- Contiene la hora de acceso, a través de este dato es posible detectar horas pico del sitio y con esto programar manutención y/o actualizaciones.
- El tipo de Navegador (“Netscape”, “Explorer” u otro) y Sistema Operativo del usuario que nos visita; esto nos permite saber si nuestros documentos con “JavaScript”, “Applets” o “Flash” han sido apreciados en su totalidad. (Véase también Como opera un navegador y sus consideraciones).
- Si fue exitoso el envío de información o el usuario recibió un mensaje de página no encontrada (el afamado 404 NOT FOUND)

Y otros datos más... [24]

Por lo general la mayoría de los registros de acceso manejan una estructura estándar desarrollado por la NCSA - Centro Nacional de Aplicaciones de Superordenadores de EE.UU. (www.ncsa.uiuc.edu). A continuación, se explican los campos que puede tener un registro de acceso, tomando como ejemplo un servidor de la compañía “Terra”.

- Servidor: Por defecto será la dirección IP (200.38.26.54) del visitante. Para obtener el nombre de dominio (host.terra.es), hay que configurar el servidor para que haga una petición de resolución (traducción) de la dirección IP en el nombre de dominio de acuerdo con el DNS (Sistema de Nombre de Dominio). Normalmente, la respuesta se obtiene para 50-70% de las peticiones.

5.1.3 host.terra.es

-Identificación: "-" Casi nunca se utiliza. Originalmente, por el año 1984, fue pensado para autenticar a las máquinas UNIX conectadas a Internet mediante los antiguos protocolos FTP, Archie, Telnet, etc.

-Usuario Autorizado: "-" Otro campo pasado de moda. Originalmente fue pensado para contener el nombre autenticado del visitante para poder acceder a directorios protegidos con contraseña.

-Sello De Fecha: Fecha en formato dd/mmm/yyyy, hora en formato hh:mm:ss y las horas de diferencia del tiempo del servidor con el Meridiano de Greenwich. Por ejemplo, para Colombia 0500.

[01/Mar/2001:00:22:14 -0600]

-Petición: La petición HTTP consta de 3 partes. La primera es el método de la transacción: GET (petición normal), POST (envío de formulario) o HEAD (petición especial de un buscador). La segunda es la ruta del fichero solicitado (/EE/dic.mv). La tercera es el protocolo y la versión (habitualmente, HTTP/1.1)

"GET /EE/dic.mv HTTP/1.1"

-Código De Estado: Código de estado del servidor que describe lo que ha ocurrido con la transacción. El código más común es el 200 que significa que la página ha sido servida con éxito al solicitante. Algunos de los códigos más conocidos son: 404 Not Found - página no encontrada y 500 Internal Server Error - un fallo en el funcionamiento del servidor.

200

-Volumen Transferido: Este campo proporciona el tamaño de la transacción (fichero transferido) en bytes, en este caso 16k. Sólo las peticiones GET con el código de respuesta 200 tienen un valor en este campo. El resto de los casos se pone "0" o "-".

16059

Reunidos todos estos campos, se obtiene la línea típica del fichero log de accesos, por ejemplo, del servidor Apache (www.apache.org):

Un registro del fichero log de accesos del servidor Apache

5.1.1.54 - - [01/Mar/2001:00:22:14 -0600] "GET /EE/dic.mv HTTP/1.1" 200 16059

5.1.4 Técnica de Minería de datos (Algoritmo de Reglas de Asociación)

Las reglas de asociación nos permiten conocer la relación entre los diferentes atributos de la base de datos que estemos analizando, sea relacional, transaccional o *datawarehouse*. Las reglas obtenidas mediante estas técnicas expresan patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más atributos. [25].

La técnica aplicada para relacionar los atributos minados con reglas de asociación es la ARPH (Association Rule Hypergraph Partitioning), el cual es un algoritmo estándar basado en el algoritmo *a priori* propuesto en 1994 por Agrawal y Srikant con múltiples aplicaciones en el tratamiento de la información de diversas bases de datos.

Según el algoritmo *a priori*, se puede definir una regla de asociación como una implicación $X \rightarrow Y$, en la cual X , Y son conjuntos de datos, siendo X el conjunto de elementos antecedentes y Y el conjunto de elementos consecuente de la regla. Se pueden definir dos medidas de interés para la regla:

- *Support*: La cobertura expresa el porcentaje o fracción de registros del conjunto global de datos que satisfacen la unión de los elementos del antecedente y consecuente de la regla.
- *Confidence*: La confianza es una medida de la efectividad de la regla. Representa el porcentaje de casos en los que dado el antecedente se verifica la implicación. [26]

El programa seleccionado para la aplicación de la técnica es “WEKA (*Waikato Environment Knowledge Analysis*)”, es una aplicación desarrollada en JAVA por la Universidad de Waikato. Los datos deben ser leídos por la aplicación como texto plano en un tipo de archivo con la extensión “.arff (*Attribute-Relation File Format*)”.

Entre otros, este sistema de minería de datos provee el paquete “WEKA.associations.Apriori” que contiene la implementación del algoritmo de aprendizaje de reglas de asociación A priori. Podemos configurar este algoritmo con varias opciones: con la opción “UpperBoundMinSupport” indicamos el límite superior de cobertura requerido para aceptar un conjunto de ítems. Si no se encuentran conjuntos de

ítems suficientes para generar las reglas requeridas se va disminuyendo el límite hasta llegar al límite inferior (opción “LowerBoundMinSupport”). Con la opción “minMetric” indicamos la confianza mínima (u otras métricas dependiendo del criterio de ordenación) para mostrar una regla de asociación; y con la opción “numRules” indicamos el número de reglas que deseamos que aparezcan en pantalla. La ordenación de estas reglas en pantalla puede configurarse mediante la opción “MetricType”, algunas opciones que se pueden utilizar son: confianza de la regla, lift (confianza dividido por el número de ejemplos cubiertos por la parte derecha de la regla), y otras más elaboradas. [27].

5.2 Marco Conceptual

5.2.1 Extracción de Conocimiento (KDD):

“proceso complejo que permite la selección, limpieza, transformación y proyección de los datos; analizar los datos para extraer patrones y modelos adecuados; evaluar e interpretar los patrones para convertirlos en conocimiento”. [7, pp. 13, 14].

5.2.2 Minería de datos

La minería de datos o exploración de datos (es la etapa de análisis de "Knowledge Discovery in Databases" o KDD) es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. [25]

5.2.3 Registro de Acceso

Un registro de acceso es una lista de todas las solicitudes de archivos individuales que las personas han solicitado desde un sitio Web. Estos archivos incluirán los archivos HTML y

sus imágenes gráficas incrustadas y cualquier otro archivo asociado que se transmita. El registro de acceso (a veces denominado "datos sin procesar") puede ser analizado y resumido por otro programa. [28]

5.2.4 Web mining

En la gestión de relaciones con los clientes (CRM), la minería web es la integración de la información recopilada por metodologías y técnicas tradicionales de minería de datos con información recopilada en la World Wide Web. (Minería significa extraer algo útil o valioso de una sustancia más básica, como minería de oro de la tierra). La minería de la web se utiliza para comprender el comportamiento del cliente, evaluar la eficacia de un sitio web en particular y ayudar a cuantificar el éxito de una campaña de marketing. [29]

5.2.5 Dirección IP:

Las direcciones IP (IP es un acrónimo para Internet Protocol) son un número único e irrepetible con el cual se identifica una computadora conectada a una red que corre el protocolo IP. [30]

5.2.6 Atributo:

Los atributos son campos de datos que pertenecen a un archivo o fichero, pero no son parte de él. Por ejemplo, no cuentan en el tamaño del fichero y pueden ser copiados o modificados sin tocar el archivo en sí mismo. El sistema utiliza estos atributos para almacenar, por ejemplo, el tamaño, tipo de fichero, o fecha de la última modificación. Esto es similar a otros sistemas operativos y sus sistemas de archivos. [31]

5.2.7 Regla de asociación:

Las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos. [32].

5.3 Marco Legal

5.3.1 Artículo 61 Constitución Política de Colombia

En orden de la protección intelectual de la información en los trabajos de investigación y desarrollo, este proyecto se atiene a la normativa de protección contra el plagio. “El Estado protegerá la propiedad intelectual por el tiempo y mediante las formalidades que establezca la ley”. [14]

5.3.2 Artículo 56 del Reglamento estudiantil

El artículo 56 del Reglamento estudiantil de la Universidad Santo Tomás, en su párrafo señala: “para los trabajos escritos, los estudiantes deberán emplear debidamente las referencias bibliográficas y documentales...” [15].

Todas las fuentes de referenciación deben estar debidamente citadas y aprobadas. El desarrollo escrito de este proyecto acata las normas del AMERICAN PSYCHOLOGICAL ASSOCIATION (APA) como parámetros para la forma, redacción, citación y referenciación de este documento.

5.3.3 GPL (*General Public License*) 3.0 for Weka > 3.7.5

La licencia es desarrollada por el sistema operativo GNU. La Licencia Pública General de GNU es una licencia gratuita de copyleft para software y otros tipos de trabajos.

5.3.4 Permisos básicos.

Todos los derechos concedidos bajo esta Licencia son otorgados por el término de copyright del Programa, y son irrevocables siempre que se cumplan las condiciones establecidas. Esta Licencia afirma explícitamente su permiso ilimitado para ejecutar el Programa no modificado. La salida de la ejecución de una obra cubierta está cubierta por esta Licencia sólo si la salida, dado su contenido, constituye una obra cubierta. Esta Licencia reconoce sus derechos de uso justo u otro equivalente, según lo dispuesto por la ley de derechos de autor.

Usted puede hacer, ejecutar y propagar trabajos cubiertos que no transmita, sin condiciones mientras su licencia permanezca en vigor. Usted puede enviar trabajos cubiertos a otros con el único propósito de hacer que ellos hagan modificaciones exclusivamente para usted, o proveerle con facilidades para ejecutar esas obras, siempre y cuando usted cumpla con los términos de esta Licencia al transportar todo el material para el cual usted no controla derechos de autor. Aquellos que hagan o ejecuten las obras cubiertas para usted deben hacerlo exclusivamente en su nombre, bajo su dirección y control, en términos que les prohíban hacer copias de su material protegido por derechos de autor fuera de su relación con usted. [33]

5.4 Marco Histórico

5.4.1 Evolución de la Web

La Web inicial era eminentemente estática, es decir ficheros que estaban en un servidor de un sitio web. Hoy tenemos una Web completamente dinámica donde las páginas se generan dependiendo de las interacciones que realizan las personas en

todo el mundo, conectados en muchos casos a través de teléfonos móviles. Esta Web en la práctica es infinita, pues podemos generar un número arbitrario de páginas en un calendario o distintas representaciones de todas las bases de datos detrás de todos los sitios de comercio digital del mundo. Hoy hay más 1.500 millones de personas que usan internet, el 40% de ellas en Asia y con tasas de penetración en la población cercanas al 75% en América del Norte. Todas ellas la usan para enviar mensajes, buscar información o participar en redes sociales, entre millares de otras aplicaciones disponibles. [34].

La web ha evolucionado a un entorno dinámico en el que el flujo de la información es cada vez más complejo, rápido y diversificado, por lo tanto, resulta difícil tratar grandes volúmenes de información que provienen de diferentes servidores, para este propósito surgió el área del BIG DATA. Sin embargo, la minería de datos resulta útil aún para problemas más específicos que requieren de soluciones específicas y es trabajo de la minería de datos detectar los patrones de comportamiento de los datos que se almacenan como ficheros en un servidor. La minería de datos ha evolucionado a medida que la Web lo ha hecho. A pesar de tener limitaciones en cuanto al volumen de la información y sus aplicaciones, se han desarrollado nuevas técnicas que le han dado un espacio importante a la minería de datos en el contexto actual de la Web.

