Universidad Autónoma de Manizales
Facultad de ingenierías

Búsqueda de patrones en el comportamiento de los visitantes de la plataforma "Oferto" de la Cámara de Comercio de Armenia y del Quindío, a través de la aplicación de minería web.

Tesis para optar al título de Magíster en Gestión y
Desarrollo de proyectos de Software

Manizales
2016

Universidad Autónoma de Manizales
Facultad de ingenierías

Búsqueda de patrones en el comportamiento de los visitantes de la plataforma "Oferto" de la Cámara de Comercio de Armenia y del Quindío, a través de la aplicación de minería web.

Presentado por:

David Alberto Angarita García
Juan José Muñoz Franco

Director:

Msc. Javier Hernández Cáceres

Manizales
2016

Índice

Índice de ilustraciones y gráficos

Índice de tablas

Introducción

En la actualidad las organizaciones cuentan con información generada cada vez más rápido y de manera exponencial por el uso de las páginas web por parte de los usuarios. Dicha generación de información se debe a la publicación de sus productos y/o servicios en sus sitios web y la interacción de los usuarios con los mismos, es por esto que se vuelve necesario el análisis de dicha información con el objetivo de ser competitivo y obtener utilidades, usando como medio el mundo digital.

Tal como afirma Baeza-Yates (2005) “la información de la web es finita pero el número de páginas web es infinito”, a partir de esta premisa es claro que se cuenta con información valiosa para la gestión de la Organización. Sin embargo, para que esta información pueda tener un impacto adecuado se debe realizar un proceso con las técnicas apropiadas, ya que la mayoría de veces la información importante no se encuentra a simple vista y si no es utilizada y explotada de la forma correcta o simplemente no se hacen las búsquedas adecuadas dentro de la misma, termina por convertirse en datos sin valor.

El Censo Empresarial que adelantó la Cámara de Comercio de Armenia y del Quindío (CCAQ) en el año 2012 a 13.857 establecimientos de comercio del departamento, entre formalizados y no formalizados, suministró información sobre el panorama que en materia de conectividad y utilización de herramientas asociadas a las Tecnologías de Información y Comunicación (TIC), tienen los pequeños comerciantes del Quindío. Sólo el 2% promueve su negocio a través de alguna herramienta web, el 56% manifestó que no lo hacen, y el 42% respondió que no las conocen, así que un 98% de estos empresarios, de acuerdo con lo que indican estos datos, realizan su actividad mercantil al margen de este tipo de herramientas.

Una vez que se identificó esta situación, en el año 2015 la CCAQ realizó el primer observatorio de las TIC con una muestra de 237 empresas del departamento del Quindío, encontrando que el 61% de ellas utiliza equipos informáticos y de comunicación para desarrollar las actividades requeridas por la empresa, así mismo se identificó que un 42,1% no tiene computadores en su empresa, ya que desconoce los beneficios de implementarlos en su actividad empresarial, además solo el 29,30% hace uso de una página web propia y del total solo un 24,60% está asociada a alguna página que pueda promocionar su negocio, también se pudo identificar que el 68,10% de las empresas afirma no comercializar sus productos en internet.

A partir de este contexto, se concluye que la gran mayoría de empresas del Quindío, especialmente las micro empresas, es decir aquellas que tienen como máximo 10 trabajadores o activos de hasta 500 salarios mínimos mensuales legales vigentes y las pequeñas, es decir aquellas que tienen de 11 a 50 empleados o entre 501 y menos de 5.000 salarios mínimos mensuales legales vigentes según la Ley 905 de Agosto 2 de 2004, no participan en un proceso de conectividad y aprovechamiento de las herramientas que las TIC ofrecen de manera adecuada, en especial para los procesos de mercadeo y venta de su oferta de productos y servicios. Así mismo, que a pesar de que el Quindío es un departamento considerado uno de los principales destinos turísticos del país, su estructura comercial presenta debilidades en cuanto a capacidad para alinearse con las nuevas tendencias y hábitos de compra que ha ido ganando no sólo el turista, sino el ciudadano local, como resultado de todas las alternativas que ofrecen las TIC; entre estas, poder ubicar información oportuna y pertinente sobre condiciones especiales que apliquen para cierta oferta de productos, como las ofertas y descuentos especiales, bonos, y otras herramientas de captación y fidelización de clientes, que pueden ser fácilmente encontrados, a través de un sitio web o una aplicación para dispositivos móviles.

Por lo tanto, se llevó a cabo la implementación de la plataforma “Oferto” que busca mejorar los procesos de mercadeo y venta de los empresarios del departamento a través de herramientas de *marketing* móvil y el aprovechamiento de funcionalidades complementarias que les permitirán incrementar ventas, fidelizar clientes y conformar una red social de comerciantes y compradores en favor de sus establecimientos de comercio, plataforma en la cual los consumidores pueden encontrar todos los productos y/o servicios que hay actualmente con descuento en el departamento, teniendo a la fecha 1.185 empresas registradas en dicha plataforma y en promedio 2.793 visitantes mensuales.

Sin embargo, existen actualmente problemas a los que se ven enfrentados en “Oferto”, tales como la dificultad para determinar información relevante dentro de los accesos generados en la plataforma, la forma en la que se debe mejorar la distribución de los productos y/o servicios, la identificación de la intención de compra de los visitantes, las relaciones entre fechas y tendencias de consumo, la probabilidad de venta de los productos que son publicados, las características de los visitantes y sus intereses, entre otros.

Lo anterior debido a que los datos no son fáciles de identificar, evaluar y analizar; de allí la importancia de la *web mining* o minería web, ya que dicha área de investigación tiene un enfoque importante en el movimiento económico que se ha generado a través del comercio electrónico, buscando resolver los problemas antes mencionados; todo esto con el fin de identificar si existen o no patrones de comportamiento de los visitantes y si son encontrados, a través de estos ayudar a resolver los problemas planteados por la organización; para establecer futuros proyectos encaminados al mejoramiento continuo de la plataforma.

Así pues, se hace necesario ejecutar este análisis a través de técnicas de minería de datos, ya que la inmensa cantidad de información con la que se cuenta no puede ser procesada de manera

manual, además dichas técnicas permitirían identificar patrones no evidentes en los datos, logrando, de esta forma, ayudar a resolver los problemas del negocio que de otra manera consumirían mucho tiempo o simplemente no serían posible responder.

1. Referente contextual

1.1 Área Problemática

La Cámara de Comercio de Armenia y del Quindío (CCAQ) es una entidad privada sin ánimo de lucro que tiene como fin brindar servicios de calidad, pertinentes y oportunos para 11.365 empresarios de la región (renovados al 31 de Marzo del 2016), dicha entidad busca ser un ente que permita unir fuerzas privadas, públicas, académicas, productivas y las demás necesarias para generar competitividad, desarrollo e innovación en ellos, teniendo la ejecución de proyectos como una de las principales actividades para el logro de tal fin, por medio de su departamento comercial y de proyectos, dentro de los cuales se han obtenido cifras en programas y proyectos de intervención y apoyo social con alrededor de 74.000 beneficiarios, así mismo en programas y proyectos de reactivación económica una cifra cercana a 1500 entre personas y empresas, de igual manera la CCAQ actúa como gerente o interventor de 111 proyectos de infraestructura.

Actualmente, tal como lo informa en su página web, ejecuta otros programas como los que se describen a continuación:

Arte de lo hecho a mano: iniciativa que busca el fortalecimiento competitivo del sector artesanal del Quindío mediante estrategias de diseño del producto, herramientas de comercialización y mejoramiento de la capacidad gerencial

Iniciativa Kaldia: ruta competitiva del café en el Quindío. Apuesta por el café de calidad, micro lotes, formación continua de caficultores en el proceso de cafés especiales y educación del consumidor.

Iniciativa Artemis: ruta competitiva del sector cuero en el Quindío. Apuesta por el cuero de calidad para abastecer industrias de alta gama. Especialización en procesos de curtiembres.

Rutas del paisaje cultural cafetero: proyecto de desarrollo del clúster de turismo, bajo la implementación de un modelo competitivo, sostenible y participativo para el fortalecimiento de las Mipymes "micro, pequeña y mediana empresa" y el Destino.

Fortalecimiento empresarial para desplazados: fortalecimiento socio-empresarial, productivo y comercial de unidades productivas por población en situación de desplazamiento forzado, ubicadas en el Quindío.

Competitividad del clúster turístico del Quindío: mediante el diseño de la oferta, la capacitación del talento humano y la certificación en calidad de las empresas del sector.

Quindío, destino de experiencias: propuesta de innovación y diversificación de la oferta de producto turístico en el marco del paisaje cultural cafetero (PCC) para un mercado local e internacional.

Apoyo a empresas de reciente creación: alianza de cooperación para fortalecer las microempresas de reciente creación quindianas en respaldo y acompañamiento, permitiendo llegar a nuevos mercados.

Mejoramiento competitivo del sector: apoyo a empresas del sector agroindustrial del Quindío para incrementar su capacidad comercial y aprovechar de manera eficiente las oportunidades de mercado

Proyecto TIC: mejoramiento de procesos de mercadeo y venta de los establecimientos mediante una herramienta de conectividad basada en el uso de una aplicación de marketing móvil.

Programa cafés diferenciados: fortalecimiento de la cadena productiva de cafés especiales del departamento del Quindío, componente agroindustrial, eslabón tostión y molienda, bebidas y otros complementarios.

Proyecto de competitividad: Gestionado por la Comisión Regional de Competitividad e Innovación del Quindío y cofinanciado por INNPULSA por medio del convenio IFR 005-007.

Como se observa, para la CCAQ es relevante el acompañamiento de las empresas desde todas las actividades económicas posibles, por lo que generar estrategias que permitan facilitar la toma de decisiones en cuanto a lo que los consumidores desean tiene un alto impacto para la organización, es de allí que nace la plataforma "Oferto", la cual se compone de tres herramientas como lo son un sitio web (www.Oferto.co), una aplicación móvil llamada "Oferto", disponible en las tiendas de aplicaciones y un sitio web autoadministrable entregado a cada empresa que hace parte de la plataforma, a través del cual pueden administrar sus productos y/o servicios, con el fin de poder publicar fácilmente las promociones que deseen ofrecer a través de este servicio, además sirve como el sitio web de la empresa en caso de no contar con uno.

A continuación, se representa y describe la estructura general de "Oferto":

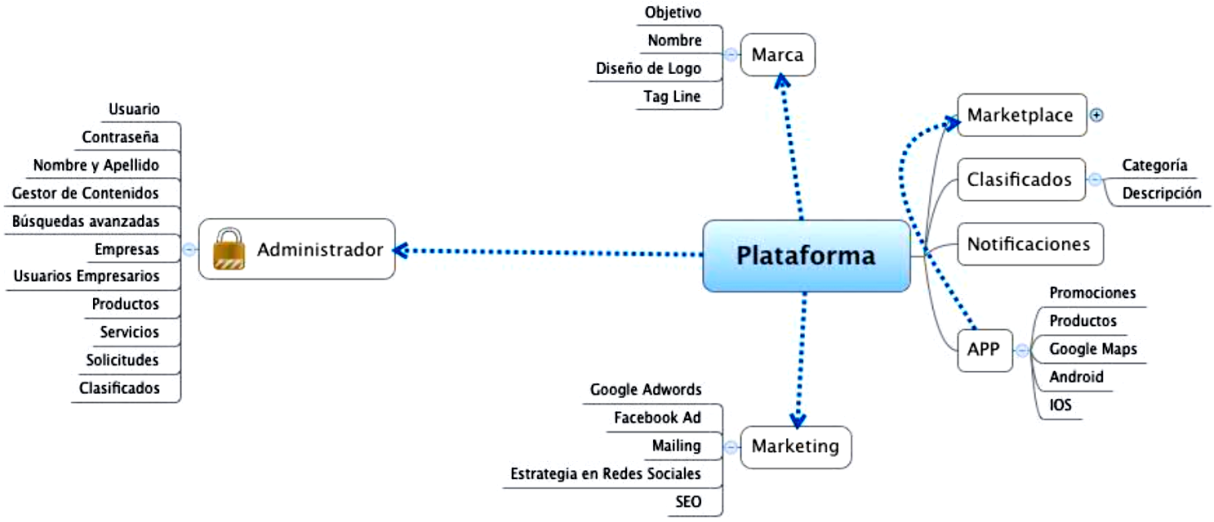


Ilustración 2: Estructura “Oferto”

En términos teóricos, lo anterior puede describirse de la siguiente manera:

- **Marketplace y App:** es el portal web www.Oferto.co y la aplicación móvil en la cual el visitante encuentra la oferta de servicios y productos de la región; esta incluye:
    - Las ofertas destacadas.
    - Productos y servicios que publican las empresas.
    - Categorías: comercio, turismo, inmobiliario, automotriz, servicios, etc.
    - Buscador avanzado de productos, servicios y empresas:
        - Palabras clave
        - Productos y/o servicios.
        - Categorías
        - Sector
        - Ubicación
        - Rangos de precios

- Consulta de promociones.
- Suscribirse a recibir promociones vía e-mail.
- Reservar productos.
- Calificar a las empresas
- Ubicación geográfica de las empresas.

- **Sitio web empresa:** las empresas cuentan con una página web propia con las siguientes características:
  - Administrador de contenidos.
  - Nombre de la empresa.
  - Datos de ubicación y contacto.
  - Categorías de los productos y/o servicios
  - Productos y servicios.
  - Promociones.
  - Mailing a clientes.
  - Subdominio
  - Recibir notificaciones de contacto desde el marketplace y la app
  - Datos del representante.

Dentro de la plataforma se tienen los tres tipos de usuarios que se describen a continuación:

- **Administrador del sistema:** persona encargada de administrar todas las secciones de la plataforma web realizando actividades como:
  - Gestión de contenidos
  - Administración de las empresas
  - Administración del marketplace

    - Administración de las ofertas.
    - Administración de los usuarios.
    - Generación de reportes exportables a Excel y por rangos:
        - Número de empresas registradas
        - Número de ofertas publicadas
        - Número de visitas a las ofertas
        - Ofertas más consultadas.
        - Reservas generadas por medio de la plataforma.
        - Pines, códigos o identificadores generados de compras o reservas.
        - Usuarios

- **Visitante**: es la persona que entra al marketplace, mini sitios o hace uso de la app para ver y reservar las ofertas de las empresas. Estos usuarios podrán etiquetar empresas como favoritas y recibir notificaciones por parte de ellas.
- **Empresario**: crea perfil como empresa, publica sus productos y servicios en el mini sitio y crea promociones en el marketplace y app.

Basados en datos entregados por los administradores de la plataforma, en promedio se han recibido 2793 visitas mensuales de consumidores que buscan y reservan los productos y servicios que se publican a través de “Oferto”, esto ha hecho que se cuente con información generada por la publicación de productos y servicios por parte de los empresarios, a lo que se le suma la interacción que tienen los visitantes con ellos al realizar su proceso de navegación dentro de la plataforma, sin embargo dicha información actualmente no está siendo usada para dar respuesta a preguntas que se plantean los responsables de la plataforma a la hora de proponer mejoras en los proceso de publicación, actualización y mejoramiento de la misma, como lo son: ¿Cómo mejorar

la distribución de los productos y/o servicios en la plataforma?, ¿Cuáles son las intenciones de compra de los consumidores del departamento?, ¿Qué relación puede existir entre las fechas y las tendencias de consumo?, ¿Qué probabilidad hay de que un producto que se publique sea comprado?, ¿Cómo es un visitante recurrente?, ¿Cuáles son los intereses de los visitantes? , ¿Cómo identificar visitantes con necesidades similares?

Por todo lo anterior, es importante llevar a cabo un proceso de minería web que apunte a la consolidación de la información almacenada hasta el momento y lograr avanzar en la búsqueda de identificación de la existencia o no de patrones de comportamiento de los visitantes, para ayudar a dar respuesta a las preguntas planteadas por la organización, a través de estos, en caso de ser encontrados, con el fin de establecer futuros proyectos encaminados al mejoramiento continuo de la plataforma.