5.4.2 Evolución de la Minería de Datos

La idea de Minería de Datos no es nueva. Ya desde los años sesenta los estadísticos manejaban términos como Data Fishing, Data Mining (DM) o Data Archaeology con la idea de encontrar correlaciones sin una hipótesis previa en bases de datos con ruido. A

principios de los años ochenta, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro entre otros, empezaron a consolidar los términos de Minería de Datos y KDD. Esta tecnología ha sido un buen punto de encuentro entre personas pertenecientes al ámbito académico y al de los negocios. La evolución de sus herramientas en el transcurso del tiempo puede dividirse en cuatro etapas principales:

- Colección de Datos (1960).
- Acceso de Datos (1980).
- Almacén de Datos y Apoyo a las Decisiones (principios de la década de 1990).
- Minería de Datos Inteligente. (finales década de los 1990). [35]

Así pues, la minería de datos da sus inicios como una metodología para la recolección de información, en una época en la que el crecimiento tecnológico y el concepto de redes de información empiezan a tomar poder a nivel global. Más adelante se empieza a hablar acerca de la accesibilidad de la información como un facilitador, en cuanto la archivística y la desaparición del papel como fuente primaria de almacenamiento de grandes volúmenes de información. Todo esto impulsando al crecimiento de las industrias y el nacimiento de una ideología esbelta para la reducción de desperdicios y la optimización de los procesos. A principios de la década de los 90 se empieza a entender el gran valor de la información almacenada y el potencial de la información que se consideraba inútil como lo es la información de los registros de acceso y error de los servidores web. Información que luego sería aprovechada para la toma de decisiones pertinentes en las organizaciones. Es cuando nace entonces el concepto de minería de datos inteligente, en el que la extracción del conocimiento a partir de los patrones observados del análisis de un conjunto de datos, se

convierte en el campo de estudio más amplio para el tratamiento de la información y la toma de decisiones de mejores muchos más acertadas.

Debido al crecimiento acelerado de la Web, su globalización e infinita gama de aplicaciones y oportunidades, en los últimos años la minería de datos aplicada a la información recolectada de la Web ha tenido una relevancia mucho mayor a otros modelos de extracción de conocimiento. El Web Mining ha despertado gran interés en la actualidad, particularmente debido a los avances de la comunidad científica en distintas líneas de investigación relacionadas con Data Mining orientado a la World Wide Web. En los últimos años esto se ha potenciado fuertemente en virtud del gran aumento en volumen del tráfico, tamaño y complejidad de las fuentes de información disponibles en la Web y el reciente interés en el desarrollo del comercio electrónico (e-commerce). [36]

5.5 Estado del Arte

“Práctica de Minería de Datos, Introducción al Weka” desarrollado por José Hernández Orallo y Cesar Ferri Ramirez en la Universidad Politécnica de Valencia en el Año 2016. El informe ya es distribuido como una cartilla (libro pequeño) y surge como una herramienta para guiar al estudiante a familiarizarse con un entorno de minería de datos avanzado mediante el software libre conocidos como WEKA, ya que para la época resultó una herramienta muy útil para el tratamiento de la información, así que básicamente el propósito del documento es ofrecer una guía metodológica para la aplicación de diferentes técnicas de minería de datos en el entorno de WEKA. El documento presenta los resultados

de la aplicación de la metodología a partir de ejercicios prácticos pertinentes a cada técnica de minería de datos compatible con los algoritmos del software. Esta guía a pesar de no ser tan reciente, mantiene su vigencia y se puede utilizar en la actualidad con fines académicos. [27]

ESTUDIO WEBMINING Y DATAMINING DEL LOG HUAWEI LTDA, realizado por Javier Ulises Arias Rivera y Cesar E. Polo Castro en la Universidad de Magdalena en el año 2012. Este trabajo nace con la problemática del gran desaprovechamiento de los grandes volúmenes de datos de las organizaciones, ya que dicha información puede aprovecharse para hacer un análisis de minería de datos. Con el fin de aprovechar los conocimientos en Web mining, los autores decidieron realizar un modelo de aplicación para el análisis de un Log sucio de uno de los servidores de la empresa HUAWEI. Su objetivo era obtener información, a partir de un log generado por el servidor Web huawei.com, para que la toma de decisiones sea la correcta dentro de la organización. Como resultado de la aplicación de técnicas de agrupamiento y reglas de asociación a priori, lograron establecer las reglas con mayor cobertura y confianza, concluyendo así qué sitios eran los más visitados por los usuarios de la página web y sus tendencias, es decir en cómo usaban la página web. El trabajo ofrece información relevante en cuando a la metodología del modelo de web mining que aplicaré en este proyecto. [37].

El artículo científico publicado en la revista IEE “*Association Rule Mining for Web Usage Data to improve Websites*” realizado por Avadh Kishor, Ajeet Kumar, Ashish K. Maurya en la Universidad, UP, India en el año 2014. El artículo científico debe su realización a la creciente popularidad de la aplicación de técnicas de reglas de asociación en la minería de datos de uso de la web. El documento propone un modelo de reglas de asociación para

minar conocimiento con el uso de un algoritmo a priori (k-itemset). El proceso consiste en la extracción de reglas de asociación que pueden ser expresadas en información útil. Entonces, el conocimiento transaccional del uso de la web es considerado para el alcance de este propósito. Como resultado se obtuvo un sistema autónomo para la extracción de información a partir de reglas de asociación con un alto grado de confianza. [38]

6. Marco Metodológico

Este proyecto tiene un enfoque mixto para el estudio de patrones de comportamiento de los usuarios de la página web Americanmotos.com. La información será extraída de un web log, es decir que la escritura de la información consiste en una línea de código para cada entrada o fichero, por lo tanto, la información se considerará con atributos. Al realizar la limpieza y procesamiento para su análisis, transformaremos los atributos en valores cuantitativos que nos permitirán un análisis numérico de la información.

6.1 Fundamentos Epistemológicos.

La investigación y aplicación del proyecto tiene un enfoque mixto, ya la información extraída se considera como atributos, pero será procesada con valores numéricos que caracterizan cada atributo. Se considerará el tratamiento de un web Access log con más de tres mil ficheros (líneas de código). Esta información es referente a las peticiones del usuario, la IP con la que realiza la consulta, la fecha y la hora y además la ruta de acceso o URL.

6.2 Diseño de la investigación

Este proyecto inicia como una investigación exploratoria, de recopilación de conceptos que inciden en la aplicación de un modelo de minería de datos para el uso de la web. Con ayuda de los

conceptos y guías metodológicas se procederá una etapa de investigación explicativa, ya que al realizarlos experimentos para el tratamiento de la información, se buscarán respuestas para la solución de los problemas que se presentan durante la realización del proyecto en el ámbito teórico y metodológico, finalmente la investigación terminará como una investigación descriptiva, ya que ofrecerá conocimiento a partir de los resultados de la aplicación del modelo de minería. Este conocimiento se traduce a partir de patrones de comportamiento, correlaciones y tendencias descritas por las reglas de asociación encontradas.

6.3 Fases o etapas del proyecto.

6.3.1 Selección del log

Se enlistaran una serie de log gratuitos, que nos permitirán obtener una cantidad de registros sobre las acciones que realizan los usuarios de la página web, para determinar con dichos datos las tendencias que permitan la toma de decisiones y aumento de ventas en la compañía a estudiar.

Con este objeto se recurrió a la búsqueda de compañías que suministraran el ingreso al log de sus páginas web, entre las que se encontraban:

- Servicios integrales de tecnología.
- Garcillantas boulevard
- Comercializadora Toque de ángel
- American Motos

En el proceso de selección del log se tomó en cuenta la disponibilidad de la compañía, diversidad de portafolio, expansión en el mercado, tradición en el sector y expansión de canales de distribución. Para cumplir con el objetivo elegimos la compañía American motos, la cual fue fundada en el año 1994 por el señor Carlos Díaz, se encuentra ubicada en la Carrera 22 # 28-43

Bucaramanga. También distribuye más de 20 marcas de llantas con un catálogo de cerca de 1000 modelos de las mismas.

Actualmente comercializa llantas, cascos, lubricantes y repuestos para motos, para brindar a sus clientes el mayor respaldo y servicio en el momento de realizar sus compras.



FIGURA 1. AMERICAN MOTOS

6.3.2 Configuración software Analog

Se realizó la descarga del software libre Analog con el fin de determinar el tráfico y cantidad de peticiones de la página web Americanmotos.com.

Para tal propósito se realizó una previa configuración de parámetros del software para obtener posteriores resultados:

LOGFILE access-log

Debido a que este parámetro inicia con # es considerado por lo tanto analog no lo determina la momento de ejecutarse.

LOGFILE logs/*

Con este parámetro pedimos que sean seleccionados todos los registros del log del servidor.

LANGUAGE spanish

Nos indica que todos los reportes sean visualizados en idioma español

OUTFILE reporte.html

Declaramos el nombre del archivo donde serán generados los reportes del software, para este caso en la extensión HTML.

HOSTNAME "American Motos"

Este parámetro visualiza el encabezado de los resultados en este caso American Motos.

LOGO <https://www.americanmotos.com/gif>

En esta línea podemos visualizar una imagen para el momento de arrojar los resultados de Analog

LINKINCLUDE pages

El parámetro indica que los resultados obtenidos tengan links HTML.

BASEURL <http://www.americanmotos.com>

El parámetro indica en que URL serán obtenidos los datos a analizar.

FAILURE ON

Este parámetro nos indica que serán generados los reportes tanto positivos como negativos de la página web American Motos.

Con la configuración de los parámetros de Analog ya queda disponible la opción de ejecución del log para posteriores resultados.

6.3.3 Organización y limpieza del log.

El procedimiento de limpieza del log es clave dentro de la depuración de datos, cuya organización y manejo es crítico, impactando directamente los resultados obtenidos, es por esta razón que se decidió utilizar una herramienta que permitiera maniobrar dentro del mar de información y cuya interfaz es más amigable con el usuario, derivando en la implementación del programa Microsoft Excel, debido a que los filtros y la función adjudicada a “Texto en columnas”, permitieron manejar de forma independiente los diferentes atributos para alcanzar con propiedad, los valores adjudicados a cada campo, teniendo en cuenta el procedimiento previo realizado con ANALOG donde se logró analizar los datos del servidor web adjudicado a American Motos S.A

En primera medida se estudió el log sucio del American Motos, para su previo análisis. En la siguiente ilustración se podrá observar el log original:



The screenshot displays the Microsoft Excel application window. The title bar reads "log american motos - Excel". The ribbon is set to the "Inicio" (Home) tab, showing various options for font, alignment, and styles. The status bar at the bottom indicates "Hoja1" and "Inicio".

	A	B	C	D	E	F	G	H	I	J
	(IP del cliente)	[dia]	[hora]	código de estado	tamaño del recurso	cabecera de petición del Referer	cabecera de petición para el agente de usuario			
1	157.55.83.125	1	-0500	404	208		Mozilla/5.0 (compatible; bingbot/2.0; http://www.bing.com/)			
2	157.55.83.125	1	-0500	200	638		Mozilla/5.0 (compatible; bingbot/2.0; http://www.bing.com/)			
3	202.40.55.59	1	-0500	200	1125		Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0)			
4	119.63.193.32	1	-0500	200	1125		Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0)			
5	186.1.176.12	1	-0500	200	653	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
6	186.1.176.12	1	-0500	304	-	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
7	186.1.176.12	1	-0500	304	-	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
8	186.1.176.12	1	-0500	200	245388	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
9	186.1.176.12	1	-0500	200	245388	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
10	186.1.176.12	1	-0500	200	1986284	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
11	186.1.176.12	1	-0500	200	245388	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
12	186.1.176.12	1	-0500	200	245388	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
13	186.1.176.12	1	-0500	200	245388	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
14	186.1.176.12	1	-0500	200	245388	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
15	186.1.176.12	1	-0500	200	245388	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
16	186.1.176.12	1	-0500	200	29	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
17	186.1.176.12	1	-0500	200	29	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			
18	186.1.176.12	1	-0500	200	29	https://www.americanmotos.com/product-186-1-176-12	Mozilla/5.0 (Windows NT 5.1; rv23.0) Gecko/20100101 Firefox/23.0			

FIGURA 3. TABULACIÓN DEL LOG. ELABORACIÓN PROPIA

6.3.4. Configuración de ejecutable en WEKA

Para la posterior ejecución del software WEKA y con ayuda de los registros tabulados en Microsoft Excel se procedió a crear el archivo (arff) extensión utilizada para documentos en weka, en los cuales se definirán las variables a utilizar con el parámetro @attribute para la compañía American Motos se definieron los siguientes:

@attribute día: Indica los días en los que se realizaron los registros.

@attribute Mes: Registra los meses en que se realizaron las visitas a la página web.

@attribute año: Registra los datos del año.

@attribute Hora: Indica la hora de consecución de las visitas.

@attribute minuto: Este parámetro registra el minuto de ingreso al registro.

@attribute segundo: Muestra los datos en cada segundo ingresado.

@attribute código: indica el código de verificación del registro, ya sea positivo, negativo o en proceso de ejecución.

@attribute url: Registra las url visitadas por los usuarios de la página web.

american motos 2018 Bloc de notas

Archivo Edición Formato Ver Ayuda

Correlación MOTOS

```
@attribute IP{"157.55.33.125","202.46.55.59","119.63.193.132","186.82.19.218","190.90.196.164","66.249.73.11","100.43.83.134","157.55.35.106","66.249.75.133","181.51.91.60.180","66.249.66.133","208.87.233.180","186.1.182.30","190.90.87.162","202.46.53.36","181.55.114.182","199.30.24.150","181.241.29.125","199.30.25.89","186.1.177.139","181.52.144.192","181.135.110.56","180.76.5.98","190.84.56.12","190.90.122.6","203.24.148.50","186.1.171.67","190.255.82.69","190.158.175.187","190.66.225.68","202.46.60.17","202.46.54.44","190.157.219.109","186.82.163.196","202.46.54.35","180.76.5.147","202.46.60.17","157.55.35.46","201.245.190.2","202.46.60.19","157.55.33.25","181.129.131.87"}
@attribute dia {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31}
@attribute Mes {Noviembre,Diciembre}
@attribute Año {2017}
@attribute Horas {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23}
@attribute Minutos {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52}
@attribute Segundos {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52}
@attribute Código {200,304,404}
@attribute URL{"https://www.americarmotos.com/product-page/h660","https://www.americarmotos.com/llantas","https://www.americarmotos.com/product-page/m7","https://www.americarmotos.com/product-page/si-7-syn-2t-ing-oil","https://www.americarmotos.com/product-page/sport-rain","https://www.americarmotos.com/product-page/t30","https://www.americarmotos.com/product-page/sh-mx30","https://www.americarmotos.com/product-page/bandas-de-freno-japan","https://www.americarmotos.com/product-page/power-pure-2ct","https://www.americarmotos.com/product-page/s-23-freak","https://www.americarmotos.com/product-page/city-grip","https://www.americarmotos.com/product-page/celerity-4t-20v50-titanio","https://www.americarmotos.com/product-page/s-06-one","https://www.americarmotos.com/product-page/revenge-speed-7","https://www.americarmotos.com/product-page/motul-7100-4t","https://www.americarmotos.com/product-page/h660","https://www.americarmotos.com/blog","https://www.americarmotos.com/product-page/battlax-s21"}
@data
```

181.129.131.87	1	Diciembre	2017	16	2	0	200	https://www.americarmotos.com/product-page/h660
181.129.131.87	1	Diciembre	2017	16	2	0	200	https://www.americarmotos.com/product-page/h660
181.129.131.87	1	Diciembre	2017	16	2	0	200	https://www.americarmotos.com/product-page/si-7-syn-2t-ing-oil
181.129.131.87	1	Diciembre	2017	16	6	0	200	https://www.americarmotos.com/product-page/m5
181.129.131.87	1	Diciembre	2017	16	7	0	200	https://www.americarmotos.com/product-page/me-77
181.129.131.87	1	Diciembre	2017	16	8	0	200	https://www.americarmotos.com/product-page/k761
181.129.131.87	1	Diciembre	2017	16	8	0	200	https://www.americarmotos.com/product-page/m5
181.135.177.7	1	Diciembre	2017	6	51	0	200	https://www.americarmotos.com/product-page/trailmax
181.135.177.7	1	Diciembre	2017	6	51	0	200	https://www.americarmotos.com/product-page/revenge-speed-7
181.135.177.7	1	Diciembre	2017	6	51	0	200	https://www.americarmotos.com/product-page/revenge-speed-7
181.135.177.7	1	Diciembre	2017	6	52	0	200	https://www.americarmotos.com/product-page/motul-7100-4t
181.135.177.7	1	Diciembre	2017	6	52	0	200	https://www.americarmotos.com/product-page/motul-7100-4t
181.135.177.7	1	Diciembre	2017	6	52	0	200	https://www.americarmotos.com/product-page/h660
181.135.177.7	1	Diciembre	2017	6	52	0	200	https://www.americarmotos.com/product-page/h660
181.135.177.7	1	Diciembre	2017	6	52	0	200	https://www.americarmotos.com/blog
181.135.177.7	1	Diciembre	2017	6	52	0	200	https://www.americarmotos.com/blog
181.135.177.7	1	Diciembre	2017	6	52	0	200	https://www.americarmotos.com/product-page/battlax-s21
181.135.177.7	1	Diciembre	2017	6	52	0	200	https://www.americarmotos.com/product-page/battlax-s21

FIGURA 4. PARÁMETROS WEKA. ELABORACIÓN PROPIA.

6.3.5. Análisis en ANALOG y WEKA

Extraeremos conocimiento útil suministrado por los software Analog y WEKA para posteriormente y con ayuda de las reglas de asociación dar resultados del proyecto como se evidenciara a continuación:

6.4 Estadísticas del servidor web de American Motos

Las peticiones en cuestión fueron realizadas del 15 de noviembre del 2017 4:32 am al 15 de diciembre del 2017 4:42 pm (30 días).

6.4.1 Resumen General

- Peticiones exitosas: 5,848
- Promedio de peticiones exitosas por día: 191
- Peticiones exitosas por página: 621
- Promedio de peticiones exitosas por página por día: 20
- Peticiones no logradas: 112
- Archivos diferentes solicitados: 4,639
- Sistemas diferentes atendidos: 261
- Transferencia total: 1.63 gigabytes
- Promedio de transferencia por día: 54.73 megabytes

6.4.2 Informe Mensual

Con ayuda del software libre analog, se presenta un informe de las peticiones realizados en el tiempo de estudio del proyecto.

Cada unidad (■) representa 10 peticiones por páginas o fracción.

mes	No. pet.	Págs.	
Nov 2017	3343	310	■
Dic 2017	2505	311	■

FIGURA 5. INFORME MENSUAL ANALOG. ELABORACIÓN PROPIA

En este informe se logra observar el mes de mayor tráfico para la página, que fue el mes de Diciembre con un 1,003% más que el mes anterior lo cual no muestra diferencia significativa en el momento de estudio, así como también el mes de mayor número de peticiones por parte de los usuarios que fue el mes de Noviembre con un 33.45% más referente a Diciembre.

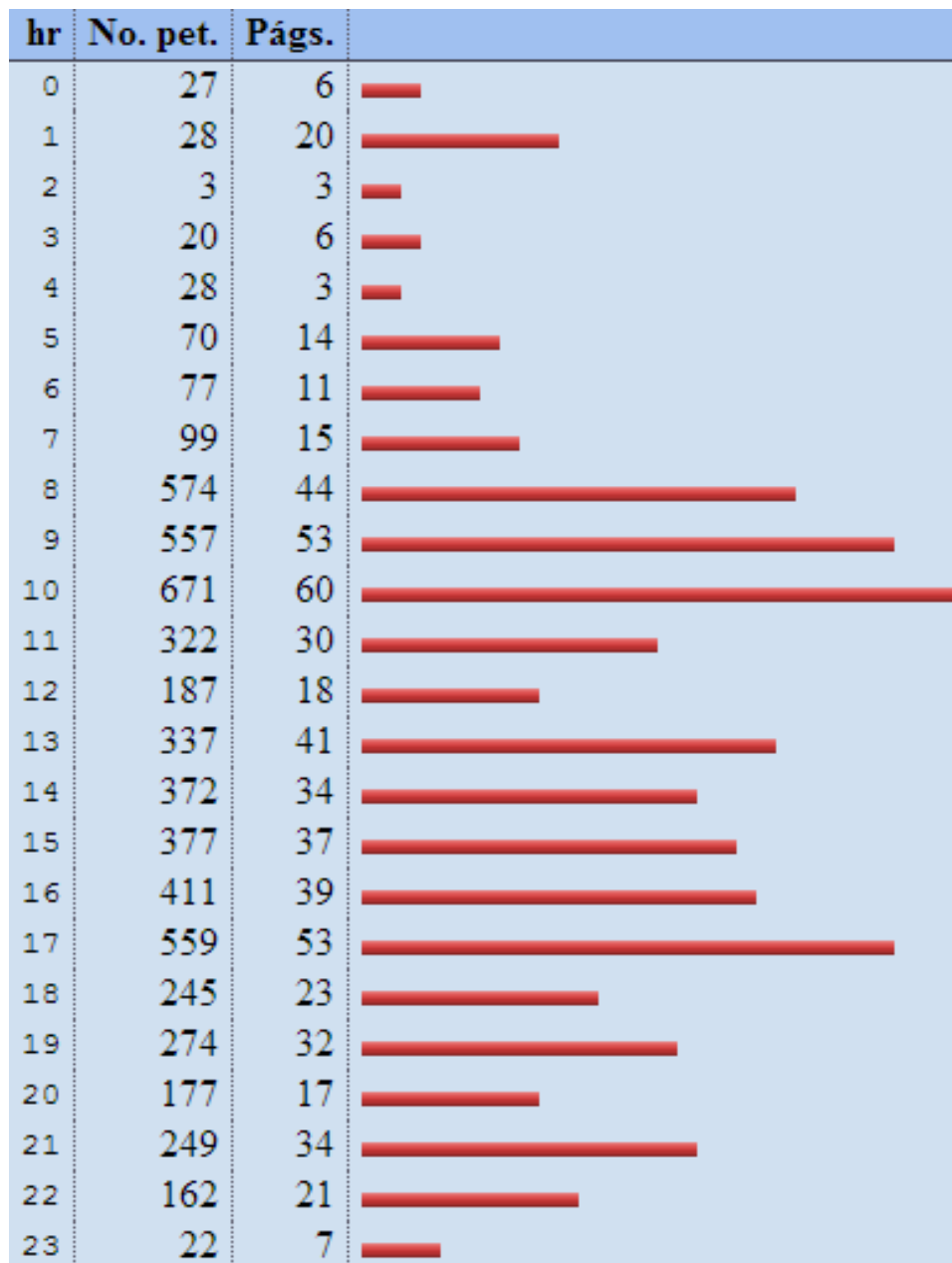


FIGURA 7. INFORME HORARIO. ELABORACIÓN PROPIA

De este reporte se puede concluir que la hora donde se presentaron más ingresos a la página es a las 10:00 am con 671 números de peticiones por 60 páginas con un incremento del 275.45% respecto al promedio horario que es de 243.6, lo que podemos deducir es que las horas laborales son las horas con mayor ingreso al sitio.

6.4.5 Informe de códigos de estado

Este reporte enlista los códigos de estado HTTP de todas las peticiones.



Las áreas están dibujadas por número de peticiones.

FIGURA 8. CÓDIGOS DE ESTADO. ELABORACIÓN PROPIA

En la figura anterior se enlistan los tres códigos de estado, destacándose el código 200 que significa petición exitosa, con un 94.24% de los registros totales, en menor medida sigue el código 304 (sin modificar desde el ultimo acceso) con un 3,87% y terminando con el código 404 (documento no encontrado) con un 1,88% de las peticiones. Lo que nos indica la estabilidad por parte del servidor para American Motos.

Listado de los códigos de estado, por orden numérico.

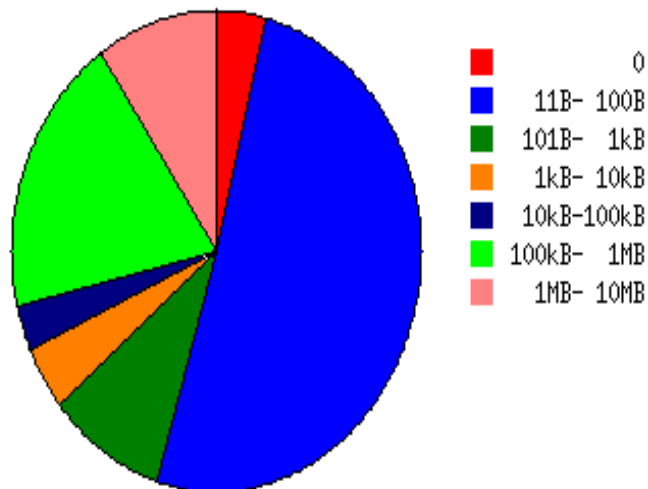
No. pet.	cód. de estado
5617	200 OK
231	304 Sin modificar desde el último acceso
112	404 Documento no encontrado

FIGURA 9. PETICIONES DE ESTADO. ELABORACIÓN PROPIA

Se puede observar de manera numérica que los aciertos obtenidos en la página fueron de 5617 entre los demás observamos que el servidor tiene una gran eficiencia con respecto a los contenidos fallidos.