1.2 Antecedentes

1.2.1 *Predicting User's Web Navigation Behavior Using Hybrid Approach*

Este proyecto realizado por Narvekara y Sakina Banu (2015) del Departamento de Ingeniería Informática D.J.Sanghvi de la Facultad de Ingeniería, Mumbai, India, tenía como propósito aumentar la precisión al momento de predecir el comportamiento de navegación de los usuarios, reduciendo la complejidad de la predicción y buscando resultados eficientes, mediante la minería web de uso, clasificando los modelos de predicción en dos categorías, como lo son los modelos basados en caminos y los modelos basados en rutas, teniendo como principal restricción su base en la historia y los conocimientos previos del usuario.

En esta investigación se describe el Modelo oculto de Markov, el cual tiene un corte estadístico en el que se asume que el sistema a modelar es un proceso de Markov de parámetros desconocidos, en el que el objetivo es determinar patrones desconocidos, utilizando la Regla Dempsters, que es una extensión de la teoría de la probabilidad para describir incertidumbre sobre la evidencia, además de un hibrido entre los modelos antes mencionados Los resultados del proyecto indicaron que el sistema mejora la precisión de la predicción sin comprometer el tiempo de predicción, manejando dos niveles; el nivel uno usado para la formación y el dos para la predicción, funcionando mejor el modelo hibrido para una mayor secuencia y una mayor predicción, logrando aumentar tanto el tráfico, como la ayuda al servidor a gestionar los recursos.

1.2.2 *Effective web log mining and online navigational pattern prediction*

El segundo proyecto reseñado, fue el realizado por Guerbas, Addam, Omar Zaarour, Nagi, y otros (2013) del Departamento de Ciencias de la Computación de la Universidad de Calgary, Canadá, que tenía por objetivo principal proponer un marco referente de minería web para predecir el comportamiento de la navegación en línea, revisando cada uno de los pasos dentro del proceso de minería web como son la limpieza, preparación, pre-procesamiento de los datos, el procesamiento y el análisis de los resultados.

En medio del proceso se revisaron diferentes algoritmos como los de agrupamiento, DBSCAN, K-*Means, PageGather, Optics*, para identificar en ellos, las ventajas y desventajas en el proceso de identificación del comportamiento de los visitantes de un sitio web, lo cual evidenció diversas dificultades en los cierres de sesión de los usuarios, por lo que se propuso un

tiempo medio de 30 minutos para una sesión, con el fin de buscar falsos positivos dentro de la información. Para esto, se postuló el algoritmo DBSCAN para el proceso de agrupación, ya que no es necesario tener un conocimiento previo sobre el número de agrupaciones y también porque detecta valores atípicos. Lo anterior permite contar con diferentes posibilidades para evaluar tanto sesiones como patrones de visitantes.

1.2.3 *Dynamic Recommendation System Using Web Usage Minning for E-Commerce*

Un tercer trabajo que se reseña, es el ejecutado por Prajyoti Lopes y Bidisha Roy., (2015) del *Departament of Computer Engineering, St Francis Institute of Technology* de Mumbai. Este trabajo se enfocó en generar sugerencias de productos en tiempo real sin importar si los clientes se registraron ante la plataforma o no.

Para cumplir con dicho objetivo, se desarrolló un proceso de minería de datos con el fin de descubrir patrones y reglas ocultas, dichos patrones se definen con los datos del servidor de acceso y algunos datos adquiridos proactivamente. Posteriormente, se limpiaron los datos y se enviaron al sistema de sugerencias, que mediante el uso de tres técnicas distintas para las proponer sugerencias, le proporciona al usuario nuevos productos con base en *patrones* descubiertos con la información previa.

Gracias a este trabajo se logró incrementar las sugerencias exitosas que la plataforma proporciona a un valor de entre el 80% y el 85%.

1.2.4 Sistema de apoyo para la acreditación de la calidad de programas académicos de la universidad de caldas, aplicando técnicas en minería de datos

Este proyecto fue realizado por Juan Carlos González Cardona de la Universidad Autónoma de Manizales (2011). La investigación tuvo como objetivo determinar los factores de la alta deserción estudiantil que se presentaba en la Universidad de Caldas. Para conocer dichos factores, se utilizó la metodología CRISP – DM con la herramienta para minería de datos *RapidMiner*.

Los resultados obtenidos en esta tesis mostraron los factores de riesgo y de protección para los estudiantes de cada programa presencial dentro de la universidad.

1.2.5 Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo.

Esta tesis fue presentada por Juan Miguel Moine de la Universidad Nacional de La Plata en Buenos Aires, Argentina (2013). El objetivo de este trabajo fue realizar un análisis comparativo entre las metodologías de minería más difundidas. Para realizar dicho análisis se elaboró un cuadro comparativo con las características de cada metodología. Con el cuadro en mención se analizó el nivel de especificación de las tareas, los escenarios de aplicación, las actividades que componen cada fase del proceso y las actividades destinadas a la dirección del proyecto.
Como conclusión de esta tesis, se señaló que si bien las metodologías CRISP – DM y *Catalyst* cumplen con muchos de los puntos a evaluar, se sugieren algunos otros para mejorar con respecto a la dirección del proyecto.

1.2.6 Análisis del comportamiento del usuario web

La tesis doctoral realizada por Pablo Enrique Román Asenjo en la Universidad de Chile (2011), tuvo como objetivo la aplicación de teóricas sobre la neurofisiología en la toma de decisiones para describir el comportamiento de navegación de los usuarios web, teniendo que identificar patrones de comportamiento del usuario web como parte del proceso. Esto dejó como resultado la evidencia de una posible aplicación de teorías sobre la toma de decisiones para explicar el comportamiento de navegación de dichos usuarios.

1.2.7 Clasificación difusa para descubrir perfiles de usuarios en la web

En este trabajo realizado por Oscar Fernando Bedoya Leiva en la Universidad del Valle (2013), tuvo como fin proponer una nueva estrategia que combina técnicas de *clustering* y clasificaciones difusas, para predecir el comportamiento que en términos de navegación tendría un usuario. Teniendo como resultado las ventajas en cuanto a la exactitud en la tarea de clasificación.

1.2.8 Análisis de algoritmos de aprendizaje automático para la caracterización de usuarios de la Web

Esta tesis fue presentada por Moisés Torres García (2015) perteneciente a la Universidad Politécnica de Madrid, la investigación tuvo como objetivo realizar un estudio de las fases, técnicas y metodologías que se usan actualmente en el campo del análisis de datos y el aprendizaje automático.

Teniendo como base el estudio anterior, se identificaron las más utilizadas y se aplicaron sobre un ejemplo perteneciente al ámbito de la minería web. De este proyecto se pueden resaltar los resultados obtenidos al aplicar técnicas para dotar a los datos de un formato que facilite su posterior análisis, gracias a esta preparación previa se pudo incrementar el poder predictivo de las variables. De igual manera, se pudo concluir que existe una relación estrecha entre la posibilidad de predicción y el segmento de mercado, ya que aquellas webs con un segmento amplio, presentaron mayores dificultades a la hora de caracterizar a los usuarios.

1.3 Justificación

La creciente complejidad en la toma de decisiones por parte de los usuarios que ingresan a los servidores, dentro del amplio campo de la competitividad entre las empresas que brindan productos o servicios, obliga a buscar técnicas que permitan aprovechar toda la información a la que se tiene acceso con el fin de generar conocimiento al servicio de las empresas. Así pues, en la actualidad, existen sistemas exploratorios de los *weblogs* (*Software Analog*), o a partir del uso de *cookies* como el Google *Analytics* que aportan, no solo elementos de ámbito descriptivo, sino poco personalizable, esto debido a que solo trabajan con una pequeña porción de los datos comparado con lo que se tiene en el servidor, lo que genera dificultades, ya que estos son capturados de manera remota por medio de rastreadores o etiquetas, disminuyendo su exactitud, ya que depende de la admisión de las cookies en el navegador del visitante, también problemas con algunos firewall u otro sistemas de seguridad , falta de trazabilidad de muchos de los archivos procesados y dificultad al necesitar reprocesar los datos. Mientras que al trabajar con los log del servidor no es necesario hacerlo en tiempo real, además la cantidad y calidad de datos con los que se cuenta es mayor y solo se puede obtenerlos con el visto bueno del administrador del

sistema, sumándole que puede ser implementado el proceso sin necesidad de tocar las páginas de la plataforma, facilidad para reprocesar los datos y se evitan problemas con sistemas de seguridad, entre otros.

En el departamento del Quindío prevalece una combinación de desconfianza y desconocimiento sobre los beneficios que se pueden percibir mediante el uso de nuevas tecnologías, particularmente sobre el uso del internet y las aplicaciones móviles como un canal publicitario y de ventas entre comerciantes y clientes. Esto respaldado por las cifras arrojadas en el primer observatorio de las TIC realizado por la CCAQ en el año 2015, en las que se evidencia que solo el 26,30% de las empresas tiene una página web propia, mientras que un 20% participa en otras páginas o plataformas como es "Oferto" y el 68,10% de los empresarios afirma no comercializar sus productos en internet; por otra parte en cuanto a la inversión de sus utilidades en la implementación de las TIC, el 42,60% señaló que invierte entre el 5 y el 20%, mientras que el 46,30% no especificó cuánto representaba el porcentaje de inversión.

Es debido a esta razón que la Cámara de Comercio de Armenia y del Quindío inició la búsqueda de técnicas que permitan la adquisición de un conocimiento específico para ponerlo al servicio de sus empresarios, con el fin de impulsar sus ventas e incrementar la confianza en el uso de plataformas tecnológicas y, por ende, en todos los aspectos relacionados con la economía del sector.

Adicionalmente, llevar a cabo un proceso de minería web al archivo de *log* del servidor de la plataforma "Oferto" es, no solo importante, sino necesario debido a que posiblemente permitirá identificar datos vitales para mejorar el funcionamiento y/o rendimiento de la plataforma a partir de la información generada por la utilización de la web por parte de los visitantes, para permitir

así la CCAQ continuar con su esfuerzo de modernizar los servicios que se le brindan a los empresarios del departamento.

Todo lo anterior podría ser logrado, pero depende de la calidad del conjunto de datos a analizar, ya que un *log* con datos erróneos arrojará resultados equívocos que no reflejan la realidad del sector, por lo que solo será posible garantizar una mejora continua en este proceso a través de la retroalimentación continua de los resultados de Oferto por parte de sus representantes y las empresas, para identificar las señales que arroje el mercado.

1.4 Formulación del problema: pregunta de investigación

¿Cuáles son los patrones de comportamiento más representativos de los visitantes de la plataforma Oferto de la CCAQ que permitan mejorar la competitividad y la rentabilidad de las empresas de la región?

1.5 Objetivos

1.5.1 Objetivo General

- Identificar la existencia de patrones de comportamiento de los visitantes de la plataforma “Oferto” de la CCAQ combinando técnicas de minería web al log de acceso generado por esta.

1.5.2 Objetivo Específicos

- Construir el *Data Warehouse* a partir del proceso de ETL (*Extract, Transform and Load*).
- Aplicar la metodología CRISP-DM sobre el *Data Warehouse* diseñado con el fin de encontrar posibles patrones de comportamiento de los visitantes.
- Caracterizar las sesiones de usuario que se realizan en la plataforma "Oferto".
- Validar los resultados obtenidos en el proceso de minería web siguiendo la metodología CRISP-DM.

1.6 Resultados esperados

Resultados y/o productos	Beneficiarios
Data Warehouse	Integrantes del proyecto Administrador del sistema de información
Resumen del proceso de la metodología CRISP-DM aplicado.	Integrantes del proyecto
Sesiones de usuario de la plataforma "Oferto" caracterizadas.	Cámara de Comercio de Armenia y del Quindío Empresas de sectores integrados a la plataforma "Oferto"
Resultado de la validación de las sesiones caracterizadas	Administrador del sistema de información. Cámara de Comercio de Armenia y del Quindío
Patrones de comportamiento de los visitantes de la plataforma "Oferto".	Cámara de Comercio de Armenia y del Quindío Empresas de sectores integrados a la plataforma "Oferto"

Tabla 1: resultados esperados

2. Estrategia Metodológica

2.1 Metodología Cross-Industry Standard Process for Data Mining (CRISP DM)

El desarrollo del proyecto planteado estará basado en la aplicación de la metodología Cross-Industry Standard Process for Data Mining (CRISP DM) probada para orientar trabajos de minería de datos. A continuación, se describen cada una de las etapas, según Rodríguez (2010) e IBM SPSS Modeler (2012):

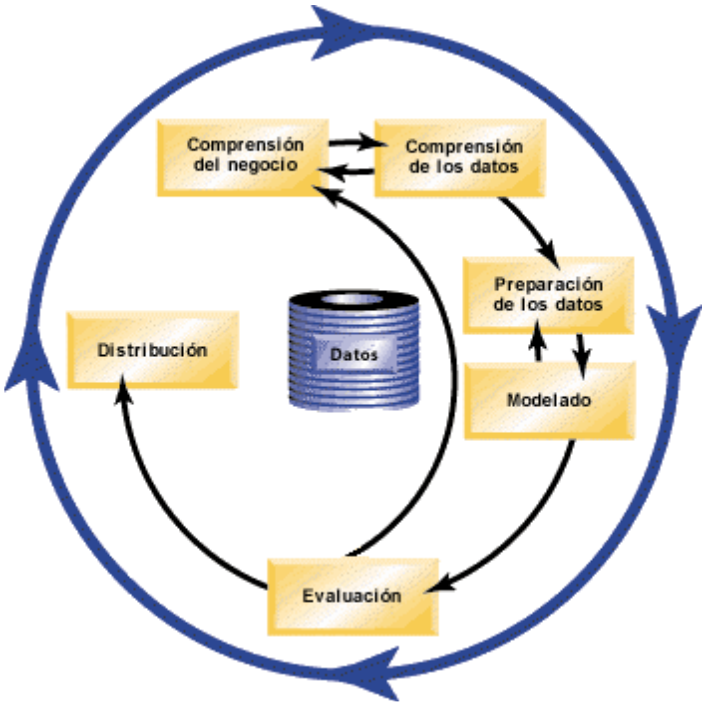


Ilustración 3: Ciclo vital CRISP-DM

Etapa 1: Comprensión del negocio

En esta etapa se busca comprender la situación actual de la organización, con el fin de identificar los recursos, problemas y objetivos que tiene con respecto al proceso a ejecutar, buscando de esta manera determinar la estructura de la organización, describir el área problemática, identificar cómo se hace actualmente el proceso y poder definir adecuadamente la información con la que se

va a trabajar, las restricciones con las que se cuenta y todo aquello que intervenga en el proceso de minería. Para lograr llevar esto a cabo es necesario realizar el inventario de los recursos con los que se cuenta, identificar los requisitos, supuestos y restricciones que se tienen dentro de la organización, hacer un proceso de identificación de riesgos y contingencias, establecer la terminología a usar dentro del proceso con el fin de evitar malos entendidos con expresiones o palabras que pueden tener diferente significado para ambas partes y finalmente plantear un plan de ejecución del proyecto.

Etapa 2: Comprensión de los datos

En esta etapa se busca, principalmente, estudiar a fondo los datos con los que se va a hacer el proceso de minería, es aquí donde se debe lograr elegir acertadamente el recurso con el que se va a llevar a cabo todo el proceso, dentro de esta elección podemos tener claro que se deben ejecutar los siguientes pasos:

Acceder a los datos: se debe tener claro que pueden haber datos existentes, adquiridos y/o adicionales, dependiendo de los objetivos planteados.

Explorar los datos: esta actividad le permitirá establecer hipótesis de ser necesarias y dar forma a las tareas de transformación de datos.

Determinar la calidad de los datos: se busca identificar errores en los datos, valores perdidos u otro tipo de incoherencias que dificulten el futuro análisis de ellos. Así mismo, describir los resultados obtenidos para dejarlos documentados y usar esto durante el proceso.