6.4.6 Informe del tamaño de los archivos

Este reporte enlista los tamaños de los archivos.



Las áreas están dibujadas por número de peticiones.

FIGURA 10. GRÁFICO DE BITES. ELABORACIÓN PROPIA

Listado de los tamaños por bites, referente al número de peticiones de los usuarios de American Motos. Lo que evidencia que la tendencia de bites descargada por los usuarios de de entre 11-100Kb con un 52% de todas sus peticiones.

tamaño	No. pet.	%bytes
0	234	
1B- 10B	0	
11B- 100B	2955	0.01%
101B- 1kB	543	0.02%
1kB- 10kB	256	0.05%
10kB-100kB	177	0.83%
100kB- 1MB	1128	17.67%
1MB- 10MB	555	81.42%

FIGURA 11. TAMAÑO DE DESCARGA. ELABORACIÓN PROPIA

En la figura anterior podemos observar, el tamaño de los bites descargados por parte de los usuarios de la página web, donde se destacan las descargas entre 1-10 Mega bites con 555 peticiones con el 81.42% de la descarga total de los usuarios.

También se aprecia que el tamaño de descarga más atractivo para los usuarios es el comprendido entre 11-100 bites con 2955 peticiones, aunque no representa un margen alto en la cantidad de descarga total con tan solo 0,01%.

En segunda medida resaltan las descargas entre 100Kb- 1Mb con 1128 peticiones que representan el 17.67% de las descargas totales de los usuarios de American Motos.

6.4.7 Informe de tipos de archivo

Este reporte enlista las extensiones de los archivos pedidos.



Las áreas están dibujadas por cantidad de tráfico.

FIGURA 12.EXTENSIONES USADAS. ELABORACIÓN PROPIA

Listado de las extensiones que representan, por lo menos, 0.1% del tráfico, ordenadas por cantidad de tráfico.

No. pet.	%bytes	extensión
1684	99.09%	.jpg [JPEG graphics]
292	0.83%	.swf
3872	0.07%	[no listadas: 3 extensiones]

FIGURA 13. PETICIÓN DE EXTENSIÓN. ELABORACIÓN PROPIA

Es claro que la extensión con mayor número de peticiones fue el .jpg con un 99.09% ya que, si visitamos la web, la mayor parte de los enlaces e imágenes están en esa extensión, algunos enlaces se encuentran en extensión .swf y por eso resalta en los resultados con el 0,83% del total de las extensiones.

6.5 ANALISIS DE LOS DATOS MEDIANTE WEKA

Luego de la limpieza del log mediante Microsoft Excel, se prosiguió con la creación del archivo .arff, el cual posee los campos de estudio: IP, fecha, código de estado y URL, para que el programa WEKA lo reconozca. A continuación, se mostrará la creación del .arff

```

american motos: Bloc de notas
Archivo Edición Formato Ver Ayuda
relation MOTOS
@attribute IP{"157.55.33.125","202.46.55.59","119.63.193.132","186.82.19.218","190.90.196.164","66.249.73.11","100.43.83.134","157.55.35.105","66.249.75.133","181.51.91.60.180","66.249.66.133","208.87.233.180","186.1.182.30","190.90.87.162","202.46.53.36","181.55.114.182","199.30.24.150","181.241.29.125","199.30.25.89","186.1.177.139","181.52.144.192","181.135.110.56","180.76.5.98","190.84.56.12","190.90.122.6","203.24.148.50","186.1.171.67","190.255.82.69","190.158.175.187","190.66.225.68","202.46.63.92.118","202.46.54.44","190.157.219.109","186.82.163.196","202.46.54.35","180.76.5.147","202.46.60.17","157.55.35.46","201.245.190.2","202.46.60.19","157.55.33.25","181.51.91.60.180"}
@attribute dia {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31}
@attribute Mes {Noviembre,Diciembre}
@attribute Año {2017}
@attribute Horas {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23}
@attribute Minutos {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52}
@attribute Segundos {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52}
@attribute Codigo {200,304,404}
@attribute URL{"https://www.americanmotos.com/product-page/h660","https://www.americanmotos.com/llantas","https://www.americanmotos.com/product-page/m7","https://www.americanmotos.com/product-page/si-7-syn-2t-ing-oil","https://www.americanmotos.com/product-page/sport-rain","https://www.americanmotos.com/product-page/t30","https://www.americanmotos.com/product-page/sh-mx30","https://www.americanmotos.com/product-page/bandas-de-freno-japan","https://www.americanmotos.com/product-page/power-pure-2ct","https://www.americanmotos.com/product-page/s-23-freak","https://www.americanmotos.com/product-page/city-grip","https://www.americanmotos.com/product-page/celerity-4t-20w50-titanio","https://www.americanmotos.com/product-page/s-06-one","https://www.americanmotos.com/product-page/battlax-s21"}
@data
181.129.131.87 1 Diciembre 2017 16 2 0 200 https://www.americanmotos.com/product-page/h660
181.129.131.87 1 Diciembre 2017 16 2 0 200 https://www.americanmotos.com/product-page/h660
181.129.131.87 1 Diciembre 2017 16 2 0 200 https://www.americanmotos.com/product-page/si-7-syn-2t-ing-oil
181.129.131.87 1 Diciembre 2017 16 6 0 200 https://www.americanmotos.com/product-page/m5
181.129.131.87 1 Diciembre 2017 16 7 0 200 https://www.americanmotos.com/product-page/me-77
181.129.131.87 1 Diciembre 2017 16 8 0 200 https://www.americanmotos.com/product-page/k761
181.129.131.87 1 Diciembre 2017 16 8 0 200 https://www.americanmotos.com/product-page/m5
181.135.177.7 1 Diciembre 2017 6 51 0 200 https://www.americanmotos.com/product-page/trailmax
181.135.177.7 1 Diciembre 2017 6 51 0 200 https://www.americanmotos.com/product-page/revenge-speed-7
181.135.177.7 1 Diciembre 2017 6 51 0 200 https://www.americanmotos.com/product-page/revenge-speed-7
181.135.177.7 1 Diciembre 2017 6 52 0 200 https://www.americanmotos.com/product-page/motul-7100-4t
181.135.177.7 1 Diciembre 2017 6 52 0 200 https://www.americanmotos.com/product-page/motul-7100-4t
181.135.177.7 1 Diciembre 2017 6 52 0 200 https://www.americanmotos.com/product-page/h660
181.135.177.7 1 Diciembre 2017 6 52 0 200 https://www.americanmotos.com/product-page/h660
181.135.177.7 1 Diciembre 2017 6 52 0 200 https://www.americanmotos.com/blog
181.135.177.7 1 Diciembre 2017 6 52 0 200 https://www.americanmotos.com/blog
181.135.177.7 1 Diciembre 2017 6 52 0 200 https://www.americanmotos.com/product-page/battlax-s21
181.135.177.7 1 Diciembre 2017 6 52 0 200 https://www.americanmotos.com/product-page/battlax-s21

```

FIGURA 14. ARCHIVO ARFF. ELABORACIÓN PROPIA

En las primeras líneas se puede observar, la declaración de las variables que utilizaremos en el archivo, entre las llaves van los valores que tomara cada registro del archivo. En este formato, que es estándar de WEKA, podemos observar los registros, que se listan después de @data. Dichos registros fueron tomados del archivo tabulado previamente en Excel para mayor maniobrabilidad de los datos obtenidos.

6.5.1. Frecuencia de los datos

Mediante el software asociado a la actividad de minería de datos WEKA, se realizó un análisis más detallado al log obtenido, logrando así obtener datos concluyentes arrojados de la página de American Motos S.A, reflejando gran afluencia, tanto de clientes como de personas particulares. Es por ello que decidimos analizar la frecuencia de los datos, para saber, cuáles son las IP que más solicitudes hacen al servidor, y desde que recurso acceden a ella.

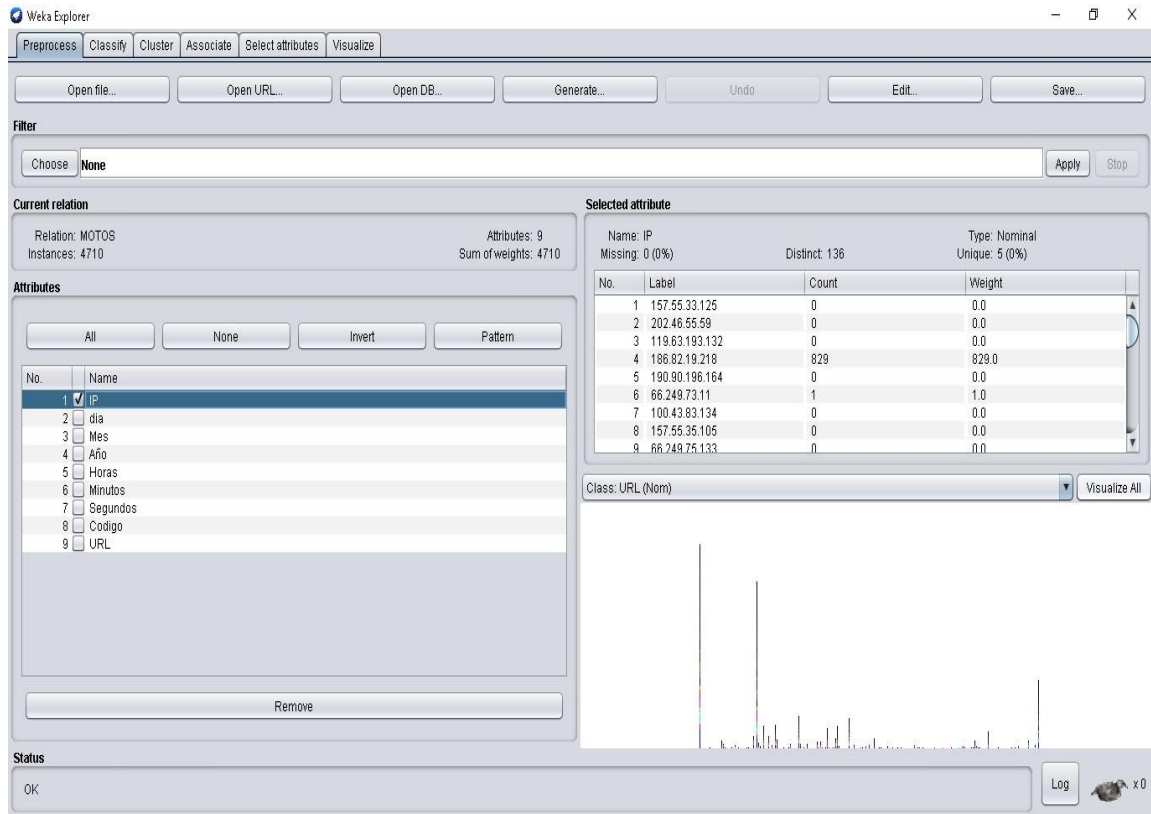


FIGURA 15. PANTALLA INICIO WEKA. ELABORACIÓN PROPIA

En esta vista se observa al lado izquierdo las variables a estudiar como son: IP, día, mes, año, horas, minutos, segundos, código y URL. Al lado inferior derecho se observa un gráfico con respecto a la variable utilizada para el estudio y la cantidad de veces en que se repite dicho registro.

Analizando la gráfica anterior, se puede inferir que las IP con más solicitudes al servidor fueron:

FRECUENCIA	
IP	FRECUENCIA
186.82.19.218	829
190,90,196,164	669
190.205.244.25	285
190,90,124,222	154
186,179,106,114	123
186,1,176,66	120
190,90,42,236	120
190.214.49.4	112

FIGURA 16. FRECUENCIAS IP. ELABORACIÓN PROPIA

Realizando una revisión a todas las IP, que hacen consultas al servidor, se encontraron que una gran parte de estas IP, no pertenecen a la empresa. Mediante la herramienta de software libre DNSStuff.com pudimos saber de qué operador y de qué país provienen algunas de estas direcciones, encontrando así, solicitudes desde países como Venezuela y Ecuador. Además, proveedores como Colombia Móvil, Telefónica, Dynamic Broadband Service y ETB.

Mostramos algunas evidencias a continuación:

WHOIS/IPWHOIS Lookup Results for 186.82.19.218
Share

Results for Target: 186.82.19.218

Created Date : 20081028
Updated Date : 20100305
WHOIS Server: whois.arin.net

Discovered Nameservers
Registrar Information

NS3.TELMEXLA.NET.CO | 190.157.8.71
NS2.TELMEXLA.NET.CO | 190.157.8.71
[NOT DETECTED]

*Please note these results are obtained from third party databases (whois.arin.net)

Contact Information

Registrant
Administrative Contact
Technical Contact

11111 - Bogota - DC
CO
Cra 7 # 63-44 Piso 6, 00, 00
10 - Bogota - DC
Network Security Team
ATI

FIGURA 17. IP MÁS FRECUENTE. ELABORACIÓN PROPIA

En la imagen anterior se muestra la IP que más se dirigió a la página de la compañía American Motos

WHOIS/IPWHOIS Lookup Results for 190.205.244.25 [Share](#)

Results for Target: 190.205.244.25

Created Date :	20070613
Updated Date :	20070613
WHOIS Server:	whois.arin.net

Discovered Nameservers DNS1.CANTV.NET 200.44.32.10 DNS2.CANTV.NET 200.109.76.11	Registrar Information [NOT DETECTED]
--	--

*Please note these results are obtained from third party databases (whois.arin.net)

Contact Information

Registrant	Administrative Contact	Technical Contact
Segunda Avenida de los Palos Grandes, 000, Entre Av. Fr 1060 - Caracas - MI VE CANTV COR Los Palos Grandes- Chacao, Caracas Venezuela, 000, -	Alexander Martinez	LUM

[▶ Detailed WHOIS Response](#)

FIGURA 18. IP DE MAYORES VISITAS EXTRANJERA. ELABORACIÓN PROPIA

En la ilustración se observa que la IP es de origen Venezolano y en la clasificación de frecuencias, ocupa la tercera posición con 285 interacciones.

[◀ Return to tools](#)

WHOIS/IPWHOIS Lookup Results for 190.214.49.4 [Share](#)

Results for Target: 190.214.49.4

Created Date :	20130822
Updated Date :	20130822
WHOIS Server:	whois.arin.net

Discovered Nameservers [NOT DETECTED]	Registrar Information [NOT DETECTED]
---	--

*Please note these results are obtained from third party databases (whois.arin.net)

Contact Information

Registrant	Administrative Contact	Technical Contact
CHILE OY Y CHIRIBOGA C.C. BAHIA MALL LOCAL 84, , - GUAYAQUIL - EC EC Edificio Droira, s/n, esquina	Evelin Gavilanes	VWR

[▶ Detailed WHOIS Response](#)

FIGURA 19. IP ECUATORIANA. ELABORACIÓN PROPIA

Se visualiza que la IP es de origen ecuatoriano, y se encuentra ubicado en la casilla 10 por orden de frecuencias en la página de American Motos.

6.6 Datos generados por WEKA

La ventana que nos muestra el programa, después de la ejecución del archivo es la que mostramos a continuación:

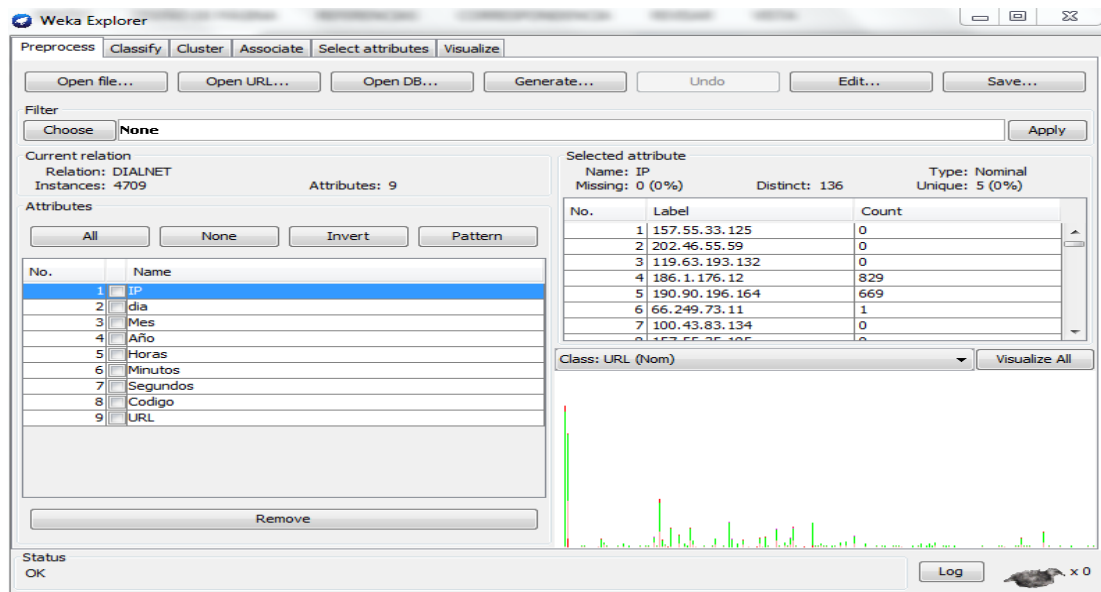


FIGURA 20. WEKA. ELABORACIÓN PROPIA

7. Resultados

Este proyecto inicia como una investigación exploratoria, de recopilación de conceptos que inciden en la aplicación de un modelo de minería de datos para el uso de la web. Con ayuda de los conceptos y guías metodológicas se procederá una etapa de investigación explicativa, ya que, al realizarlos experimentos para el tratamiento de la información, se buscarán respuestas para la solución de los problemas que se presentan durante la realización del proyecto en el ámbito teórico y por medio de las reglas de asociación se darán los siguientes resultados.

7.1 Graficas generadas por WEKA

WEKA nos permite obtener una gráfica por cada atributo declarado en el archivo:

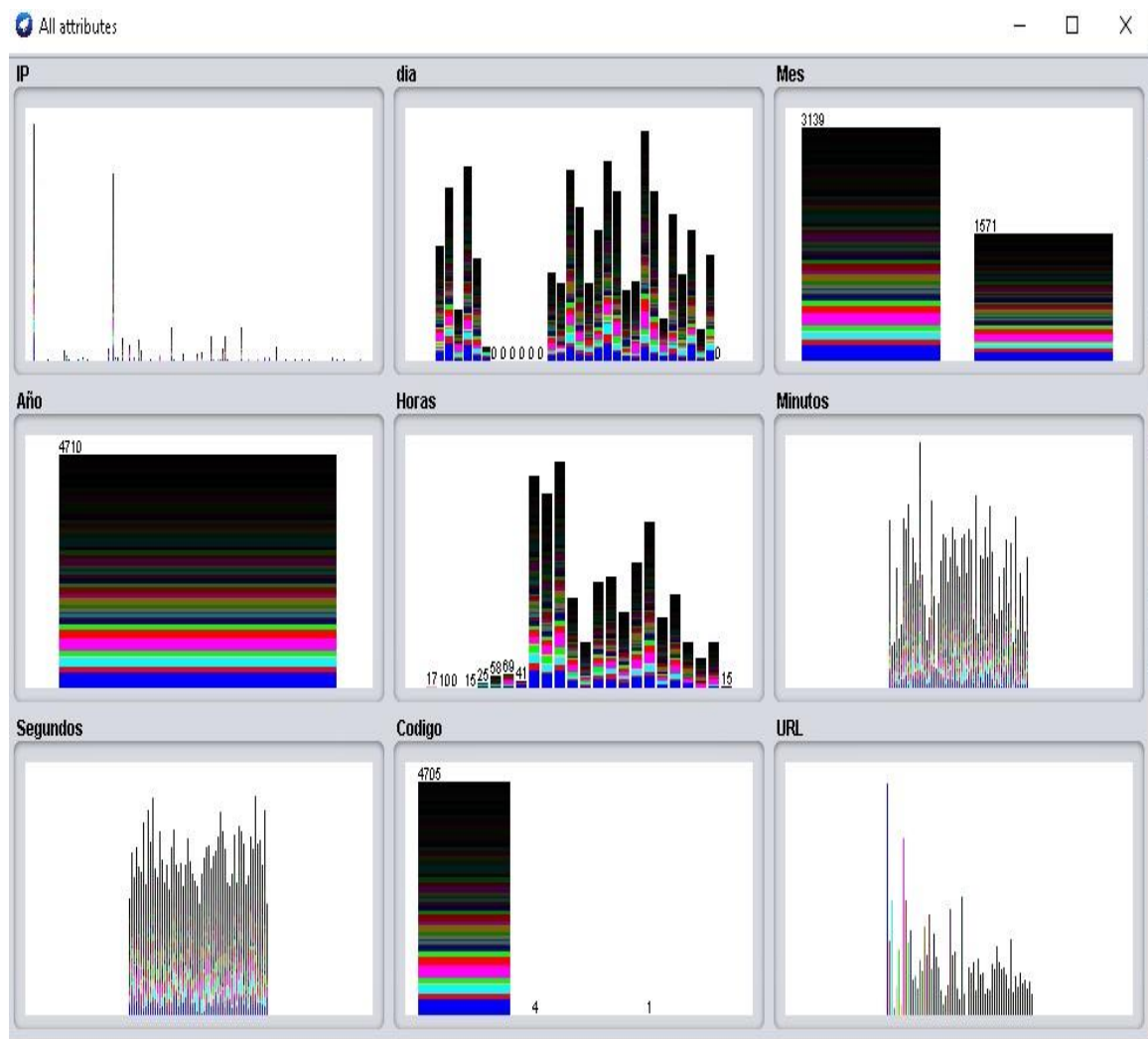


FIGURA 21. GRAFICAS WEKA. ELABORACIÓN PROPIA

En la figura anterior se observan distintos gráficos, dependiendo del tipo de variable estudiada. Las barras indican la variable estudiada, mientras que los colores indican la IP usada en cada variable para dar un mejor entendimiento de las situaciones presentadas.

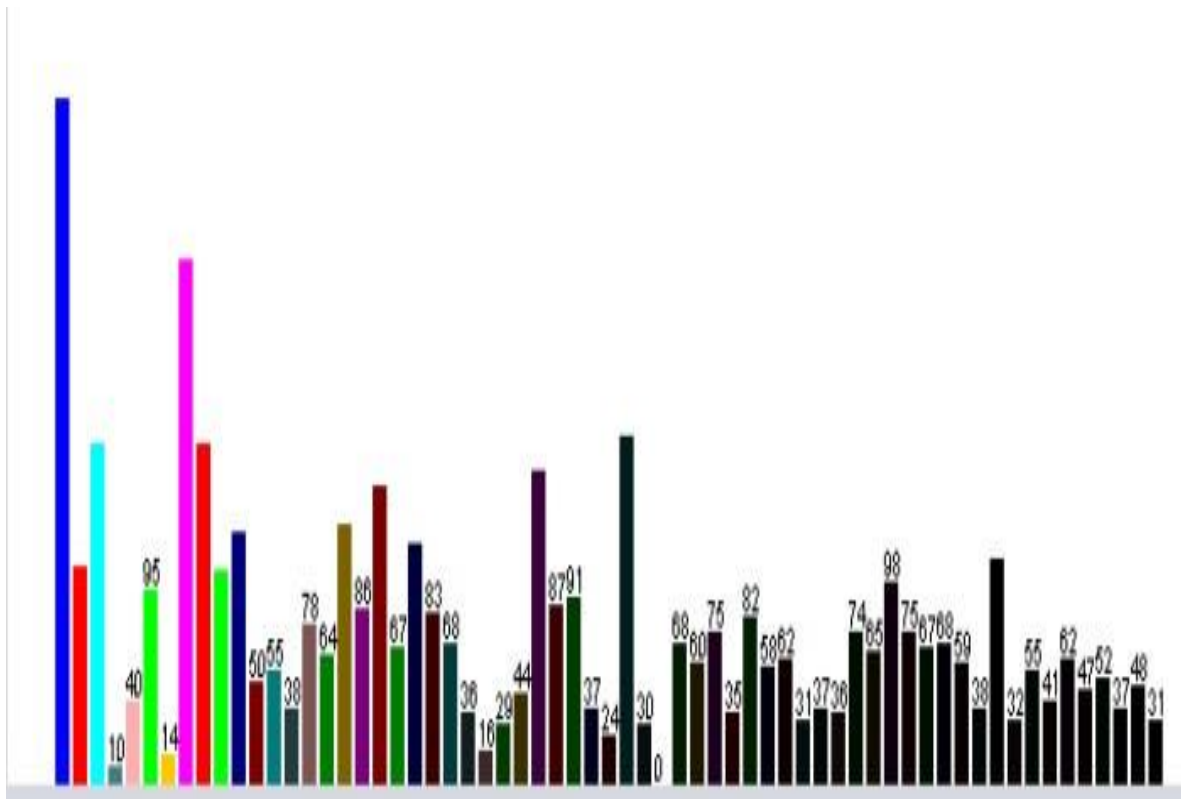


FIGURA 22. URL MÁS VISITADAS. ELABORACIÓN PROPIA

Mediante este grafico observamos la frecuencia de las peticiones de los usuarios en la página web American Motos. En ella cada columna representa una URL de la página y el número representa la cantidad de visitas a cada URL, los colores más llamativos representan las IP más representativas en dichas visitas.