Etapa 3: Preparación de los datos

Esta es la etapa en la que normalmente se ejecuta entre el 50% - 70% del tiempo, aquí se debe realizar primero una selección clara de los datos con los que se va a trabajar durante el proceso, luego de esto se debe llevar a cabo la limpieza de los mismos, con el fin de evitar tener datos que generen ruido dentro del proceso, una fusión de conjuntos y/o registros de los datos, seleccionar una muestra de un subconjunto de datos, hacer una agregación de registros, una derivación de nuevos atributos, la clasificación de los datos para el modelado, eliminación o sustitución de valores en blanco o ausentes, división en conjuntos de datos de prueba y entrenamiento

Etapa 4: Modelado

Durante esta etapa cobra valor todo lo hecho anteriormente, los datos que se han venido preparando se utilizan con el fin de plantear el modelo adecuado, para esto se debe:

- Generar un diseño de comprobación
- Generar varios modelos
- Llevar a cabo la configuración de los parámetros

Para después de esto llevarlos a una de las herramientas analíticas con el fin de empezar a obtener resultados que permitan dar respuesta a los problemas planteados, este modelado se suele ejecutar en múltiples iteraciones, con el fin de poder comparar resultados y establecer resultados óptimos, haciendo un seguimiento de los parámetros establecidos.

Etapa 5: Evaluación

En esta etapa se formalizará la evaluación en función de si los resultados cumplen con los criterios establecidos en un comienzo, en esta etapa se debe tener totalmente claro qué era lo que se buscaba con el proceso, para, de esta manera, establecer criterios de toma de decisiones con

respecto a si los resultados están expresados con claridad, si hay información especial dentro de ellos, si es posible aplicar dichos resultados en la organización y de qué manera hacerlo.

Etapa 6: Distribución

Es la etapa donde se busca dar a conocer a la Organización y a sus interesados el conocimiento adquirido del proceso que se llevó a cabo y cómo implementarlo dentro de ella, además de la entrega de un informe final con el proceso que se realizó y los resultados obtenidos del mismo.

2.2 Cronograma

2.2.1 Fases y Actividades

Actividades Tesis			
Comprensión del negocio	Determinación de objetivos de la empresa		2 Sem.
	Valoración de la situación actual		
	Inventario de recursos		
	Identificación de requisitos, supuestos y restricciones		
	Identificación de riesgos y contingencias		
	Terminología a usar		
	Determinación de objetivos de la minería de datos		
	Producción de un plan de proyecto		
Comprensión de los datos	Recopilación de datos iniciales		1 Mes
	Descripción de los datos		
	Exploración de los datos		
	Verificación de la calidad de los datos		
Preparación de datos	Selección de los datos	Inclusión y exclusión de datos	1 Mes
	Limpieza de datos		
	Construcción de nuevos datos	Derivación de atributos	
	Integración de datos	Tareas de integración	
	Formato de datos		
Modelado	Selección de técnica de modelado	Modelado de supuestos	1 mes y medio

	Generación de diseño de comprobación		
	Generación de modelos		
	Configuración de parámetros		
	Ejecución de modelos		
	Descripción del modelo		
	Evaluación del modelo	Seguimiento de los parámetros revisados	
Evaluación	Evaluación de los resultados		2 Sem.
	Proceso de revisión		
Informe final	Elaboración de informe final		2 Sem.
	Revisión del informe final		

Tabla 2: fases y actividades

2.2.2 Distribución de las actividades

CRONOGRAMA WEBMINING OFERTO																							
Nombre del proyecto Búsqueda de patrones en el comportamiento de los visitantes de la plataforma "Oferto" de la Cámara de Comercio de Armenia y del Quindío aplicando minería web.																							
Integrantes	David Alberto Angarita García (DAAG) - Juan Jose Muñoz Franco (JJMF)																						
Fecha Inicio																							
Fecha Finalización																							
Mes	Mes 1				Mes 2				Mes 3				Mes 4				Mes 5				Mes 6		
Fase - responsable\Semana	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Comprensión del negocio - DAAG																								
Comprensión de los datos - DAAG																								
Preparación de datos - JJFM																								
Modelado - JJFM																								
Evaluación DAAG - JJFM																								
Elaboración Informe Final DAAG - JJFM																								
Socialización resultados DAAG - JJFM																								

Tabla 3: distribución de las actividades

3. Desarrollo

3.1 Referente Teórico

3.1.1 La WEB o *World Wide Web*

En 1990 Berner – Lee y Cailliau crearon un sistema de documento de hipertexto o hipermedios enlazados que eran accesibles a través de Internet; esto se constituyó como la primera versión de

la *World Wide Web* o WWW. Cuando un texto en la pantalla de un computador conduce a un usuario o a otro texto relacionado, se le conoce como Hipertexto. Su forma más habitual es en documentos que contienen hipervínculos o referencias cruzadas automáticas que van a otras secciones del texto o a otros documentos.

Al combinarse el hipervínculo con una red de datos y un protocolo de acceso, se pueden referenciar los recursos de distintas formas, ya sea como parte de un documento o visitarlo con un programa de navegación. Los hipervínculos hacen parte fundamental de la arquitectura de la WWW, pero estos no se limitan al HTML o la Web. Cualquier medio o documento electrónico puede emplear un hipervínculo o una variación del mismo.

De la misma manera se puede definir *Hipermedia* como un conjunto de métodos o procedimientos para escribir, diseñar, o componer contenidos que tengan texto, video, audio, mapas u otros medios, y que además tenga la posibilidad de interactuar con los usuarios. Mientras que *hipermedios* se refiere a la conexión entre documentos de diversos tipos de medios. Parte de la arquitectura de la Web se basa en el protocolo de transporte de hipertexto (*hypertext transport protocol -http*) y el uso de buscadores para invocar dicho hipertexto. Existe una confusión al creer que Internet y web son sinónimos; Internet es la red de redes donde reside toda la información, mientras que la web es un subconjunto que solo abarca las páginas que son accesibles desde un navegador. Bajo la tutela del Internet están los protocolos, juegos, correos electrónicos, etc. Pero estos elementos no hacen parte de la Web.

Los buscadores Web tienen un papel importante en lo que es conocido como “navegar por la Web*”*, ya que ellos son los encargados de buscar el hipertexto y organizarlo para que los usuarios puedan acceder a ellos de manera fácil. Los buscadores recuperan información de los servidores

Web en forma de documentos o páginas Web, de esta forma se encuentran los enlaces y el usuario está en la capacidad de navegar.

La Web está en constante evolución y su popularidad está en aumento, provocando que sea el medio preferido para publicar información y gracias al desarrollo de nuevos protocolos sirve como puente para que el comercio electrónico llegue a sus clientes.

La gran difusión de contenidos, la posibilidad de relacionar documentos de diversas fuentes y la combinación de diversas maneras de presentar la información, se constituyen como los puntos fuertes del uso de la Web como medio de comunicación.

Desde el inicio de la Web, se ha tratado de incluir información adicional en la Web para describir y darle significado a la relación de los enlaces y los datos. Esta nueva información se conoce como Web Semántica y hace que sea posible la evaluación automática por parte de las maquinas, para lograr dicho propósito esta información debe ser escrito en un lenguaje formal. El objetivo general de la Web Semántica es mejorar la interoperabilidad entre sistemas reduciendo la mediación humana en los procesos.

Debido a la interacción entre servidores y buscadores Web, existe una gran cantidad de información disponible acerca del comportamiento de los usuarios. Esta información es relevante y útil a la hora de extraer conocimiento, pero dicha tarea no es fácil debido a que muchos de los datos son semiestructurados o no estructurados. Adicionalmente, se debe tener en cuenta que la información puede estar dispersa a través de varios servidores o que las paginas contengan información multimedia que no puede ser interpretada por las maquinas.

Teniendo como base las dificultades escritas, según Kosala, R. and Blockeel, H (2000) la minería Web se clasifica de la siguiente forma:

Minería del contenido: tiene por objeto encontrar patrones de los datos de las páginas web.

Minería de la estructura: su objeto es descubrir el modelo subyacente a la estructura de enlaces de la web.

Minería del uso: consiste en la aplicación de técnicas de minería de datos a los registros de uso de repositorios de datos de la web.

3.1.2 Funcionamiento de la WEB

El proceso comienza cuando se ingresa una URL en el navegador Web o se está siguiente un enlace de hipertexto desde otra página. Posterior a la petición del navegador, se inician una serie de comunicaciones que no son perceptibles para el usuario, dicha comunicación trae los datos necesarios para visualizar la página. Cuando la URL llega al navegador, se debe traducir la URL a una IP mediante una tabla llamada DNS. La IP se convierte en la dirección del servidor al que se debe acudir para conseguir los datos de la página, a continuación se realiza una petición HTTP solicitando los recursos que el usuario está solicitando, con dichos recursos que llegan del servidor Web, el navegador renderiza la página según este descrito en el código HTML y CSS. Como paso final se incluyen imágenes y otros recursos para darle el aspecto definitivo a la página Web.

3.1.3 Caché de páginas WEB

Cuando los navegadores van a realizar una petición HTTP, primero deben revisar si la página ha sido cargada con anterioridad y si la página no ha cambiado desde la última carga, si esto sucede,

el navegador cargara esta versión que esta almacenada en el disco duro del computador. A esto se le llama Caché de la página Web y especialmente útil para reducir el tráfico en internet.
Par aprovechar al máximo este sistema, los diseñadores reúnen todo el código CSS y JavaScript en archivos asociados al sitio Web, de esta forma el navegador utilizará al máximo el Caché del disco duro y reducirá las peticiones realizadas al servidor. (World Wide Web: 2016)

3.1.4 Necesidad del Data Mining Y Web Mining

Actualmente se cuenta con grandes bases de datos y una variedad de información que se encuentra al interior de las organizaciones. Dicha información representa la historia de la empresa, representa el cuerpo de conocimiento y en últimas son las transacciones o situaciones que se han producido. Una práctica común al interior de las organizaciones, es tomar decisiones basadas en experiencias pasadas o en lo históricos de la empresa, pero el gran volumen de información y la falta de estructura de los datos dificulta dicho análisis. Es de suma importancia analizar los datos con el fin de conocer el pasado, entender el presente y predecir el futuro. Con el fin de facilitar el proceso de toma de decisiones, los datos deben ser guardados en repositorios que tenga una estructura de búsqueda y uso unificado, estos repositorios reciben el nombre de "*Data warehouse*" o "Almacenes de Datos".

La manera tradicional de analizar la información existente de las empresas, era mediante consultas a las bases de datos mediante lenguajes generalistas como el SQL, en donde se producía sobre la base de datos operacional de las aplicaciones de gestión. Esta forma de

consultas es supremamente rígida ya que solo permite realizar consultas previamente establecidas y no permite que las consultas sean escalables.

Un paso adelante se encuentran las herramientas OLAP, dichas herramientas si bien permiten el análisis de datos cruzados o agregados, no permiten la identificación de reglas o de patrones que son útiles a la hora de tomar decisiones al interior de la empresa. Se debe tener presente que el valor de los datos se encuentra en el conocimiento que podemos extraer de ellos.

Cuando las bases de datos contienen datos que son nominales, textuales, multimedia o datos que no se integran bien con los sistemas de información transaccional, los paquetes estadísticos pierden funcionalidad y usabilidad. Esta situación también se puede presentar cuando las bases de datos están compuestas por tablas con millones de registros, que presentan una alta dimensionalidad o que no tiene un formato común de consulta.

Los resultados esperados de un proceso de minería de datos son un conjunto de reglas, ecuaciones, árboles de decisión, redes neuronales, grafos probabilísticas que pueden ayudar a la toma de decisiones.

La minería de datos (DM o *Data Mining*) es parte de un proceso más grande que se dedica a la extracción de conocimiento a partir de los datos. Dicho proceso es llamado “Descubrimiento de Conocimiento en Bases de Datos” o KDD por sus siglas en inglés (*Knowledge Discovery from Databases).* El poder de la Minería de Datos radica en la integración de las ventajas de la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, usando como materia prima las bases de datos (Molina: 1998). El proceso de Minería de Datos está siendo muy usado en este momento para identificar patrones de comportamiento dentro de la WWW (Han & Kamber: 2006).

Actualmente en la WWW se puede encontrar una gran cantidad de información a disposición de todo el mundo, pero es precisamente dicha cantidad la que presenta un problema a la hora de realizar análisis. En la mayoría de los casos resulta imposible analizar los resultados de una búsqueda como las que realizan los buscadores Web de manera manual. Suponiendo que desea analizar los resultados de buscar la palabra "Minería" en Google, el resultado de dicha búsqueda arroja alrededor de 13.200.000 enlaces. Ante la gran dificultad de realizar este tipo de análisis, la Minería de Datos toma ventaja de los datos de navegación que dejan todas las operaciones de consulta de los usuarios, dichos datos están compuestos por IP, tipo de navegador, tiempo de acceso, páginas visitadas y sistema operativo. Este tipo de análisis hace referencia a una división de la Minería de Datos llamada "Minería Web" o "Web *Mining*" (Kosala & Blockeel: 2000), que consiste en aplicar los conceptos de minería a los servicios Web (Srivastava: 1999).
Algunas herramientas para *data mining* son las siguientes: WEKA (Hernández Orallo & Ferri Ramírez: 2005), ADaM (Adam *Documentation*: 2015), Orange (Demsar, Zupan, Leban et al: 2014), TariyKDD (Calderon, A.:2007), SPSS (IBM Corporation:2011), Webalizer (Webalizer, :2015), DBMiner (Han, Fun, Wang et al: 1996), *RapidMiner* (RapidMiner: 2015), DB2 *Intelligent Miner WEKA, SAS Enterprise Miner, STATISTICA Data Miner WEKA* Hernández Orallo & Ferri Ramírez: 2005)

3.1.5 La Minería de Datos (Data Mining)

Teniendo en cuenta autores como Servente & García-Martínez (2002), Perichinsky & García-Martínez (2000: 107) se denomina "Minería de Datos" a :

> "un conjunto de herramientas y técnicas aplicadas al proceso de extraer y presentar conocimiento implícito, que anterior a el proceso era desconocido, que es potencialmente útil y comprensible por los humanos, esto basados en conjuntos de datos de gran magnitud, con la finalidad de identificar o predecir tendencias y comportamientos; y describir de forma automatizada modelos previamente desconocidos"

El término “Minería de Datos Inteligente” se refiere puntualmente, según Evangelos & Han, (1996) y “Michalskí et al (1998) a “la aplicación de métodos de aprendizaje automático para descubrir y enumerar patrones que se encuentren en dichos datos”, para lograr esto, se desarrollaron un alto número de métodos de análisis de datos teniendo como base de ellos la estadística.

En la medida en que se incrementaba la cantidad de información almacenada en las bases de datos, estos métodos empezaron a enfrentar problemas de eficiencia y escalabilidad y es aquí donde aparece el concepto de minería de datos. Una de las diferencias entre al análisis de datos tradicional y la minería de datos es que, según Hernández Orallo (2000) “el primero supone que las hipótesis ya están construidas y validadas contra los datos, mientras que el segundo supone que los patrones e hipótesis son automáticamente extraídos de los datos”

En dos trabajos más recientes es definida como un proceso que permite, a través del uso de herramientas especializadas, el procesamiento de grandes volúmenes de datos posibilitando, analizar y descubrir el conocimiento a partir de estos. De acuerdo con De Bie (2011) en la gran mayoría de las organizaciones actuales, en las que existe una gran diversidad y dispersión de datos y sus tamaños son considerables, esta tarea no podría ser eficaz sin el uso de la minería de datos, por otra parte, Larose (2014) afirma que la importancia del proceso de descubrir correlaciones, patrones, tendencias significativas y nuevas examinando cuidadosamente una gran cantidad de datos almacenados en los repositorios utilizando tecnologías de reconocimiento de patrones y técnicas estadísticas y matemáticas.

3.1.6 Aplicaciones de la Minería

Desde hace varios años se percibe un incremento en el uso de la minería de datos en todo tipo de campos en busca de la extracción de conocimiento. Hernández Orallo, Ramírez & Ferri Ramírez (2005) mencionan los siguientes campos en los que se usan las técnicas de minería.

Aplicaciones financieras y banca:

- Obtención de patrones de uso fraudulento de tarjetas de crédito.
- Determinación del gasto en tarjeta de crédito por grupos.
- Cálculo de correlaciones entre indicadores financieros.
- Identificación de reglas de mercado de valores a partir de históricos.
- Análisis de riesgos en créditos.

Análisis de mercado, distribución y, en general, comercio:

- Análisis de la cesta de la compra (compras conjuntas, secuenciales, ventas cruzadas, señuelos, etc.).
- Evaluación de campañas publicitarias.
- Análisis de la fidelidad de los clientes. Reducción de fuga.
- Segmentación de clientes.
- Estimación de stocks, de costes, de ventas, etc.

Seguros y salud privada:

- Determinación de los clientes que podrían ser potencialmente caros.
- Análisis de procedimientos médicos solicitados conjuntamente.
- Predicción de qué clientes contratan nuevas pólizas.
- Identificación de patrones de comportamiento para clientes con riesgo.

- Identificación del comportamiento fraudulento.
- Predicción de los clientes que podrían ampliar su póliza para incluir procedimientos extras (dentales, ópticos…).

Educación:

- Selección o captación de estudiantes.
- Detección de abandonos y de fracaso.
- Estimación del tiempo de estancia en la institución.

Procesos industriales:

- Extracción de modelos sobre el comportamiento de compuestos.
- Detección de piezas con trabas. Modelos de calidad.
- Predicción de fallos y accidentes.
- Estimación de composiciones óptimas en mezclas.
- Extracción de modelos de costo.
- Extracción de modelos de producción.

Medicina:

- Identificación de patologías. Diagnóstico de enfermedades.
- Detección de pacientes con riesgo a sufrir una patología concreta.
- Gestión hospitalaria y asistencial. Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones.
- Recomendación priorizada de fármacos para una misma patología.

Biología, bioingeniería y otras ciencias:

- Análisis de secuencias de genes.
- Análisis de secuencias de patrones.

- Predecir si un compuesto químico causa cáncer.
- Clasificación de cuerpos celestes.
- Predicción de recorrido y distribución de inundaciones.
- Modelos de calidad de aguas, indicadores ecológicos.

Telecomunicaciones:

- Establecimiento de patrones de llamadas.
- Modelos de carga en redes.
- Detección de fraude.

Otras áreas:

- Correo electrónico y agendas personales.
- Clasificación y distribución automática de correo, detección de correo spam, gestión de avisos, análisis del empleo del tiempo.
- Recursos humanos: selección de empleados.
- Web: análisis del comportamiento de los usuarios, detección de fraude en el comercio electrónico, análisis de los *log* de un servidor web.
- Turismo: determinar las características socioeconómicas de los turistas en un determinado destino o paquete turístico, identificar patrones de reservas, etc.
- Tráfico: modelos de tráfico a partir de fuentes diversas: cámaras, GPS.
- Hacienda: detección de evasión fiscal
- Policiales: identificación de posibles terroristas en un aeropuerto
- Deportes: estudio de la influencia de jugadores y de cambios. Planificación de eventos.
- Política: diseño de campañas políticas, estudios de tendencias de grupos, etc.

3.1.7 Técnicas de la Minería de Datos

En cuanto a las técnicas de minería de datos se encuentran múltiples clasificaciones como la mostrada a continuación:

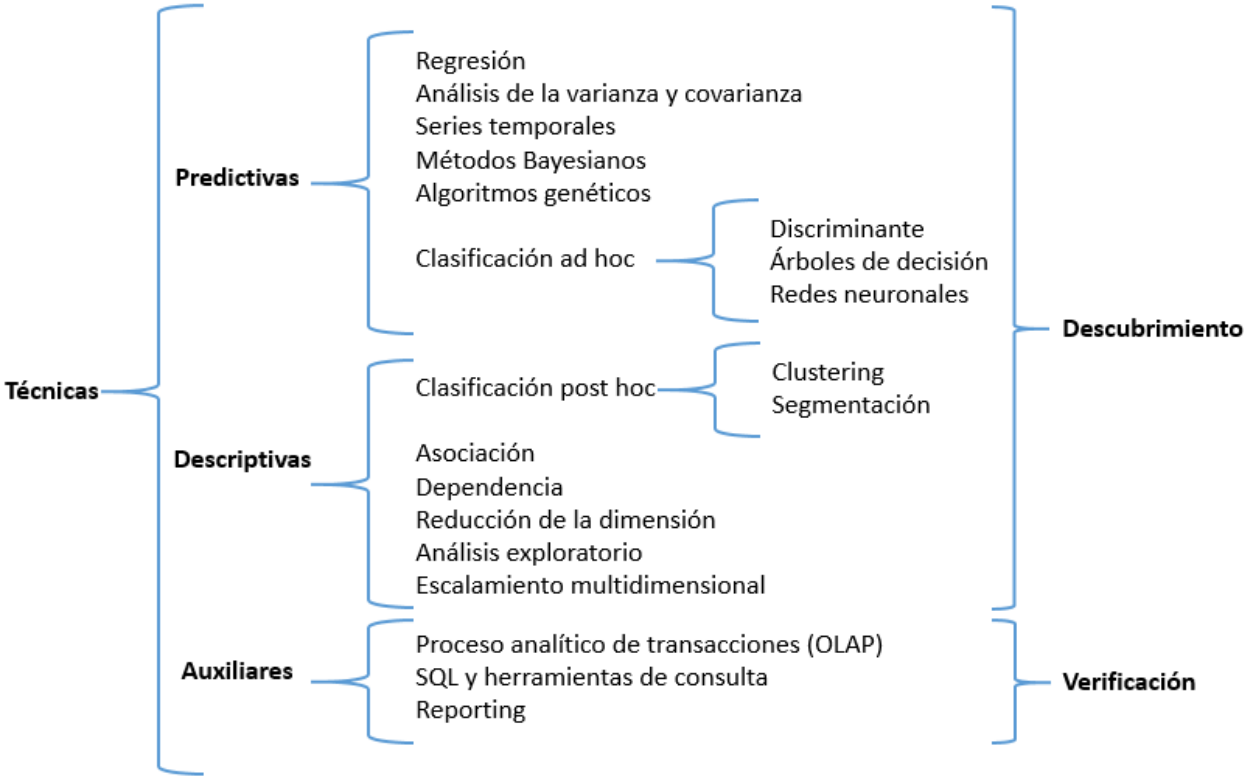


Ilustración 4: Clasificación de técnicas de minería de datos
Fuente: Minería de datos conceptos, técnicas y sistemas Pág. 9.

Según Sosa & Sosa (2014) la primera técnica denominada "predictivas" se refiere a aquellas en las que las variables pueden clasificarse inicialmente en dependientes e independientes con base en un conocimiento teórico previo. Algunos algoritmos son los de tipo de regresión, árboles de decisión, redes neuronales, algoritmos genéticos y técnicas bayesianas. Por otra parte, las "Las descriptivas" son aquellas en las que todas las variables tienen al inicio el mismo nivel o grado de pertenencia. Se crean automáticamente iniciando del reconocimiento de patrones. En este grupo

podemos encontrar técnicas de segmentación, agrupación (clustering), reducción de la dimensionalidad, entre otras. Por último, “las auxiliares” son más limitadas y usadas de apoyo superficial. Se basan en técnicas de estadística descriptiva, consultas e informes dirigidas generalmente a la presentación y verificación.

Se toma de García & Molina (2012) algunas de las técnicas existentes más representativas como lo son:

Clustering. (“Segmentación”): también llamada agrupamiento, permite la identificación de tipologías o grupos donde los elementos guardan gran similitud entre sí y muchas diferencias con los de otros grupos. Así se puede segmentar el colectivo de clientes, el conjunto de valores e índices financieros, el espectro de observaciones astronómicas, el conjunto de zonas forestales, el conjunto de empleados y de sucursales u oficinas, etc. La segmentación está teniendo mucho interés desde hace ya tiempo dadas las importantes ventajas que aporta al permitir el tratamiento de grandes colectivos de forma pseudoparticularizada, en el más idóneo punto de equilibrio entre el tratamiento individualizado y aquel totalmente masificado, vale la pena mencionar que García & Molina (2012) menciona tres algoritmos relevantes como lo son *clustering* numérico (K-MEDIAS), *clustering* conceptual (COBWEB) y *clustering* probabilístico (EM).

Reglas de Asociación: tste tipo de técnicas se emplea para establecer las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes; pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Son utilizadas cuando el objetivo es realizar análisis exploratorios, buscando relaciones dentro del conjunto de datos. Las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos. Debido a sus

características, estas técnicas tienen una gran aplicación práctica en muchos campos como, por ejemplo, el comercial ya que son especialmente interesantes a la hora de comprender los hábitos de compra de los clientes y constituyen un pilar básico en la concepción de las ofertas y ventas cruzada, así como del "*merchandising*". En otros entornos como el sanitario, estas herramientas se emplean para identificar factores de riesgo en la aparición o complicación de enfermedades. Para su utilización es necesario disponer de información de cada uno de los sucesos llevados a cabo por un mismo individuo o cliente en un determinado período temporal. Por lo general esta forma de extracción de conocimiento se fundamenta en técnicas estadísticas, como los análisis de correlación y de variación. Uno de los algoritmos más utilizado es el **algoritmo *A priori***, que se presenta a continuación.

La predicción: Este proceso busca determinar los valores de una o varias variables, partiendo de un conjunto de datos. La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión. Por ejemplo, para predecir el sueldo de un graduado de la universidad con 10 años de experiencia de trabajo, o las ventas potenciales de un nuevo producto dado su precio. Se pueden resolver muchos problemas por medio de la regresión lineal, y puede conseguirse todavía más aplicando las transformaciones a las variables para que un problema no lineal pueda convertirse a uno lineal.

Regresión no lineal: en muchas ocasiones los datos no muestran una dependencia lineal. Esto es lo que sucede si, por ejemplo, la variable respuesta depende de las variables independientes según una función polinómica, dando lugar a una regresión polinómica que puede planearse agregando las condiciones polinómicas al modelo lineal básico. De esta forma y aplicando ciertas transformaciones a las variables, se puede convertir el modelo no lineal en uno lineal que puede resolverse entonces por el método de mínimos cuadrados.

Árboles de Decisión: es una metodología del aprendizaje supervisado. La representación que se utiliza para las descripciones del concepto adquirido es el árbol de decisión, que consiste en una representación del conocimiento relativamente simple y que es una de las causas por la que los procedimientos utilizados en su aprendizaje son más sencillos que los de sistemas que utilizan lenguajes de representación más potentes, como redes semánticas, representaciones en lógica de primer orden etc.

Árboles de Predicción: Los árboles de predicción numérica son similares a los árboles de decisión mencionados anteriormente, excepto en que la clase a predecir es continua. En este caso, cada nodo hoja almacena un valor de clase consistente en la media de las instancias que se clasifican con esa hoja, en cuyo caso estamos hablando de un árbol de regresión, o bien un modelo lineal que predice el valor de la clase, y se habla de árbol de modelos.

Redes de neuronas: buscan detectar y aprender complejos patrones y características dentro de los datos. Aprenden de la experiencia y del pasado, y aplicando tal conocimiento a la resolución de problemas nuevos. Este aprendizaje se obtiene como resultado del adiestramiento (*training*) y este permite la sencillez y la potencia de adaptación y evolución ante una realidad cambiante y muy dinámica.

3.1.8 Minería Web (Web Mining)

Debido al crecimiento acelerado que ha tenido la WWW, la cantidad de información y de datos disponibles en ella ha aumentado a un rimo exponencial desde su aparición en 1990. El objetivo inicial de la WWW era en palabras de su creador Berners-Lee el principal objetivo de la web fue "El concepto de la Web integró muchos sistemas de información diferentes, por medio de la formación de un espacio imaginario abstracto en el cual las diferencias entre ellos no existían. La Web tenía que incluir toda la información de cualquier tipo en cualquier sistema."(11:2008).

Para lograr la comunicación entre maquinas Berners-Lee desarrolló un identificador único que sirve para referirse a cualquier tipo de recurso dentro de la Web llamado URL (*Uniform Resource Locutor*).

El FTP es un protocolo de red que permite el intercambio de archivos entre maquinas basado en la arquitectura Cliente – Servidor. Desde un cliente se puede realizar una conexión para recibir o enviar archivos sin importar el sistema operativo que el cliente este usando. Basado en el protocolo FTP fue creado el protocolo HTTP ((*Hypertext Transfer Protocol)* para la transferencia de Hipertexto. Si bien el Data Mining y la Web Mining tiene similitudes, existen grandes diferencias que las separa a la hora de realizar el análisis de los datos. Esto se debe a que la Web es poco estructurada por naturaleza a diferencia de las bases de datos relaciones. Esto provoca que las técnicas de la minería de datos no se puedan aplicar directamente sino que deban modificarse para superar el problema de estructura. Algunos otros problemas que presentan los datos en la Web son los siguientes:

- La falta de contexto en la información y en las bases de datos.
- Separar la información relevante de que no lo es.
- Sobrecarga de información.

Es por esta razón que, según Salton, G. y McGill, M. J., (1983), la minería Web toma elementos como la recuperación de información teniendo en cuenta el procesamiento del lenguaje natural (*Natural Language Processing - NLP*) la inteligencia artificial y el aprendizaje automático, que son propios de otras áreas de investigación.

Actualmente existen herramientas que permiten que los datos de navegación del usuario que están guardados en la cookies o el log de servidor, sean más entendibles transformándolos en reportes o resúmenes. Existe una confusión acerca del concepto de “Minería Web”, debido a que

muchas herramientas proporcionan vistas, resúmenes y elementos estadísticos que pueden ser de gran importancia a la hora de administrar plataformas Web, pero que se quedan cortas en el momento de responder preguntas más complejas o en el momento de extraer patrones de comportamiento de los usuarios.

Las verdaderas herramientas de Minería Web, deberían proporcionar información oculta a simple vista. Deberían poder responder preguntas como ¿cuáles serían los visitantes más adecuados para una nueva línea de productos?, ¿cuál es el perfil de los visitantes de una página determinada?, ¿qué organización del portal favorece las compras?, ¿qué páginas web fomentan el abandono del sitio web?

Según Molina (2002) algunos resultados que se pueden obtener con la Minería Web son, entre otros, que el 85% de los clientes que accede a la página:

- /productos/home.html y a /productos/noticias.html
- también acceden a /productos/historias_suceso.html.

Esto podría indicar que existe alguna noticia interesante de la empresa que hace que los clientes se dirijan a historias de suceso. Este resultado permitiría detectar la noticia sobresaliente y colocarla quizá en la página principal de la empresa.

Los clientes que hacen compras en línea cada semana en la página /compra/producto1.html tienden a ser de sectores del gobierno. Esto permitiría proponer diversas ofertas a este sector con el fin de potenciar sus compras.

El 60% de los clientes que hicieron una compra en línea en /compra/producto1.html también compró en /compra/producto4.html después de un mes. Esto indica que se podría recomendar en la página del producto 1 comprar el producto 4 y ahorrarse el costo de envío de este producto.

3.1.8.1 Definición de Web *Mining*

En términos generales la definición de *Web Mining* es muy similar a la del proceso de KDD, con la diferencia que la fuente de datos es la Web. De acuerdo con Scotto (2004) la *Web Mining* "tiene como propósito descubrir información oculta en los datos que produce la Web" (1-3). Debido a la naturaleza de la información en la Web, esta toma elementos de procesos de recuperación de la información como lo son la *Information Retrieval – IR* y de la *Informarion Extraction – IE.*

Gran parte de la Web está compuesta por texto, imagen, vídeo, metadatos o hiperenlaces, dando lugar a investigaciones que usan el término Multimedia *Data Mining* como una subdivisión de la Web *Mining* para tratar estos datos en específico.

3.1.8.2 Áreas o categorías de Web *Mining*

Como lo explican (Baeza-Yates, R., Poblete, B. 2005) en el caso de la minería web los datos con los que se va a trabajar pueden ser obtenidos desde varios lados, como los son el cliente, los servidores proxy, la base de datos corporativa de la empresa a la cual pertenece el sitio o del cliente. Partiendo de esta premisa, los datos obtenidos de un sitio web se pueden clasificar en tres tipos principalmente:

Minería Web de Estructura: Donde los datos describen la organización del contenido del sitio web en su interior. Dentro de lo que se incluye la forma en la que están distribuidos los enlaces tanto internos como externos al sitio, y la estructura jerárquica del mismo.