7.1.1 Productos más visitados por los usuarios de la página web American Motos

Mediante la frecuencia de las URL se identifican los productos más atractivos para los clientes así como su tendencia a la compra de productos relacionados.

<https://www.americanmotos.com/product-page/h660>, frecuencia 333 visitas

H-660



FIGURA 23. LLANTA MÁS VISITADA

Llanta doble propósito para motos enduro, proporciona mejor agarré en situaciones de humedad, es uno de los productos de mayor rotación de American Motos en el año 2017.

<https://www.americanmotos.com/product-page/h-923> , 255 visitas

H-923



FIGURA 24.LLANTA H923

Llanta para terreno urbano, flexible que proporciona mejor agarre en las curvas. Es un producto nuevo de la compañía y es comercializado por American Motos desde septiembre del 2017.

<https://www.americanmotos.com/product-page/z8> , 170 visitas.



FIGURA 25. LLANTA Z8

Roadtec Z8 Interac, llanta de altas velocidades en cualquier tipo de terreno ideal para motos de alto cilindraje, es la llanta más vendida por American Motos en la clasificación de alto cilindraje lo cual evidencia que los datos obtenidos por el log de la compañía no son muy distantes de sus ventas físicas en el almacén.

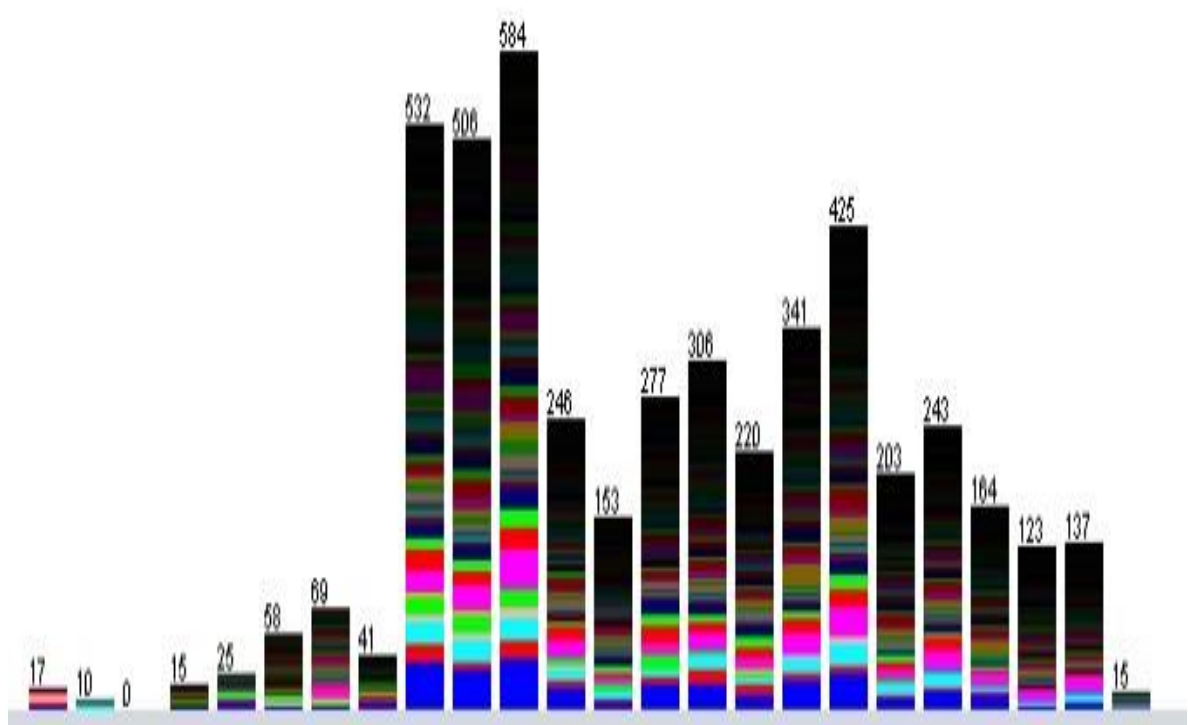


FIGURA 26. TRAFICO IP-HORA. ELABORACIÓN PROPIA

En el anterior grafico se aprecia las horas con que mayor frecuencia visitan la página web American Motos, las barras indican las horas de acceso, empezando desde la hora 0 y terminando en la hora 23, la altura representa la cantidad de visitas por parte de los usuarios.

Se destacan como horas de alto flujo las comprendida entre las 8 y las 10:59 de la mañana, y las 16 a 18 de la tarde. Las horas frías comprenden entre las 23 a 7 de la mañana, horas de menos actividad de los usuarios de American Motos.

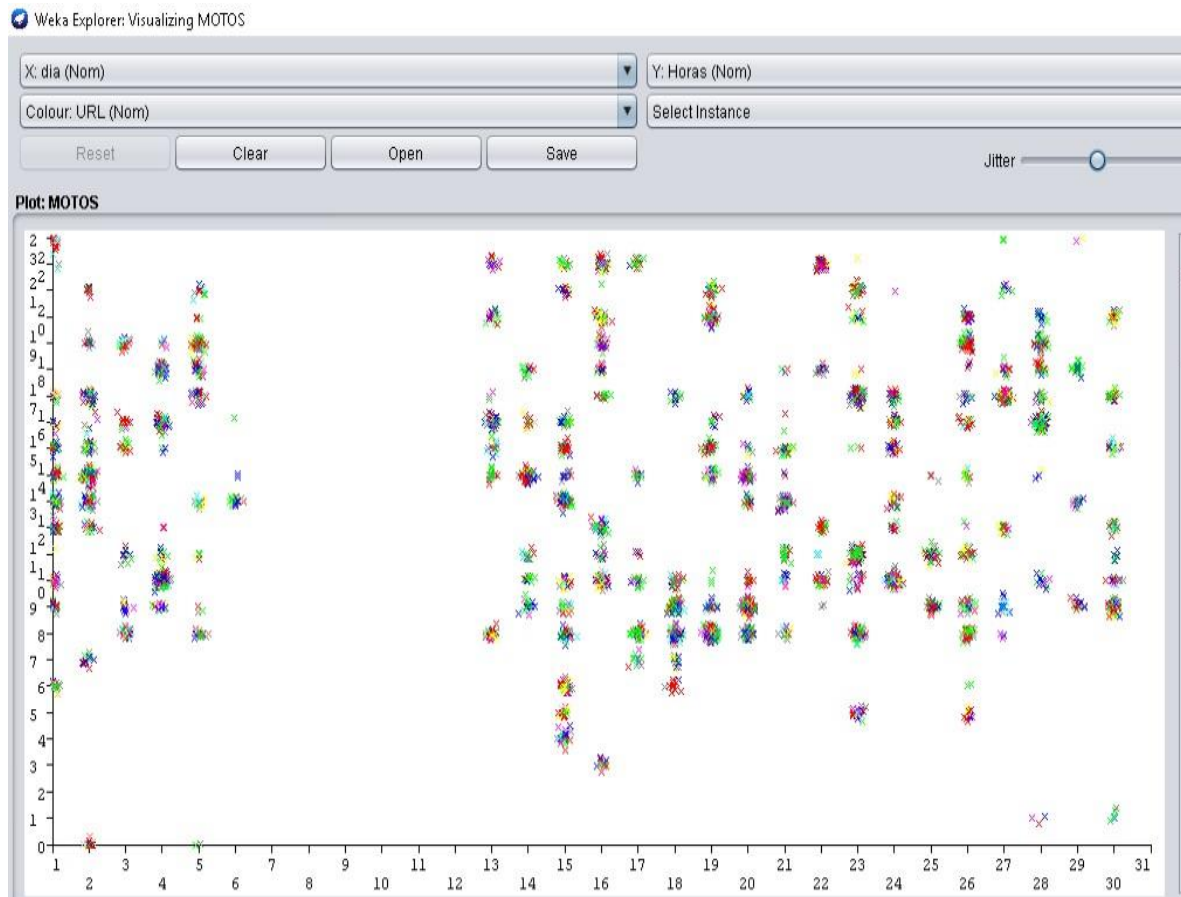


FIGURA 27. MATRIZ HORA-DÍA. ELABORACIÓN PROPIA

Como podemos observar en la anterior matriz publicada, se evidencian la cantidad de visitas y peticiones de los usuarios respecto al día y la hora de ejecución.

En la zona de las X's, se ubican los días estudiados en el proyecto, y en la zona de las Y's, las horas de ingreso de los usuarios, los colores representan las URL de las visitas.

El objeto de la matriz es encontrar zonas de mayor acumulación de respuesta para determinar las horas de mayor flujo con respecto a los productos de mayor cantidad de visitas por los usuarios de la página web.

En ella podemos evidenciar una mayor acumulación de visitas en los primeros 5 días del mes.

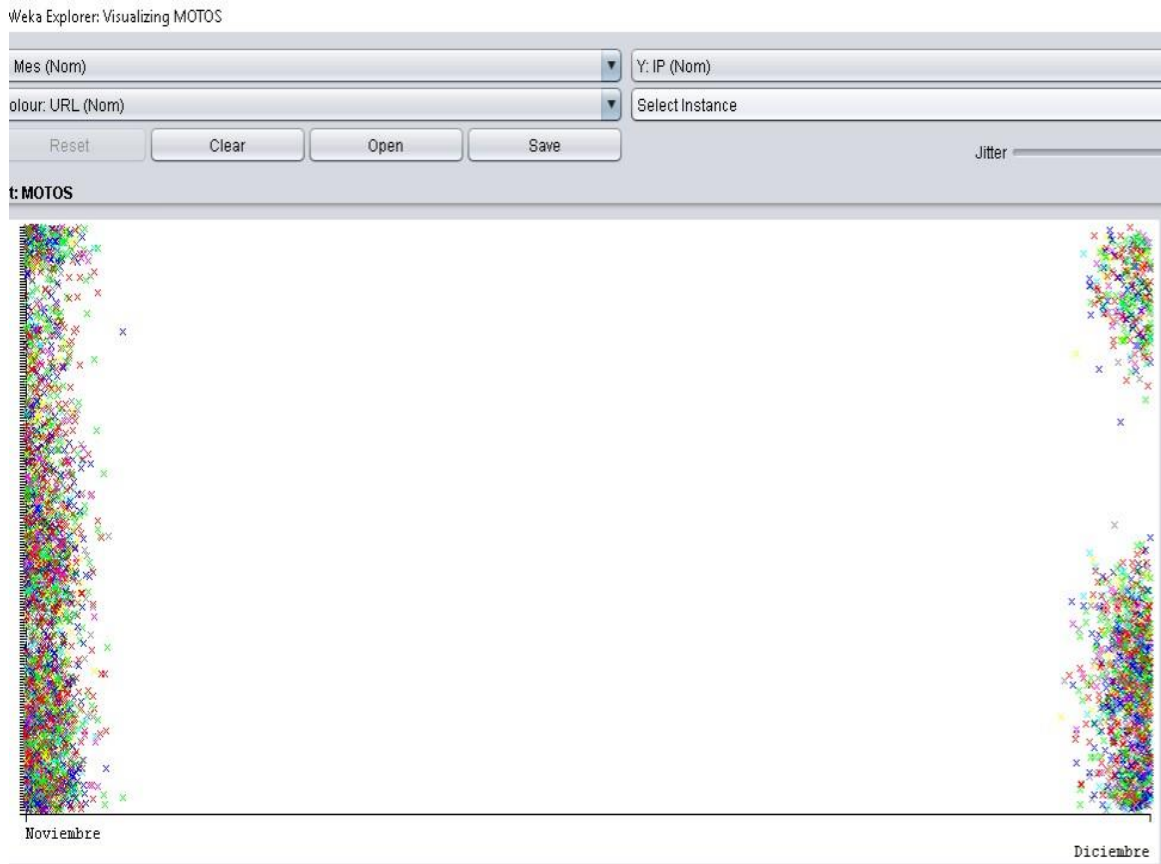


FIGURA 28. MATRIZ TRÁFICO POR MES. ELABORACIÓN PROPIA

En el anterior grafico se visualizan la cantidad de visitas entre los meses de noviembre y diciembre con sus respectivas URL's, observándose una mayor cantidad de visitas para el mes de noviembre.