Minería Web de Contenido: Son todos aquellos datos reales que se le dan al usuario. Esto quiere decir, datos que se almacenan en los sitios web, que en su mayoría son textos e imágenes entre otros. Este tipo de datos son importantes pero tienen un gran nivel de dificultad en su procesamiento, debido a que son multimedial.

Minería Web de Uso: Estos datos son aquellos que describen el uso que le dan los visitantes a un sitio, el cual queda registrado en los Log de acceso de los servidores en los que está el sitio web.

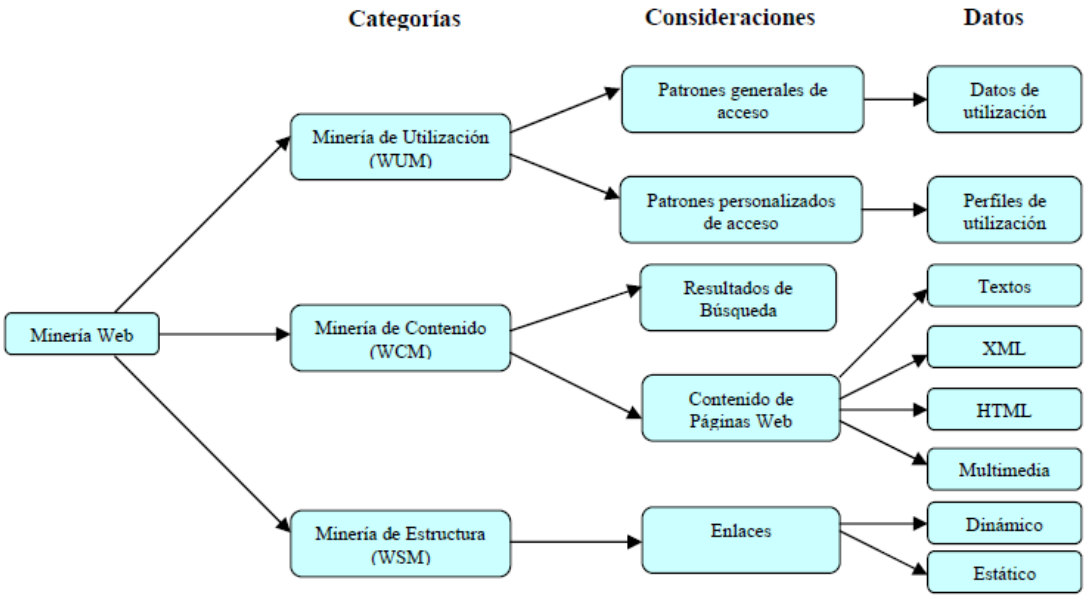


Ilustración 5: Categorías de minería web
Fuente: Dürsteler, J.C (2005)

3.1.8.2.1 Minería Web de Estructura

Representa el grado de dificultad que tienen los visitantes de un sitio web para encontrar la información, busca determinar si la estructura del sitio es simple o muy profunda, si los elementos están situados adecuadamente dentro de la página, si la navegación es perceptible,

cuáles son los lugares menos visitadas y busca establecer una relación con el espacio que ocupan en la página central.

Según Kosala, R. and Blockeel, H. (2000) este modelo se puede usar con el fin de dar una categorización a los sitios web, además de permitir generar el grado de similitud que hay entre sitios web diferentes.

3.1.8.2.2 Minería Web de Contenido

Está centrada en el contenido, buscando que la forma de escribir el contenido del sitio web sea más adecuado para los visitantes, teniendo como parte de su finalidad identificar si los temas que se tratan en el mismo son de interés o no, además para mejorar la relevancia del sitio web. Esta área de la minería se basa en la recuperación de la información a través de la exploración semántica de los documentos, usualmente a través de minería de textos y el análisis semántico de los mismos, sin embargo este proceso también puede ser basado en documentos semiestructurados.

3.1.8.2.3 Minería Web de Uso

Su fin es recoger datos y a partir de estos identificar patrones concernientes a los contenidos del sitio web y a las búsquedas que realizan los visitantes en ellos, buscando con ello mejorar la navegación dentro del sitio web, donde inicialmente se busca establecer los objetivos desde el punto de vista del negocio, para luego reunir los datos que se tendrán en cuenta para el proceso y análisis posterior, los cuales pueden ser obtenidos del Log del servidor, datos de facturación, de marketing, del cliente entre otros. Luego de tener dichos datos se deberá realizar el proceso de limpieza y selección de aquellos con los que se busca identificar las sesiones de usuario, reconstruyendo la secuencia de páginas a las que accedió el visitante a través de la información

que queda almacenada en el Log. Todo lo anterior permite establecer estrategias para mejorar el diseño de la colección de recursos existentes.

Se debe tener en cuenta que una sesión puede ser definida como todas aquellas páginas consultadas por un mismo visitante durante una sola visita al sitio. Al final de esta fase se tendrán ficheros sobre los que se aplicarán las herramientas que permitirán extraer la información.

Luego de esto se iniciará la búsqueda de patrones de comportamiento de los visitantes, para esto es necesario aplicar las técnicas y métodos adecuados de acuerdo al problema que se busca resolver y de aquellos datos con los que se cuenta.

3.1.8.3 El proceso de Web Mining

Según Kosala y Blockeel (2000), la *Web Mining* se compone de las siguientes fases:

Descubrimiento de las fuentes

Durante esta fase se recuperan datos de las fuentes textuales de la Web, como son los correos, grupos de noticias, RSS, documentos de HTML o documentos de hipertexto. Esta búsqueda se basa en la creación de índices de documentos basados en palabras claves definidas y priorizadas en función de ciertos criterios de relevancia.

Selección y pre-procesado de la información

Durante esta etapa, los datos se seleccionan y se transforman con el fin de eliminar cualquier dato no relevante y de obtener una representación más entendible para los analistas.

Generalización

Durante esta fase se realiza el proceso de *Web Mining* en sí, su intención es descubrir información oculta útil a partir de los datos seleccionados en la fase anterior. Para realizar este proceso, la Web Mining toma elementos de IR (*Information Retrieval*) algunas técnicas para la categorización y la clasificación de textos; y ha desarrollado algunas técnicas propias, como por ejemplo el análisis de caminos (*web paths*) usado para extraer secuencias de patrones de navegación desde archivos log.

Análisis: validación e interpretación de los patrones minados

Esta fase se ocupa de presentar la información minada de manera entendible para los interesados en la información. Para dicho propósito, se emplean técnicas de representación estadística para facilitar su análisis y contrastar el conocimiento minado con el conocimiento que se tenía anteriormente. En esta parte del proceso se socializarán y analizarán los resultados a través de gráficas y documentación con aquellas personas que viven el día a día interactuando con la plataforma y los empresarios, buscando así determinar la relación entre la realidad y lo obtenido luego de la aplicación de los procesos de minería, para finalmente obtener conclusiones y posibles puntos de partidas sólidos para mejoras con respecto a este proceso.

3.1.9 Comercio Electronico (E- Commerce)

El comercio electrónico es “cualquier transacción en la que las partes interactúan electrónicamente en vez de tener un contacto físico”. Sin embargo esta definición se queda corta a la hora de explicar el fenómeno que ha sido el comercio electrónico en los últimos años. Las

estructuras organizacionales modernas han derribado sus fronteras con los clientes y con los proveedores, de tal forma que los procesos empresariales ahora son transversales a estos tres actores.

Este cambio se ha dado en parte a las nuevas formas de hacer negocio, entre esas el comercio electrónico. El comercio electrónico habilita el intercambio y soporte en una escala global. Habilita a las compañías para llegar a un mayor número de clientes sin intermediaros o grandes superficies, hace a las empresas más flexibles en sus operaciones internas y les permite seleccionar los mejores proveedores sin importan su ubicación geográfica.

Algunas de los tipos de comercio electrónico son:

- *business-business* (Empresa a Empresa)
- *business-consumer* (Empresa a Consumidor)
- *business-administration* (Empresa a Gobierno)
- *consumer-administration* (Consumidor a Gobierno)

Una interacción B2B (*Business to Business*) puede demostrarse mediante una relación entre una empresa cualquiera y un proveedor intercambiando datos y realizando pagos, mientras que la categoría de B2C (*Business to Consumer*) puede ser la más grande de todas, esta categoría abarca todas las compras por internet que se pueden realizar como si se tratara de un centro comercial. Por otro lado, la categoría B2A (*Business to Administration*) encierra todas las transacciones entre las empresas y las instituciones de gobierno. Esta categoría es relativamente nueva, pero se ha expandido rápidamente y promete ser una de las grandes oportunidades de crecimiento en nuestro país.

Por último, la categoría de C2A (Consumer to Administration), es una categoría que todavía no se ha posicionado de manera significativa, pero que apalancada por la categoría B2A se ha abierto camino mediante pequeños desarrollos.

El impacto del comercio electrónico en las empresas ha sido de tal magnitud, que en algunas se ha ofrecido un punto de quiebre en la forma de hacer negocios, convirtiéndose en una de las prioridades dentro de dichas empresas. Incluso para aquellas que no tienen como una prioridad esta nueva forma de hacer negocios, les ha sido difícil ignorar que el mercado y las expectativas de los clientes ha sido modificada notablemente por el comercio electrónico.

El comercio electrónico no es una idea futurista. Está sucediendo ahora y está sucediendo rápido. (Electronic Commerce - An Introduction: 2016).

3.2 Fase I. Comprensión del negocio

La plataforma Oferto www.oferto.co nace de una iniciativa de la CCAQ, apoyada por el MINTIC "Ministerio de Tecnologías de la Información y Comunicaciones", teniendo como su principal objetivo lograr que el grupo de establecimientos comerciales beneficiarios objeto del proyecto apropiaran los componentes que comprende la implementación técnica y comercial de la herramienta de marketing móvil: App de Ofertas y Descuentos Especiales llamada Oferto, y que a partir de éstos, mejorarán los procesos de mercadeo y venta de sus productos.

Esta herramienta de conectividad busca responder a unas de las necesidades más dicientes de la mayoría de los comerciantes del Departamento, y es la de cómo lograr incrementar ventas, cómo ampliar y lograr fidelizar su portafolio de clientes.

Pues si bien ellos realizan actividades de promoción y publicidad, lo hacen de modo tradicional y eso les supone en esta realidad actual de dinámica de consumo menor cobertura, baja efectividad y en algunos casos mayores costos. Y ello se debe principalmente a que la mayoría de estos

comerciantes no cuentan con el suficiente conocimiento y capacidades técnicas para gestionar de un modo más pertinente sus estrategias de mercadeo para un consumidor que hoy día prácticamente todo lo consulta a través de internet. Los establecimientos de comercio tenidos en cuenta para la implementación de la plataforma fueron principalmente: Tiendas de Ropa y Accesorios, Tiendas de Marroquinería y Calzado, Joyerías, Tiendas de Cosméticos y Productos de Belleza, Salas de Belleza y Centros de Estética, Tiendas que distribuyen equipos informáticos y otros relacionados,

Escuelas de Educación No Formal, Farmacias, y Tiendas de Muebles, Electrodomésticos y demás productos para el Hogar, Restaurantes y Cafés al Paso. De los cuales su mayoría se encuentran localizados en la ciudad de Armenia, más del 80%, en particular sobre la Zona Céntrica, el Centro Comercial de Cielos Abiertos, la Avenida Bolívar, y demás zonas que concentran una oferta comercial nutrida en calidad de producto, número, variedad y precio, y en menor proporción en los 11 municipios restantes del departamento del Quindío.

Siendo una plataforma web y móvil visible que permite a compradores y clientes interesados en la oferta comercial del Quindío, encontrarla en un solo lugar. Gracias a una estrategia que nos permite promocionar, centralizar, unificar y posicionar el comercio de la región ofreciendo a compradores y comerciantes una experiencia de usuario agradable, ágil y práctica en la consulta de productos, ofertas y promociones.

A continuación se puede visualizar su composición general.

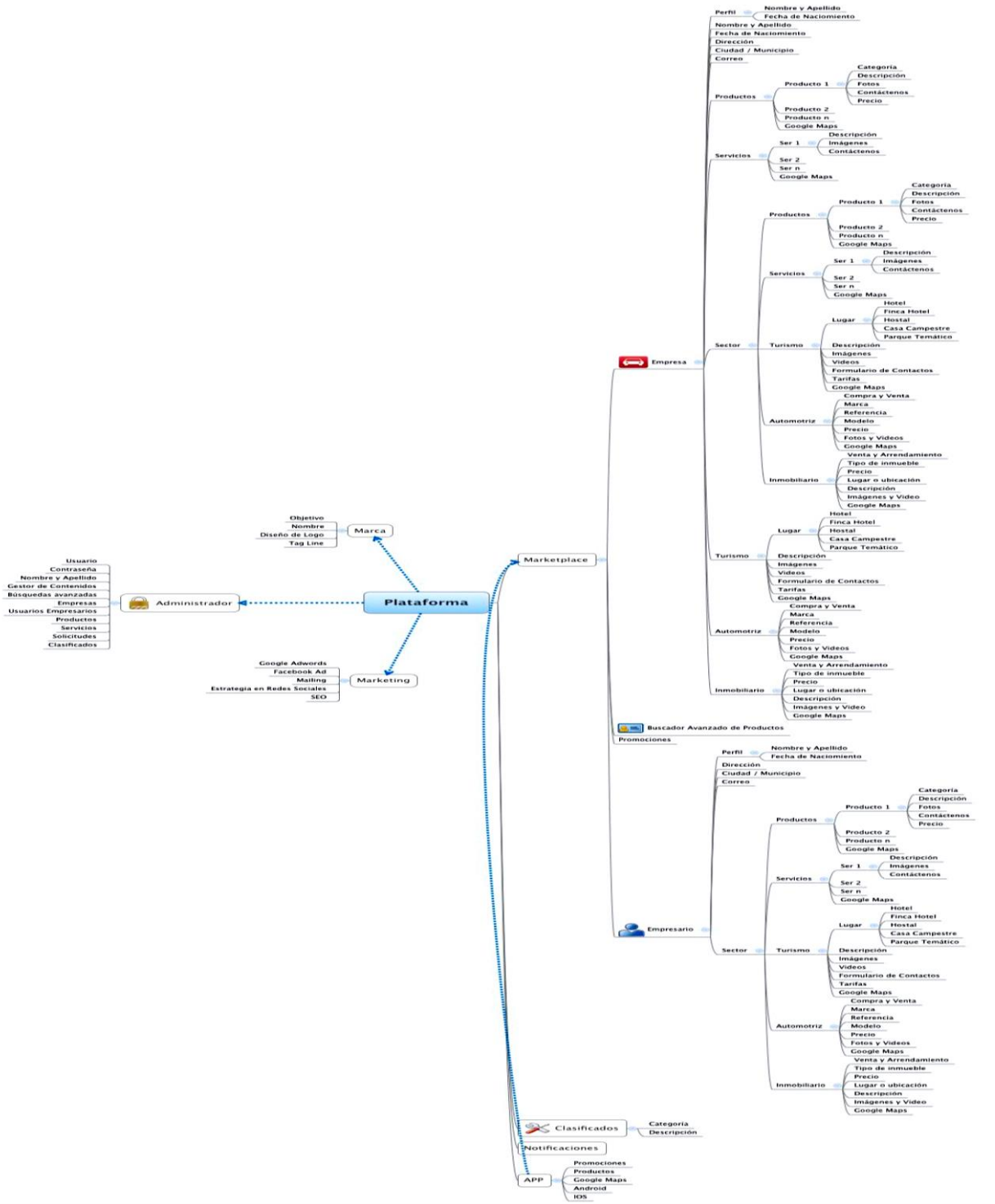


Ilustración 5: composición general plataforma oferto

Contando dentro de ella con:

Marketplace: es el portal web donde el visitante encontrará la oferta de servicios y productos de la región, dentro de él están:

- Las ofertas destacadas.
- Productos y servicios que publican las empresas.
- Comercio, turismo, inmobiliario, automotriz, servicios, etc.
- Buscador avanzado de productos, servicios y empresas.
- Consulta de promociones.
- Suscribirse a recibir promociones vía e-mail.
- Reservar productos.
- Calificar a las empresas
- Ubicación geográfica de las empresas.