El WEKA nos facilita las gráficas con respectos a las variables, aunque lo que interesa es obtener las reglas de asociación, ya que estas permiten observar la relación entre las variables de estudio, con la finalidad de crear estrategias, para obtener una mayor productividad del sitio.

7.2 Reglas de Asociación en WEKA

Las reglas de asociación nos muestran las tendencias de los usuarios de la página web American Motos, mediante su correcta interpretación podemos interpretar las acciones que posteriormente realizaran los visitantes, para la posterior toma de decisiones sobre la página web.

Obtenemos las reglas de asociación mediante el algoritmo apriori de WEKA, el cual arroja para 100 coincidencias encontradas la siguiente información:

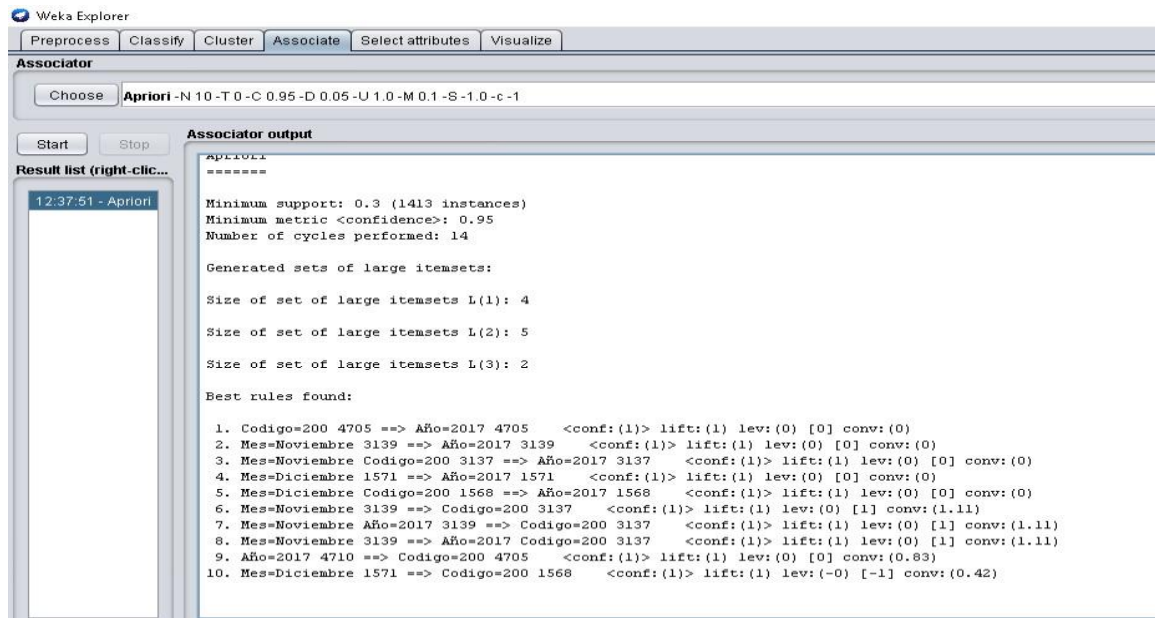


FIGURA 29. REGLAS ASOCIACIÓN. ELABORACIÓN PROPIA

En el año 2017 hubo 4705 ingresos exitosos por parte de los usuarios de American Motos con una confiabilidad del 100%

La IP 186.82.19.218 fue quien más consulto la página, con 825 visitas exitosas de las 829 registradas.

La hora con mayor flujo de visitas a la página American Motos fue las 10 con 584 registros exitosos.

Durante el mes de noviembre se registraron 3137 visitas exitosas, mientras que en diciembre se reportaron 1568 con código de estado 200.

La IP 190.205.244.25 realizo 669 consultas con un grado de éxito del 100% y se encuentra ubicada como la segunda en cantidad de registros a la página web.

De lo siguiente podemos inferir que los usuarios de la página web American motos, realizan peticiones con un flujo superior al 38% respecto a la media de visitas en el horario comprendido entre las 9 y las 11 de la mañana. Lo que nos indica una fuerte tendencia respecto a la hora de accionar de los usuarios de la compañía.

Los días de mayor cantidad de visitas a la página web American Motos son los 5 primeros días del mes con un incremento entre el 30 y 35% con respecto a la media del mes, dato relevante para posibles campañas publicitarias por parte de American Motos.

Las llantas para motos tipo enduro son las más visitadas por los usuarios, con un 40% más de visitas sobre el resto de tipos de llantas, dato que concuerda con las ventas físicas de la compañía ya que el 60% de sus ventas, son llantas para todo tipo de terreno o enduro.

Los usuarios que visitan la página web, tienden a realizar peticiones sobre los productos que ven en primer plano, es decir las imágenes o links que encuentra en el centro de la página de inicio de la compañía, con un 60% del total de visitas sobre el resto de productos. Lo que nos indica la más fuerte tendencia realizada en el estudio en el ámbito posicional de los productos.

7.3 Estrategias de mercadeo

En base al anterior análisis y los resultados arrojados por las reglas de asociación generadas por WEKA, se realizan una serie de estrategias de mercadeo para el producto, el precio, la plaza o distribución y la promoción o comunicación de la compañía American Motos.

7.3.1 Estrategias para el producto

- La compañía American Motos cuenta con línea económica de llantas las cuales son de fabricación china y tienen un margen de ganancia de cerca del 60%, los proveedores de esta línea brindan el servicio de marcado de llantas para compras superiores a las 100 unidades, por un costo adicional del 1,2 y 2 US, lo que permite lanzar una nueva serie de productos para sus clientes que sea marca propia, para dar a conocer de manera global la empresa.
- Adicionarles a los productos de la compañía servicios complementarios como servicio de montaje y mantenimiento para las ventas que así lo requieran. Dar a conocer las políticas de devoluciones y garantía para ofrecer a los clientes tranquilidad y confianza sobre el producto y la empresa.

7.3.2. Estrategias para el precio

- Lanzar al mercado la línea marca propia (American Motos), la cual se caracterizará por su bajo costo con el fin de lograr una rápida penetración, una rápida acogida y hacerla rápidamente conocida, aunque la utilidad por unidad se reduce entre el 4 y 7% respecto a la llanta fabricada con el sello del proveedor, la cantidad de ventas y de acogida de clientes aumentara gradualmente en el tiempo, logrando a corto plazo un incremento en la utilidad total de la compañía.
- Realizar compras de líneas Premium de alto costo, lo que atraerá clientes de un mercado creciente en el país como son las motos de alto cilindraje y así obtener nuevas ventas como producto de novedad.

- Ofrecer descuentos adicionales a nuestros clientes mayoristas por pronto pago, por volumen y dependiendo de la temporada ya que la cartera de la compañía tiene un retorno en algunos casos de más de 2 meses.

7.3.3. Estrategias para plaza o distribución

- Activar la casilla de ventas en línea en la página web de la compañía con el fin de aumentar las ventas totales de la compañía.
- Ofrecer el servicio de venta y asesoría por medio de correo electrónico y llamada telefónica para tener un mayor conocimiento de los clientes de American Motos.
- Brindar a los clientes servicio de montaje a domicilio para los productos en los que se pueda realizar dicho servicio.

7.3.4. Estrategias para la promoción o comunicación

- Brindar descuentos por primera compra del 10% a los clientes minoristas que realicen sus pedidos en línea, ya que la utilidad mínima por unidad de producto es del 30% y además de dar a conocer la compañía, se obtendrían ganancias adicionales debido a las ventas nuevas por este medio.
- Ofrecer a los clientes el servicio de notificaciones sobre productos nuevos y descuentos, para los visitantes que se encuentren registrados en la página web de la compañía.
- Realizar descuentos especiales en los primeros 5 días de cada mes con el fin de aumentar las ventas de los productos en línea en los momentos de mayor bonanza económica del mercado.

- Crear en la página web la opción de compras rápidas, es decir productos con descuentos especiales para productos de baja rotación por cierto lapso de horas. Para nuestro caso entre las 9 y 11 de la mañana, con la particularidad que las fechas y los productos serán determinados por los cambios del mercado.
- Participar en el evento de la feria de la rueda, evento que se realiza una vez cada año donde participan las empresas y marcas más influyentes del medio, para aumentar las relaciones comerciales de la compañía, lo que llevara a la atracción de nuevos clientes y un aumento significativo en las ventas al mayor.
- Anunciar en sitios de clasificados en internet lo que aumentara la cantidad de visitas tanto en la página web como en la tienda física de la compañía.
- Participar en el evento car audio y motos de alto cilindraje en sociedad con empresas dedicadas al tuning y personalización de motos, para dar a conocer las líneas y productos de la compañía a clientes minoristas, ya que es uno de los eventos más llamativos en la región y permite la instalación de publicidad en las locaciones destinadas a los participantes del evento.

8. Conclusiones

Mediante el análisis realizado a la página web www.americanmotos.com y con la ayuda de las técnicas de Data Mining aplicada a la web logramos identificar algunos patrones de comportamiento de los usuarios, que a simple vista pasan desapercibidos por la cantidad de información proveniente del log de la compañía.

Se logró identificar información relevante para el mejor funcionamiento de la página web, encontrando así cuales son las URL que con mayor frecuencia visitan la página y su sitio de procedencia, descubriendo que la página tiene importantes cantidades de visitas de países extranjeros a su origen, entre los que se destacan países como Venezuela y Ecuador.

Relacionamos las horas de mayor flujo de visitas comprendidas entre las 8 y 11 de la mañana con las URL e IP para tener una idea clara de los productos hacia los que se dirigen los usuarios de la página American Motos.

Definimos horarios de bajo intercambio de información, los cuales se ubicaron entre las 0 hasta las 7 de la mañana, para poder encontrar soluciones que permitan mayor regularidad en los ingresos a la página web, por medio de estrategias de mercadeo aplicadas a la compañía. Se evidencio la estabilidad del servidor lo cual permite un mayor flujo de información con el usuario, para lograr aumentar la cantidad de usuarios y clientes de la empresa American Motos.

Encontramos los productos de mayor atractivo para los clientes, quienes fueron los que en su imagen presentaban letras de color llamativo y se encontraban en la parte superior de las páginas.

Descubrimos errores en el diseño de la página web, los cuales impiden la visibilidad de algunos productos, lo que da a entender el porqué de los bajos registros y visitas de los mismos.

Este proyecto conglomeró conocimientos obtenidos a lo largo de la carrera de Ingeniería Industrial, bases de minería de datos, mercadeo, informática aplicada, fueron necesarios para obtener los resultados en dicha propuesta. Todas estas herramientas fueron utilizadas con un enfoque ingenieril para proponer mejoras y crecimiento de la página web de la compañía American Motos S.A.

9. Aportes y recomendaciones

Recomendaciones:

Eliminación de banners publicitarios, los cuales están entorpeciendo la visualización de los productos que se encuentran ubicados bajo el mismo.