Sitio web para las empresas y empresarios (minisitios): con las siguientes características:

- Administrador de contenidos.
- Categorías de las empresas. (el sistema permitirá la categorización de las empresas)
- Nombre de la empresa.
- Datos de ubicación y contacto.
- Productos y servicios.
- Promociones.
- CRM online
- Mailing a clientes.
- Subdominio
- Recibir notificaciones de contacto desde el marketplace y la app
- Datos del representante.

Ofertas y Promociones Destacadas.

- Se destacan en el marketplace y app según políticas de publicación (actualmente a consideración del web master).
- En el minisitio del empresario.
- Mailing a los clientes.
- Las promociones se podrán consultar a través de mapas.
- Si el usuario se encuentra cerca al lugar de una promoción la app le notificará en el dispositivo móvil la cercanía.
- Promociones más consultadas.

APP Móvil: Oferto cuenta con una aplicación móvil disponible en App Store, la cual cuenta con las mismas funcionalidades nombradas anteriormente en el martketplace.

3.2.1 Estructura de la organización

La CCAQ es una entidad privada sin ánimo de lucro, que promueve acciones que invitan a la formalidad y legalidad incrementando la competitividad, el fortalecimiento empresarial y el emprendimiento para el desarrollo económico del departamento del Quindío. Buscando mejoramiento del entorno, para una mejor calidad de vida a la comunidad.

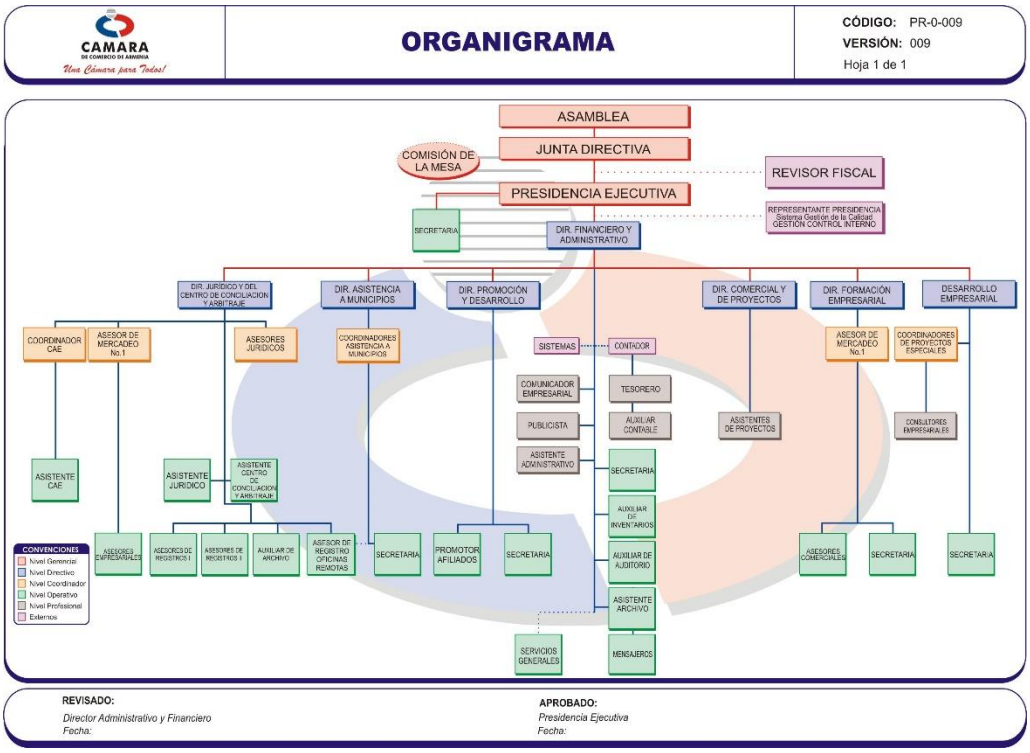


Ilustración 6: estructura de la Cámara de Comercio de Armenia y el Quindío
Fuente: página web oficial

La plataforma depende directamente del área de sistemas de la organización, la cual cuenta con un equipo de trabajo conformado por actualmente por:

Director: es la persona responsable de la dirección de la plataforma, quien toma las decisiones de las acciones que se deben llevar a cabo con la misma para alcanzar los objetivos planteados por la presidencia ejecutiva.

Profesional en mercadeo: encargado de determinar las estrategias de mercadeo que se implementan para que la plataforma tenga el impacto esperado en las empresas registradas en la misma.

Administradora de la plataforma: empresa subcontratada para el soporte, administración y actualización de la plataforma.

El equipo de trabajo mencionado anteriormente se apoya en las demás áreas de la CCAQ para su óptimo funcionamiento, dentro de estas áreas se identificaron los siguientes individuos que son claves en las propuestas que se podrían plantear al finalizar el proyecto con los resultados obtenidos como lo son:

- Presidente ejecutivo
- Directora Financiera y Administrativa
- Directora Comercial y de Proyectos

Después de aclarar la ubicación de la plataforma dentro de la CCAQ e identificar las personas claves dentro del proceso, vale la pena mencionar que está fue lanzada en septiembre del 2014 con un total de 700 empresas registradas a ella, las cuales se encuentran distribuidas en las siguientes categorías:

Categoría	Número empresas
Víveres y Abarrotes	2
Tecnología	44
Niños y Bebes	5
Mascotas	22
Moda	74
Música	3

Belleza y Cuidado Personal	5
Viajes y Turismo	4
Vehículos	28
Agro	1
Deportes	16
Gastronomía	92
Inmuebles	12
Servicios	69
Diseño, Arte y Decoración	29
Arquitectura y Construcción	2
Ferretería	15
Servicios Publicitarios	19
Fiestas y regalos	9
Hogar, Electrodomésticos y Oficina	46
Educación	11
Joyas y Accesorios	12

Seguros	7
Materiales Electricos	2
Varios	1
Juegos, Juguetes y Hobbies	1
Droguerías	12
Salud	35
Proveedores/Mayoristas	1
Servicios Empresariales	1
Institucional	1
Entretenimiento y Vida Nocturna	2
Productos Importados	1
Transporte	1
Alojamientos	38
Agencias de viaje	15
Guías de turismo	1

Tabla 4. Categorías de las empresas en CCAQ

Teniendo un total de 639 empresas en categorías, faltando 61 de ellas por elegir, lo que significa que se encuentran registradas, pero no están activas actualmente en la plataforma.

Estas empresas fueron llamadas pioneras, por lo que no deben realizar ningún tipo de pago por el uso de la plataforma y el app, además de esto se les entregaba un sitio web el cual era autoadministrable y se encontraba ligado a la plataforma, con el fin de facilitar la publicación de productos y servicios en la misma, para dar más claridad a continuación se muestra una descripción de la plataforma.

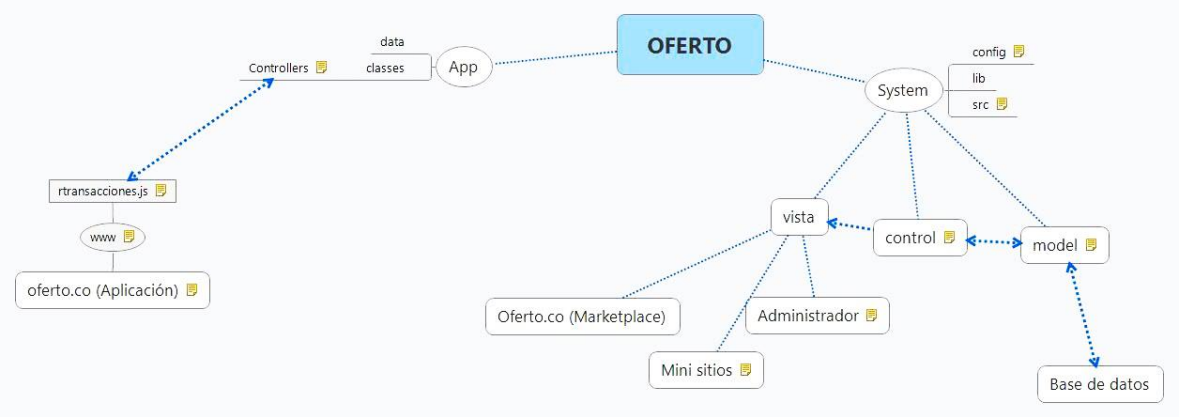


Ilustración 7: descripción de la plataforma

Actualmente, la plataforma cuenta con 1.185 empresas registradas y se busca que este número aumente, se aclara que al ser un servicio de la CCAQ para los empresarios no es vital el ingreso de recursos monetarios a través del mismo, sin embargo se debe lograr que las empresas del departamento se den a conocer a través de ella y aumenten la venta de sus productos y servicios por este medio.

3.2.2 Objetivos de la empresa

La CCAQ tiene como objetivos con la plataforma Oferto para el año 2016 los siguientes.

- Aumentar el número de visitantes de la plataforma
- Optimizar el sitio web del Marketplace de Oferto
- Aumentar las opciones de personalización de los sitios web de las empresas

- Implementar una pasarela de pagos
- Integrar la plataforma con la de Rutas del Paisaje Cultural Cafetero.

3.2.3 Valoración de la situación actual

Luego de analizar el problema, se ha evaluado la situación actual de la organización para el desarrollo del proyecto, por lo que a continuación se detallan los recursos, requisitos, supuestos y restricciones.

3.2.3.1 Personal y empresa externa

Como se identificó anteriormente, la plataforma Oferto divide su equipo de trabajo en dos entornos.

Administrativo

Conformado por el director de la plataforma que se encarga de la parte gerencial de la misma, identificando las necesidades de las empresas que hacen parte de ella y la toma de decisiones sobre el rumbo que se deberá seguir, junto con el profesional en mercadeo quien establece todas aquellas acciones a ejecutar con el fin de fortalecer el posicionamiento de la plataforma, principalmente en el departamento a través de diferentes estrategias, apoyados por las diferentes áreas de la CCAQ.

Técnico

Se encarga de dar soporte, mantenimiento y actualización de la plataforma, contando con un equipo capacitado en la gestión de servidores, bases de datos y sitios web, pero con poca experiencia en el adecuado almacenamiento y la limpieza de datos para el análisis, es importante aclarar que la plataforma fue desarrollada inicialmente por una empresa que ya no hace parte de este equipo de trabajo, lo que conlleva a algunos vacíos en cuanto a la arquitectura con la que fue creada.

Minería de datos

Encargados del proceso de minería de datos sobre los archivos de datos nombrados anteriormente.

Por todo lo anterior, se realizó una reunión con todo el equipo de trabajo, en la que se resaltó la importancia del proceso a iniciar, con el fin de contar con el conocimiento y la experiencia de cada uno de ellos en este proceso, para así llegar a obtener resultados que permitan alcanzar los logros planteados.

3.2.3.2 Datos

Como se trata de una plataforma que lleva funcionando 18 meses existe una gran cantidad de información, sin embargo el análisis de la misma es complejo, por lo que luego de reuniones con expertos y basados en la capacidad de cómputo, la cantidad de memoria disponible, el tiempo de análisis y la reserva de información por parte de la empresa administradora de la plataforma se decidió tomar una muestra de los días entre el 8 de Marzo y el 21 de Mayo del 2015.

3.2.3.3 Riesgos

Dentro del desarrollo de este proyecto se cuenta con un riesgo principal, el cual consiste en tener fallas en los datos con los que se cuenta para realizar el proceso de minería, lo que nos llevaría a aumentar las dificultades al momento de identificar patrones de comportamiento de los visitantes y por ende limitar la generación de resultados que permitan a la organización cumplir con los objetivos planteados a través del mismo, por lo que se aclaró este aspecto con los directivos de la organización. Los demás riesgos identificados podrán ser visualizados en el documento anexo de riesgos y contingencias.

3.2.4 Inventario de recursos

3.2.4.1 Recursos de hardware

Para realizar el proceso se cuenta con dos equipos de cómputo con las siguientes características:

Equipo 1:

- Marca: Lenovo Z40-70
- Procesador: Intel(R) Core(TM) I7-4510U CPU @ 2.00 GHz 2.60 GHz
- Memoria instalada (RAM): 6,00 GB (5,89 utilizable)
- Tipo de sistema: Sistema operativo Windows 8.1 de 64 bits, procesador x64

Equipo 2:

- Marca: Asus N56J
- Procesador: Intel(R) Core (TM) i7-4700HQ CPU @ 2.40GHz 2.39 GHz
- Memoria instalada (RAM): 8,00 GB (7,89 utilizable)
- Tipo de sistema: Sistema operativo Windows 10 de 64 bits, procesador x64

3.2.4.2 Recursos Software:

- Microsoft Excel: Herramienta utilizada para realizar filtros de información y algunos cálculos matemáticos sobre los datos.
- MySQL Workbench: Herramienta utilizada para la gestión de la base de datos.
- Weka: Herramienta utilizada para implementar las técnicas de minería de datos necesarias.
- NotePad++: Utilizada para la lectura y la creación de los DataSet en formato de los archivos de Weka .arff.
- RapidMiner: Herramienta utilizada para implementar las técnicas de minería de datos necesarias.
- Watson Analytics: Herramienta utilizada para mejorar las imágenes de los resultados encontrados en el proceso de minería.

- C#: Lenguaje de programación usado para crear el programa que facilite la limpieza de los datos.
- Microsoft Word: Programa usado para documentar el proceso y los resultados del desarrollo de proyecto.

3.2.4.3 Orígenes de datos y almacenes de conocimientos

Actualmente, la plataforma cuenta con una base de datos en MySQL que será usada para el proceso de limpieza, además de está, se obtendrán los log de acceso del servidor en los rangos de fechas establecidas para el proceso de minería.

El acceso a dicha información se realiza a través del web master de la empresa administradora de la plataforma, por lo que no hay dificultades en dicho proceso.

3.2.5 Requisitos supuestos y restricciones

3.2.5.1 Requisitos

Los requisitos fundamentales del proyecto son los objetivos planteados con anterioridad, los cuales dependen directamente de la calidad de los datos obtenidos de los almacenes mencionados en el punto inmediatamente anterior, además de esto se debe realizar una socialización de los resultados obtenidos con los directivos de la CCAQ.

3.2.5.2 Restricciones Legales

La plataforma cuenta con información de los usuarios registrados (consumidores y empresarios) que podría ser sensible, sin embargo estos datos no se tendrán en cuenta con el fin de proteger la identidad de los mismos.

3.2.5.3 Restricciones presupuestales

No se contará con ningún tipo de presupuesto extra para el desarrollo por fuera de lo propuesto en el proyecto, por lo que se debe seguir lo planeado debidamente.

3.2.5.4 Restricciones de datos

Los log de acceso solo pueden ser tomados del servidor de la plataforma y sólo aquellos de las fechas propuestas en la selección de datos, esto a través del web master de la empresa contratista. Del mismo modo, no se pueden realizar modificaciones en la estructura actual de la base de datos y tampoco se puede realizar ningún tipo de inserción de código en la plataforma ni modificaciones en la misma

3.2.6 Riesgos y contingencias

Se encontrarán de manera detallada en el documento anexo llamado riesgos y contingencias.

3.2.7 Terminología

3.2.7.1 Terminología del negocio

App Store: es el *marketplace* de aplicaciones para usuarios de Apple, a través del cual miles de desarrolladores de apps del mundo entero ofrecen sus productos y millones de usuarios pueden descargar aplicaciones gratuitas o de pago, las conocidas como Apps y juegos para iPhone/iPad

Play Store: es una plataforma de distribución digital de aplicaciones móviles para los dispositivos con sistema operativo *Android*, así como una tienda en línea desarrollada y operada por Google.

Visitante: persona que ingresa a la plataforma a interactuar con el contenido de la misma, sin necesidad de estar registrado

Marketplace: sitio web principal de la plataforma Oferto donde se publican productos y servicios de distintas empresas.

Mini-sitio: sitio web entregado a las empresas registradas a la plataforma, a través del cual gestionan la publicación de productos y servicios en la plataforma

Categoría: conglomerado de productos que cumplen con características similares

Reserva: acción que ejecuta un visitante en el momento en el que quiere separar un producto de la plataforma, para posteriormente comprarlo de manera presencial.

3.2.7.2 Terminología de Minería de Datos

CRIPS-DM: Cross Industry Standard Process for Data Mining, metodología usada en el desarrollo del proyecto.

Log: equivalente a la palabra bitácora, donde se encontrarán los registros de las interacciones de los visitantes con la plataforma.

Data Set (Almacén de datos): conjunto de datos históricos, internos o externos y descriptivos de un contexto o área de estudio, que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir y analizar los datos con el fin de ayudar en la toma de decisiones estratégicas [Hernández, Ramírez, & Ferri (2004)]

Int: tipo de dato que acepta valores numéricos, utilizados en las bases de datos.

Real: tipo de dato que acepta cadena de caracteres, comprendidos entre: letras, números y símbolos.

Date: tipo de dato que almacena fechas, para manejarlas en el proceso de minería.

Nominal: tipo de dato que puede tomar un conjunto de valores especificados previamente, utilizados en un dataset.

Discretización: es la conversión de un valor numérico en un valor nominal ordenado.

Preprocesamiento: consiste en convertir el uso, el contenido y la estructura de la información, contenida en varias fuentes disponibles de datos, en abstracciones de datos necesarias para el descubrimiento de patrones.

Datos cualitativos: son todos aquellos que contestan la pregunta "¿Cuál?" (ó "¿Cuáles?"), son etiquetas y se dividen en nominales (no se pueden ordenar) y ordinales (se pueden ordenar).

Datos cuantitativos: son los datos que se refieren a números.

Data Warehouse: es una colección de datos orientados a un dominio, integrado, no volátil y variable en el tiempo que ayuda a la toma de decisiones de la empresa u organización.

URL: sigla de *Uniform Resource Locator*, es decir, localizador uniforme de recursos. Es una secuencia de caracteres, de acuerdo a un formato estándar, que se usa para nombrar recursos, como documentos e imágenes en internet, por su localización.

ETL (Extract, Transform y Load): se refiere al proceso que permite mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos o data warehouse para analizar, o en otro sistema operacional para apoyar a un proceso de negocio.

Visitante: persona que accede al sitio web sin tener una sesión iniciada.

Hit: es cada petición que se hace al servidor solicitando un archivo

Sesión: Se define como un periodo continuo de tiempo en el que un usuario está viendo páginas o aplicaciones web dentro de un servidor o dominio

CSV: Los ficheros CSV (del inglés comma-separated values) son un tipo de documento sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y

coma en donde la coma es el separador decimal: Colombia, Ecuador, Brasil….) y las filas por saltos de línea. Los campos que contengan una, un salto de línea o una comilla doble deben ser encerrados entre comillas dobles. El formato CSV es muy sencillo y no indica in juego de caracteres concreto, ni cómo van situados los bytes, ni el formato para el salto de línea. Estos puntos deben indicarse muchas veces al abrir el fichero, por ejemplo, con una hoja de cálculo.

ARFF: es un archivo de texto que describe una línea de casos que comparte un set de atributos, fuero desarrollados por la Machines Learning del Proyecto del Departamento de Informática de la Universidad de Waikato para la herramienta de software llamada WEKA.

WEKA (Waikato Environment for Knowledge Analysis): es un software que implementa algoritmos de minería de datos que pueden aplicarse en bases de datos desde su línea de comandos o bien desde su interfaz gráfica.

3.2.7.3 Determinación de objetivos de la minería de datos

Después de identificar la información con la que se cuenta para desarrollar el proyecto se plantean junto con el director de la plataforma los siguientes objetivos:

- Encontrar patrones de comportamiento de los visitantes de la plataforma
- Identificar grupos de visitantes, según sus características comunes
- Identificar las intenciones de compra de los visitantes de la plataforma
- Encontrar datos que permitan mejorar la distribución de productos y servicios dentro de la plataforma

3.3 Fase II. Comprensión de los datos

Antes de iniciar con el proceso, cabe mencionar que los datos son el prerrequisito para el inicio de cualquier análisis y se debe tener en cuenta con respecto a ellos se debe tener en cuenta lo siguiente basado en el Delta Analítico de Davenport

Calidad en los datos: se define como la capacidad de que los datos sean usados de forma efectiva, económica y rápida para informar y evaluar las decisiones. Necesariamente la calidad de datos es multidimensional, y va más allá del nivel del registro certero de los datos para incluir factores como la accesibilidad, relevancia, oportunidad temporal, metadatos, documentación, capacidad del usuario, expectativas, costo y conocimiento de dominio específico al contexto.

La variedad: se refiere a los diferentes tipos de datos que recoge la empresa; por ejemplo, datos estructurados, no estructurados, datos de datos de texto, datos de imagen y datos de audio y video. En esta dimensión es importante considerar que las empresas inevitablemente compartirán los mismos datos en algunos casos, pero el contar con datos que solamente posee la organización y son únicos puede dar ventaja competitiva a la empresa.

La integración: se refiere a agregar datos de múltiples fuentes dentro y fuera de la organización. Esta dimensión es crítica para conjuntar la información generada por los sistemas transaccionales (como CRM, Recursos Humanos, administración de órdenes) y por los sistemas externos de la empresa. Es importante considerar que los datos que recogemos pueden contener errores o información equivocada, por lo que en esta dimensión es importante considerar la validez de los datos.

La accesibilidad: se debe poder acceder desde cualquier locación y en cualquier momento por los usuarios. Las empresas han creado Almacenes de datos empresariales - Enterprise Data Warehouse o EDW por sus siglas en inglés -, que son repositorios de datos para facilitar el acceso a los mismos a los usuarios. La velocidad de acceso es muy importante en esta dimensión. Los datos deben estar disponibles tan rápidamente como sea posible, pues con la gran cantidad de datos que se generan hoy en día, los datos pueden perder su vigencia muy rápidamente.

La estructura: una estructura de datos es una forma particular de organizar datos para que puedan ser utilizados de manera eficiente. Diferentes tipos de estructuras de datos son adecuadas para diferentes tipos de aplicaciones, y algunos son altamente especializados para tareas específicas. Los datos se encuentran de forma estructurada, semi-estructurada, cuasi-estructurada o no estructurada.

Privacidad: se refiere a las reglas de quién puede acceder a los datos.

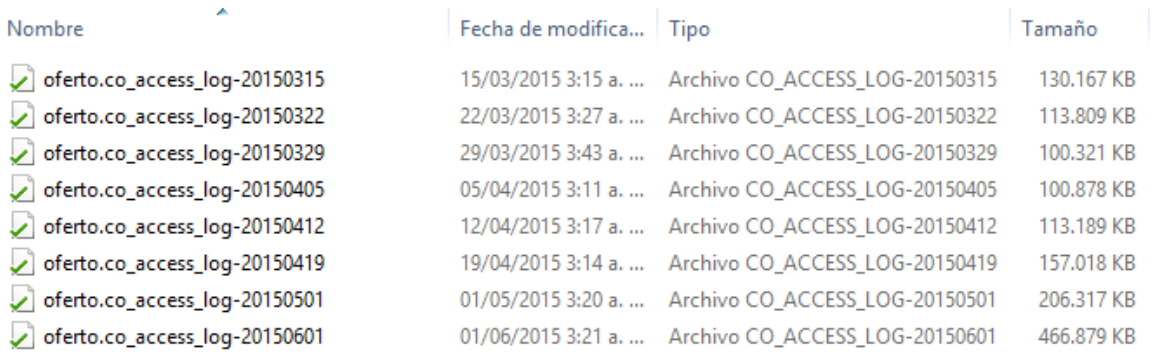
Gobernanza: significa todas las formas en que la empresa se asegura que los datos sean útiles para el análisis. Que los datos sean definidos de manera consistente, de calidad, estandarizados, integrados y accesibles.

En este caso, se realizó una recolección inicial de los datos desde sus orígenes como se describe claramente en la Fase I, donde se puede evidenciar la validez de los mismos, en cuanto a la variedad, encontramos que los datos se deberán tratar desde dos fuentes distintas como lo son la base de datos de la plataforma, en donde se encuentran datos estructurados y los log de acceso del servidor, donde se contará con datos no estructurados. Por lo que se deberá llevar a cabo un proceso de integración de los datos, buscando así obtener un data warehouse que cumpla con características que permitan llegar a obtener buenos resultados. En cuanto a la accesibilidad de los mismos siempre estará sujeta a la disponibilidad del web master de la plataforma, ya que es la única persona con los permisos de acceso a los mismos, lo que determina la privacidad de los mismos; se identifica que los datos son semi-estructurados. Además se procedió a realizar un análisis de ellos para de esta manera buscar relaciones y poder así identificar algún tipo de información oculta o completar información faltante, encontrando que inicialmente se puede pensar en relaciones entre las horas de visitas y las categorías que visitan, los días con las horas de visitas, entre otras.

3.3.1 Recopilación de datos iniciales

Se inicia el proceso de extracción de los datos basados en el proceso ETL, obteniendo los datos del log de acceso del servidor.

En múltiples trabajos sobre minería, la fuente de información principal es una base de datos, sin embargo en este caso vamos a minar el uso de un sitio web para esto se solicitó al web master descargar del servidor de la plataforma Oferto www.oferto.co los log de acceso de los días entre el 8 de Marzo y el 21 de Mayo del 2015, que fueron aquellos que se determinaron debían ser la muestra dentro del proceso de minería. En estos archivos planos se encuentran registrados todos los eventos que se realizaron por los visitantes de dicha plataforma en este tiempo, a continuación se visualizan dichos archivos.



Nombre	Fecha de modifica...	Tipo	Tamaño
oferto.co_access_log-20150315	15/03/2015 3:15 a. ...	Archivo CO_ACCESS_LOG-20150315	130.167 KB
oferto.co_access_log-20150322	22/03/2015 3:27 a. ...	Archivo CO_ACCESS_LOG-20150322	113.809 KB
oferto.co_access_log-20150329	29/03/2015 3:43 a. ...	Archivo CO_ACCESS_LOG-20150329	100.321 KB
oferto.co_access_log-20150405	05/04/2015 3:11 a. ...	Archivo CO_ACCESS_LOG-20150405	100.878 KB
oferto.co_access_log-20150412	12/04/2015 3:17 a. ...	Archivo CO_ACCESS_LOG-20150412	113.189 KB
oferto.co_access_log-20150419	19/04/2015 3:14 a. ...	Archivo CO_ACCESS_LOG-20150419	157.018 KB
oferto.co_access_log-20150501	01/05/2015 3:20 a. ...	Archivo CO_ACCESS_LOG-20150501	206.317 KB
oferto.co_access_log-20150601	01/06/2015 3:21 a. ...	Archivo CO_ACCESS_LOG-20150601	466.879 KB

Ilustración 8: Archivos planes de plataforma

Cada uno de los archivos de log cuenta con información que se debe ser analizada para llevar a cabo adecuadamente el proceso minería; a continuación se ejemplifica la forma en que inicialmente se encuentran:

```
71.47.10.48 - - [12/Apr/2015:03:17:26 -0500] "GET /files/galerias/b120140805071021.jpg HTTP/1.1" 200 61718 "http://www.bing.com/images/searc
66.249.69.44 - - [12/Apr/2015:03:17:37 -0500] "GET /main-productos-pagina-1-orden-asc-cant-30-por-precio HTTP/1.1" 200 63350 "-" "Mozilla/5.
54.187.7.109 - - [12/Apr/2015:03:17:46 -0500] "GET /main-producto-id-3714-t-camandula_en_acero HTTP/1.1" 200 32926 "-" "Mozilla/4.0 (compati
54.187.7.109 - - [12/Apr/2015:03:17:49 -0500] "GET /system/vista/skin/comercio/media/js/prototype1.7.js HTTP/1.1" 200 584260 "-" "Mozilla/4.
66.249.88.253 - - [12/Apr/2015:03:17:54 -0500] "GET /files/galerias/b120140919070704.jpg HTTP/1.1" 200 54052 "https://www.google.com/" "Mozi
71.47.10.48 - - [12/Apr/2015:03:17:58 -0500] "GET /files/galerias/b220140805071022.jpg HTTP/1.1" 200 101649 "http://www.bing.com/images/sear
180.76.5.169 - - [12/Apr/2015:03:17:59 -0500] "GET /robots.txt HTTP/1.1" 301 249 "-" "Mozilla/5.0 (Windows NT 5.1; rv:6.0.2) Gecko/20100101
180.76.6.142 - - [12/Apr/2015:03:17:59 -0500] "GET /robots.txt HTTP/1.1" 200 44 "-" "Mozilla/5.0 (Windows NT 5.1; rv:6.0.2) Gecko/20100101 F
180.76.5.64 - - [12/Apr/2015:03:18:04 -0500] "GET / HTTP/1.1" 301 239 "-" "Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/s
180.76.6.61 - - [12/Apr/2015:03:18:04 -0500] "GET / HTTP/1.1" 200 91651 "-" "Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com
66.249.69.104 - - [12/Apr/2015:03:18:09 -0500] "GET /robots.txt HTTP/1.1" 200 44 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.go
66.249.69.96 - - [12/Apr/2015:03:18:09 -0500] "POST /producto-tabla_carrito HTTP/1.1" 200 1152 "http://www.megatennis.oferto.co/main-product
180.76.6.20 - - [12/Apr/2015:03:18:14 -0500] "GET /main-productos-pagina-2-cant-15 HTTP/1.1" 200 33600 "-" "Mozilla/5.0 (compatible; Baidusp
66.249.69.5 - - [12/Apr/2015:03:18:16 -0500] "GET /main-productos-c-1749-pagina-1-orden-desc-por-nombre-cant-9 HTTP/1.1" 200 37314 "-" "Mozi
```

Ilustración 9: formato inicial de archivos

Además de los log, se contará con la base de datos de la plataforma que cuenta con un total de 50 tablas, en las que se encuentran 231302 registros. Esto con el fin de poder relacionar adecuadamente los datos obtenidos a través del log y llegar así a resultados confiables; la estructura básica de dicha base de datos se muestra a continuación:

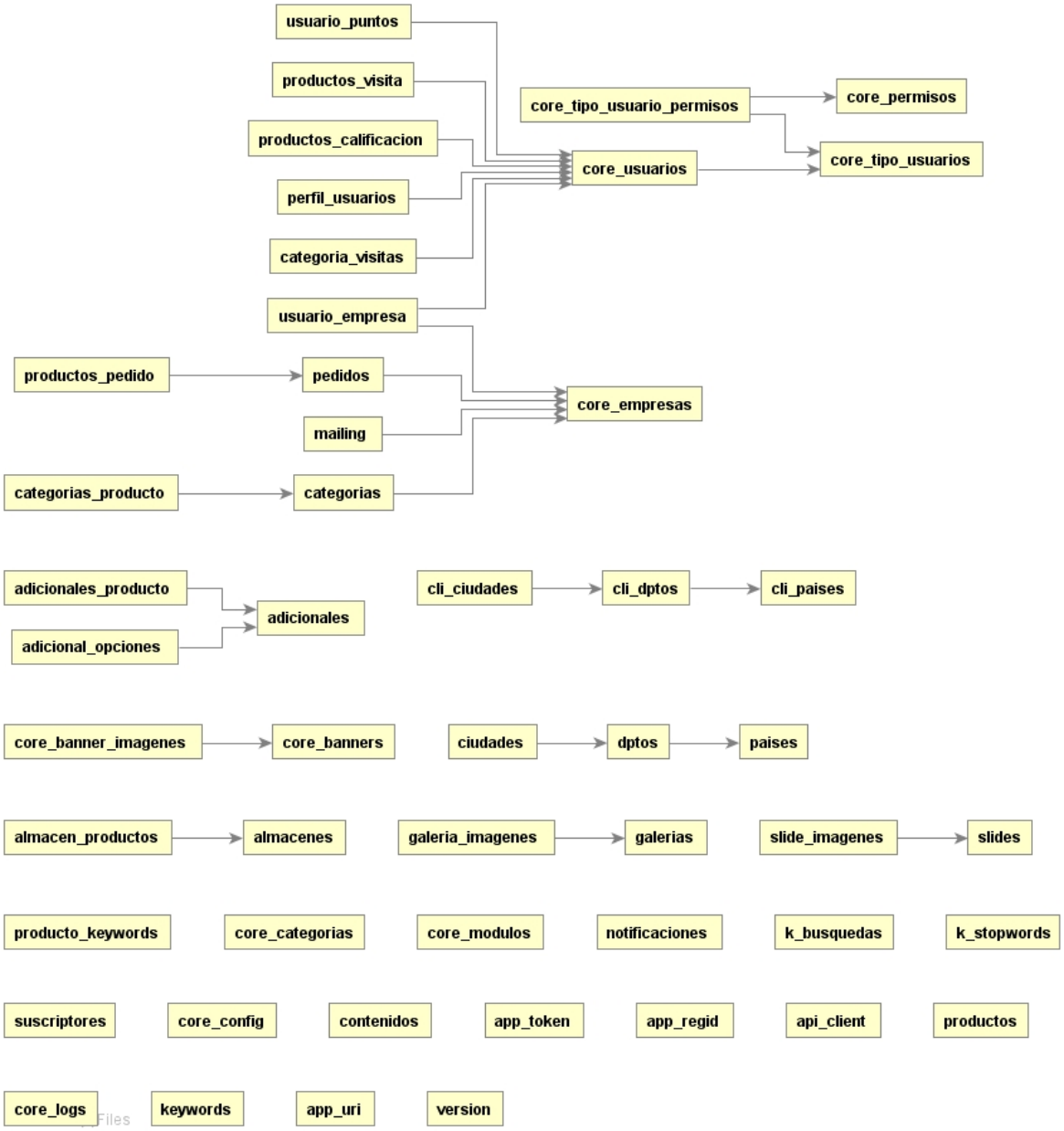


Ilustración 10: estructura básica de base de datos

Se puede pensar que estos datos son adecuados para afrontar el proceso, además de tener unas características iniciales que podrían ser suficientes para obtener conclusiones al momento de ser necesario integrar los datos obtenidos de distintos orígenes, sin embargo se debe tener en cuenta que se pueden generar problemas en dicho proceso, ya que no fueron diseñados de tal manera que la relación entre los mismo se haga fácilmente.