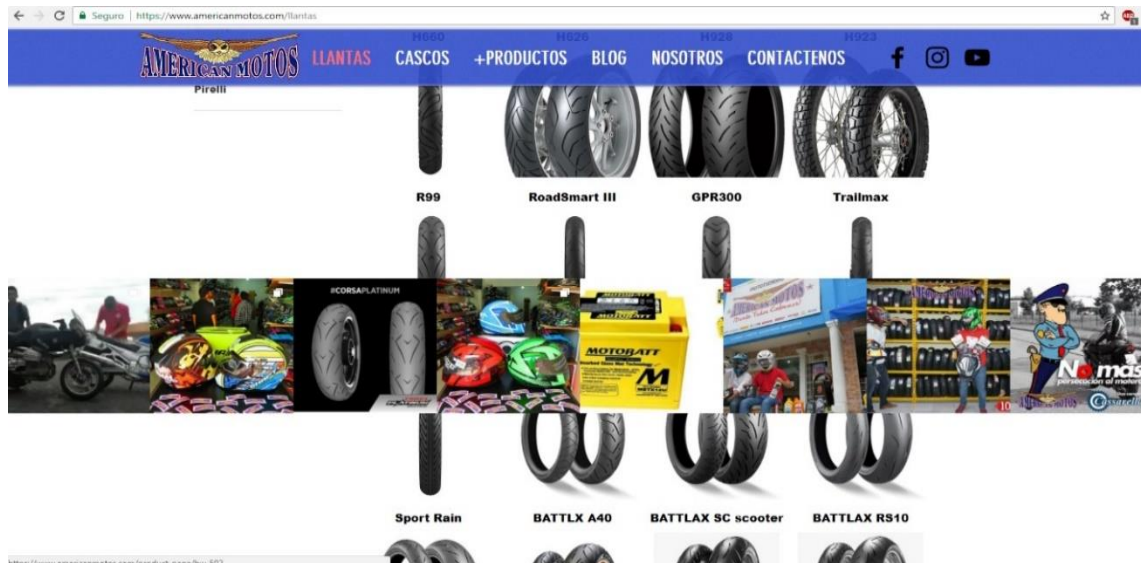


FIGURA 30. VISTA AMERICAN MOTOS WEB

Dichos banners limitan al usuario y en el pantallazo de descripción de cada producto ocultan la barra de número de contacto con la compañía, lo que lo imposibilitan a obtener una asesoría pertinente en el momento de realizar una visita al almacén físico o realizar compras por teléfono o correo electrónico.

Modificar la página para realizar compras en línea

Descripción, precio y categorización de los productos, para ayudar al cliente en la toma de decisiones. Ya que la pagina no cuenta con ningún tipo de ayuda para el usuario, que le permita saber que producto es el indicado para su moto.

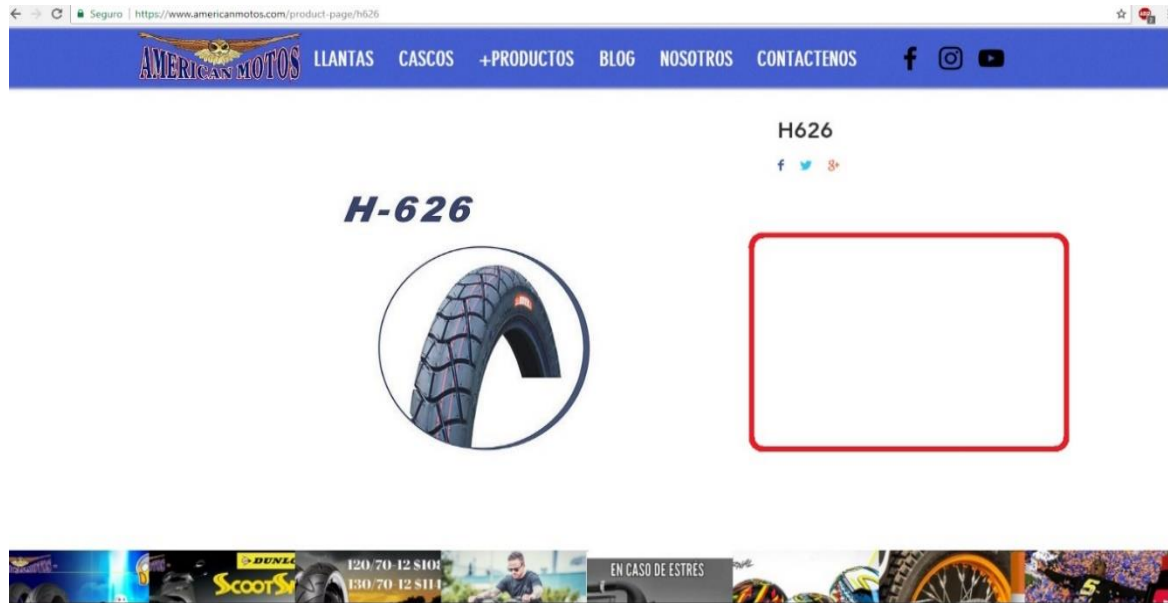


FIGURA 31. PANTALLA LLANTAS

Estrategias:

Ubicar en los productos de mayor visualización de la página, enlaces que permitan acceder a productos con pocas visitas, así como también productos alternativos y de descuento.

Realizar descuentos en las horas frías o de menor flujo comprendidas entre las 11 y 15 horas del día, y las 18 y 21 horas de la noche que incentiven las visitas en dichos tiempos.

Crear casilla de registro de usuarios, lo que permitirá a American Motos conocer la procedencia de sus clientes otorgándoles a ellos la posibilidad de obtener notificaciones vía correo electrónico o redes sociales, sobre los descuentos, productos y eventos de la página web, incrementando así el tráfico de la página y posterior cantidad de ventas en sus productos

10. Referencias Bibliográficas

- [1] J. A. M. Lenis, «El manejo de la información en la empresa,» *actualicese.com*, p. 1, 2013.
- [2] E. A. Falgueras, *Sistemas y servicios de información digital*, Barcelona: Edicions Universitat Barcelona., 2001, p. 15.
- [3] L. J. y C. O., *Autopistas Inteligentes*, Madrid: Fundesco, 1995, p. 240.
- [4] «Eginsoft,» 22 Febrero 2012. [En línea]. Available: <http://eginsoft.com/wordpress/?p=54>. [Último acceso: 10 Abril 2017].
- [5] J. E. P. Bharat y K. A., «WEBVIZ: A TOOL FOR WORLD-WIDE WEB,» Atlanta, GA, 1994.
- [6] J. A. Castañeda García, *El comportamiento del usuario de internet: Análisis de los antecedentes y consecuencias de la fidelidad.*, 2005.
- [7] J. H. Orallo, J. R. Quintana y C. F. Ramirez, *Introducción a la Minería de Datos*, Pearson Prentice Hall, 2004.
- [8] B. Gates y J. A. Bravo, *Los negocios en la era digital*, Barcelona: Plaza & Jans, 1999.
- [9] DANE y DPN, «Información mipymes censo 1990 y 2005,» 2012. [En línea]. Available: https://colaboracion.dnp.gov.co/CDT/Desarrollo%20Empresarial/INFORMACION_MI PYMES_act2012.pdf.
- [10] R. D. Martín, *Integración de modelos de reglas de asociación obtenidos de múltiples fuentes de datos*, La Habana, CU: Instituto Superior Politécnico José Antonio Echeverría, 2010.
- [11] N. Márquez, «Minería de Datos (Data Mining) vs Grandes Datos (Big Data),» *Tutaratech*.
- [12] Ecured, «Clustering,» *Ecured*.
- [13] J. Guillermo, «Las redes neuronales: qué son y por qué están volviendo,» 30 Diciembre 2014. [En línea]. Available: <https://www.xataka.com/robotica-e-ia/las-redes-neuronales-que-son-y-por-que-estan-volviendo>. [Último acceso: 20 Marzo 2017].
- [14] Art.61, «Constitución Política de Colombia,» 1991.

- [15] Art.56, Reglamento Estudiantil, Bogotá: Universidad Santo Tomás, 2008.
- [16] t. u. o. W. WEKA, «Can I use WEKA in commercial applications?,» 2017.
- [17] R. Krishnamoorthi, «Identifying user behavior by analyzing web server Access Log File,» Anna University, Tiruchirappalli, 2009.
- [18] R. T. Delarbre, «Vivir en la Sociedad de la Información: Orden global dimensiones locales en el universo digital,» *Revista iberoamericana de ciencia, tecnología, sociedad e innovación*, 2001.
- [19] «Data mining y minería web: aplicaciones de la minería de datos,» 2015. [En línea]. Available: <https://blog.es.logicalis.com/analytics/data-mining-y-mineria-web-aplicaciones-de-la-mineria-de-datos>. [Último acceso: 15 Marzo 2017].
- [20] S. S. Francia y M. N. M. García, «Minería web para el comercio electrónico,» Universidad de Salamanca, Salamanca, ES, 2009.
- [21] DANE, «Proyecciones de Población departamentales y municipales por área 2005 - 2020,» Bucaramanga, 2011.
- [22] J. Han, M. Kamber y J. Pei, *Data Mining Concepts and Techniques*, Waltham, USA: Elsevier, 2012.
- [23] Itelligent, «Itelligent,» 12 Febrero 2015. [En línea]. Available: <http://www.itelligent.es/es/mineria-web-de-contenidos-estructuras-usos/>. [Último acceso: 2 Mayo 2017].
- [24] S. C. Fuentes Reyes y M. Ruiz Lobaina, «Minería Web: un recurso insoslayable para el profesional de la información,» *ACIMED*, 2007.
- [25] «Osmosis Latina,» [En línea]. Available: <https://www.osmosislatina.com/analog/basico.htm>. [Último acceso: 5 Mayo 2017].
- [26] O. Maimon y L. Rokach, *DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK*, Israel: Springer, 2010.
- [27] D. J. Belamate, M. A. Cassani y C. P. Ricci, «Aplicación de reglas de asociación para la detección de patrones de comportamiento en sistema académico universitario.,» Córdoba, Argentina.
- [28] J. Orallo Hernández y C. Ramirez Ferri, «Prácticade Minería de Datos, Introducción al WEKA,» Universitat Politècnica de València, Valencia, España, 2006.

- [29] M. Rouse, «TechTarget,» [En línea]. Available: <http://searchsecurity.techtarget.com/definition/access-log>. [Último acceso: 2 Mayo 2017].
- [30] M. Rouse, «TechTarget,» [En línea]. Available: <http://searchcrm.techtarget.com/definition/Web-mining>. [Último acceso: 5 Mayo 2017].
- [31] «Userservers, web hosting,» [En línea]. Available: http://web.userservers.net/ayuda/soluciones/dominios/que-es-una-direccion-ip_NTk.html. [Último acceso: 3 Mayo 2017].
- [32] «HAIKU,» [En línea]. Available: <https://www.haiku-os.org/docs/userguide/es/attributes.html>. [Último acceso: 19 Marzo 2017].
- [33] T. Menzies y Y. Hu, Data Mining for Busy People, IEEE Computer, 2003, pp. 18-25.
- [34] GNU, «GENERAL PUBLIC LICENSE (GPL),» GNU, 29 Junio 2007.
- [35] R. Baeza Yates, «Tendencias en minería de datos de la Web,» *El profesional de la información*, pp. 5-10, 2009.
- [36] A. Cutro, «Mineria de datos aplicada a la encuesta permanente de hogares,» DataPRIX.
- [37] G. R. Filocamo y C. I. Chesñevar, «Formalización de Web Mining,» Universidad Nacional del Sur, Bahía Blanca, Argentina.
- [38] J. U. Arias Rivera y C. E. Castro Polo, «ESTUDIO WEBMINING Y DATAMINING DEL LOG HUAWEI LTDA,» Universidad de Magdalena, Magdalena, 2012.
- [39] A. K. Singh, A. Kumar y A. K. Maurya, «Association rule mining for web usage data to improve websites,» *International Conference on Advances in Engineering & Technology Research (ICAETR - 2014)*, pp. 1-6, 2014.
- [40] D. M. Rodriguez, Integración de modelos de reglas de asociación obtenidos de múltiples fuentes de datos, La Habana, CU: D - Instituto Superior Politécnico José Antonio Echeverría. CUJAE, 2010.
- [41] F. Laudon y J. Laudon, «Sistemas de Información,» Editorial Diana, México, 1996.

Anexos

Anexo 1. **American motos 2018.arff**

Anexo 2. **Log tabulado American motos 2018.xlsx**