En este momento la empresa no tiene planes de realizar modificaciones a la base de datos con la que actualmente trabaja, ni de realizar ningún tipo de modificación al código fuente de la plataforma, por lo que se deberá trabajar sobre la información anteriormente mencionada y buscar herramientas existentes o desarrolladas durante el proceso para llevar a cabo la limpieza e integración.

3.3.2 Descripción de los datos

3.3.2.1 Log

Cada vez que un visitante realiza una visita a un sitio web, en un archivo llamado fichero log o log file del servidor queda registrada toda la información sobre esa visita. Generalmente los log se clasifican en tres tipos:

- Log de acceso – Access log
- Log de error – Error log
- Log de referencia – Referer log

Además pueden ser guardados en dos tipos de formato como lo son dichero de log común – *Common Log File* (CLF) o en un formato de fichero de log extendido – *Extended Log File* (ELF). En este caso los log utilizados son los log de acceso con formato ELF, que tienen en total 4.255.083 de registros distribuidos como se describen a continuación:

Log´s Oferto			
Nombre del Archivo	**Inicio de Registro**	**Fin de Registro**	**Numero de Registros**
Oferto1	08/03/2015 3:03	15/03/2015 3:15	454.624
Oferto2	15/03/2015 3:16	22/03/2015 3:27	403.280
Oferto3	22/03/2015 3:27	29/03/2015 3:43	350.043
Oferto4	29/03/2015 3:43	05/04/2015 3:11	363.299
Oferto5	05/04/2015 3:11	12/04/2015 3:17	398.700
Oferto6	12/04/2015 3:17	19/04/2015 3:14	520.484
Oferto7	19/04/2015 3:15	01/05/2015 3:19	716.077
Oferto8	01/05/2015 3:20	21/05/2015 10:03	1.048.576
		Total Registros	**4.255.083**

Tabla 5: distribución de logs

Cada log se compone de 1 registro por cada interacción del visitante y cada registro de elementos.

Ejemplo Log ELF: 71.47.10.48 - - [12/Apr/2015:03:17:26 -0500] "GET /files/galerias/b120140805071021.jpg HTTP/1.1" 200 61718 "http://www.bing.com/images/search?q=shampoo+anticaspa+del+ponte+verde&id=BE44A713BC29D00079E2BC6DB00502406E99429F&FORM=IQFRBA" "Mozilla/5.0 (iPad; CPU OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) Version/8.0 Mobile/12B466 Safari/600.1.4"

- **Remotehost (host remoto):** En el ejemplo: 71.47.10.48
- **Rfc931:** el nombre del log remoto del usuario: En el ejemplo: -
- **Authuser:** el nombre con el que el usuario se identificó. En el ejemplo: -
- **Fecha: Fecha y hora de la solicitud. En el ejemplo:** [12/Apr/2015:03:17:26 -0500]
- **“Solicitud”:** la línea exacta de petición según viene solicitada desde el cliente. En el ejemplo: “GET /files/galerias/b120140805071021.jpg HTTP/1.1".
- **Estado:** el código de estado del HTTP devuelto al cliente. En el ejemplo: 200
- **Bytes:** la longitud que tiene el documento contenido. En el ejemplo: 61718

- **Referente:** URL desde donde se ha realizado la petición. En el ejemplo: http://www.bing.com/images/search?q=shampoo+anticaspa+del+ponte+verde&id=BE44A713BC29D00079E2BC6DB00502406E99429F&FORM=IQFRBA
- **Agente:** Tipo de navegador y sistema operativo usado. En el ejemplo: "Mozilla/5.0 (iPad; CPU OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) Version/8.0 Mobile/12B466 Safari/600.1.4"

3.3.2.2 Tablas

A continuación se describen cada una de las tablas de la base de datos:

Core_usuarios: almacena la información de los usuarios como lo son id, username, password, tipo, empresa, estado, email, borrado, suscrito.

Core_tipo_usuarios: es el lugar donde se establecen los tipos de usuario que van a existir dentro de la plataforma.

Core_permisos: almacena los funciones a controlar de acuerdo a los roles de los usuarios de la plataforma.

Core_tipo_usuarios_permisos: es la tabla por medio de la cual relacionan cada uno de los tipos de usuario que existen y aquellas funciones que puede ejecutar.

Usuario_puntos: almacena los puntos con los que cuenta un usuario registrado en la plataforma por sus compras, cabe aclarar que el empresario es quien decide si dar o no puntos a los compradores.

Productos_visita: almacena la cantidad de visitas que realiza un usuario que ha iniciado sesión dentro de la plataforma.

Productos_calificacion: es la tabla donde se registra la información de la calificación dada por un usuario que realizo una reserva a través de la plataforma y finalmente compro el producto y / o servicio.

Perfil_usuarios: almacena la información de la persona que administra el mini sitio dado a la empresa, datos como id, nombres, apellidos, teléfono, móvil, dirección, departamento, país y ciudad.

Categoria_visitas: relaciona las visitas hechas a una categoría por un usuario que inicio sesión en la plataforma.

Usuario_empresa: almacena el registro de los usuarios que siguen empresas y han realizado compras en las mismas.

Pedidos: almacena los datos de los pedidos (reservas) de productos y / o servicios hechos por los usuarios a las diferentes empresas, con datos como id_usuario, id_empresa, orden, fecha, nombres, apellidos, teléfono, móvil, dirección, departamento, país, ciudad, email, estado, total, método_pago, dominio, compra, código_descuento

Productos_pedido: relaciona la cantidad de cada uno de los productos con su pedido y un campo de adicional.

Mailing: registra los mails enviados a través de cada uno de los mini-sitios de los empresarios con sus respectivos datos como lo son: asunto, contenido, fecha, filtro, estado y el resultado del envió.

Adicionales: registra los adicionales creados por las empresas con su respectiva relación.

Adicionales_producto: relaciona los adicionales creados con los productos a los que se les aplica ese adicional.

Adicional_opciones: registra las opciones de los adicionales creados.

Cli_ciudades: almacena las ciudades que pueden ser usadas en la plataforma.

Cli_departamentos: almacena los departamentos que pueden ser usadas en la plataforma.

Cli_paises: almacena los países que pueden ser usadas en la plataforma.

Core_banner_imagenes: almacena la imagen, la posición y la categoría donde va a verse dicha imagen.

Core_banners: almacena los *banners* con los que cuenta la plataforma con su nombre, descripción, fecha y el id de la categoría a la que pertenece.

Ciudades: relaciona los departamentos con sus respectivas ciudades.

Dptos: relaciona los países con sus respectivos departamentos.

Países: registra los países y sus iniciales.

Almacenes: almacena la información de cada uno de los almacenes registrados por las empresas en su mini sitio, con datos como: id_almacen, id_empresa, nombre, dirección, teléfono, móvil, imagen, latitud, longitud, ubicación, id_pais, id_dpto, id_ciudad.

Almacén_productos: almacena la relación entre los almacenes y los productos.

Galería_imágenes: almacena las imágenes de cada una de las galerías de la plataforma.

Galerías: almacena las galerías de imágenes creadas dentro de la plataforma.

Slides: almacenan los slides creados para los empresarios en sus mini sitios.

Slide_imágenes: almacena la relación del slide a la imagen usada en él.

K_búsquedas: almacena las palabras por las que un usuario que ha iniciado sesión realiza.

K_stopword: almacena palabras que son consideradas vacías como artículos, pronombres y preposiciones.

Keywords: almacena las palabras claves de la plataforma

Producto_keywords: relaciona los productos con las palabras claves creadas.

Core_módulos: almacena los módulos en los que se tiene dividida la plataforma, por ejemplo: login, contenido, galería, main, etc.

Notificaciones: almacena las notificaciones *push* enviadas por el administrador de la plataforma las apps instaladas, almacenando el id de la notificación, titulo, mensaje, imagen, fecha, id_oferta, id_categoria, acción y estado.

Suscriptores: almacena los datos de los usuarios que se han suscrito para el envio de mails por parte de las empresas.

Core_config: almacena los datos de configuración base de la plataforma.

Contenidos: almacena los datos de la información puesta dentro de la plataforma y los mini sitios como títulos, términos y condiciones de uso.

Core_empresas: Almacenan la información de las empresas, como lo son id, id_categoria, nombre, sloga, descripción empresa, fecha de registro, estado, logo, skin, color, Facebook, twitter, youtube, linnkedin, título, telefonos, direccion, Skype, google maps, email, tipodominio, subdominio, dominio, web, impuesto, latitud, longitud, impuesto porcentaje, descripción, pago payu, pago_otro, entre otros que por confidencialidad no serán usados. En este momento aparecen un total de 1234 registros

Productos: almacena la información de los productos cargados por cada una de las empresas registradas a la plataforma, de los cuales se cuenta con información como el id del producto, id empresa, id galería, nombre, descripción, imagen, detalles, referencia, si se encuentra destacado, visible, su estado, si se encuentra en oferta y si es así, los datos de la oferta, el precio, cantidad de compras, las visitas que ha tenido y su calificación. Se cuenta actualmente con 10164 registros.

Core_categorias: Cuenta con la información de las categorías en las que se organiza la información en el marketplace de la plataforma, cada empresa debe elegir una categoría a la que

pertenece y es en esta categoría donde aparecen los productos que publiquen en la plataforma, de esta categoría se tiene el id, nombre y el id de la galería. En la actualidad la plataforma tiene en su Marketplace un total de 38 categorías, que solo se activan al tener productos en oferta.

Categorías_producto: permite la relación entre los productos y las categorías internas que maneja cada una de las empresas en su sitio web, allí se enlaza el id de la categoría del producto, el id de la categoría y el id del producto. Se encuentran 10827 registros en esta tabla.

Categorías: almacena las categorías creadas dentro de los sitios web de los empresarios registrados a las plataformas, se cuenta con el id de la categoría, id empresa y nombre de esa categoría interna. Al momento de realizar la revisión se encuentran 2601 categorías creadas por las empresas.

3.3.3 Exploración de datos

Teniendo en cuenta lo que se establece en la metodología CRISP-DM, esta se puede y debe personalizar de acuerdo a las necesidades del proceso, además de ser un proceso cíclico en el que es necesario en muchas ocasiones retroceder y avanzar entre sus fases, por lo que en este caso es necesario para realizar la exploración de los datos del log del servidor haber procesado en la fase de preparación de datos, para poder tener datos que se puedan explorar de forma significativa, por lo que a continuación se realizará dicho proceso.

3.4 Fase III. Preparación de los datos:

3.4.1 Selección de los datos

Como se mencionó en la fase de comprensión de los datos, se cuenta con 8 archivos de log, que tienen un total de 4.255.083 registros, además de la base de datos de la plataforma oferto que cuenta con un total 50 tablas y 231302 registros.

3.4.1.1 Base de datos

De las tablas identificadas se podrían excluir inmediatamente mailing, adicionales_producto, adicional_opciones, adicionales, core_banner_imagenes, core_banners, almacen_productos, almacenes, galería_imagenes, galerías, slide_imagenes, slides, core_modulos, notificaciones, core_config, contenidos, app_token, app_regid, api_client, app_uri, version, esto debido a que luego de analizar la información que contienen no son relevantes con respecto a los objetivos planteados para el proceso. Por otra parte, se identifican un total de 5 tablas que podrían tener información relevante para el proyecto de investigación, las cuales son core_empresas, productos, core_categorias, categorias_producto y categorias, por lo que las tablas anteriormente descritas serán utilizadas durante el proceso, esto basados en que dentro del log se debe buscar establecer relaciones que incluyen tanto a las empresas, categorías y productos, para lo que los registros almacenados en estas tablas serían de gran ayuda al momento de realizar el cruce de información y la limpieza de los archivos de log.

3.4.1.2 Archivos de log

Se utilizarán los archivos de log descritos con anterioridad, a los que primero se les hará el proceso de limpieza para poder determinar con que datos se trabajará en la fase de modelado.

3.4.2 Limpieza de datos

Es el proceso mediante el cual se busca eliminar todos aquellos datos incompletos, inconsistentes, erróneos, que nos puedan llevar a tomar decisiones o sacar conclusiones equivocadas. En este proceso fueron muchos problemas los que se debieron resolver, esto debido a que en un proceso de minería web los archivos de Log no cuentan con un formato adecuado para dar inicio inmediato al trabajo de minería; por lo que se debió adaptar, buscando que fuera

factible la aplicación de técnicas de limpieza de datos como la transformación, adición y reducción a los archivos ELF, para lograr esto se realizaron los siguientes pasos:

1. Se remueven los registros que contienen peticiones a recursos multimedia y documentos como lo son .jpg, .png, .gif, .css, .js, .txt, .ico, .jpeg, .woff, .php, .pdf, .eot, .doc, .ttf, .bmp, siendo estos los más relevantes.
2. Se remueven los registros generados por Bots, Spiders y Crawlers ya que no representan una actividad real de un visitante.
3. Se remueven todas las peticiones al servidor que no sean tipo GET ya que solo nos interesan las peticiones de consulta.
4. Se remueven todos los registros que contengan una respuesta errónea de parte del servidor. Estas respuestas se reconocen gracias al código 4XX o 5XX contenidos en la URL.
5. Se remueven todos los registros que se hayan generado de manera local desde el equipo en el cual está alojada la plataforma de Oferto.

Para este proceso se implementó un algoritmo en C# que realizaba cada uno de los procedimientos mencionados anteriormente y se analizaron 4.841.871 registros.

3.4.2.1 Integración de Log

Se integraron todos los registros contenidos inicialmente en cada uno de los 8 archivos de Log en un único archivo con el fin de realizar un solo proceso con la totalidad de los datos.

3.4.2.2 Análisis inicial del Log

Se cargó el archivo de Log unificado al software Weblog Expert, el cual genero un reporte del estado del log, dando como resultados la cantidad de hits, distribuidos en visitantes y spiders, el promedio de visitas por día, por visitante, solicitudes de cache, solicitudes fallidas, total de páginas visitadas, promedio de páginas vistas por día, vistas por visitante, total de visitantes, promedio de visitantes por día, total de IPs únicas, total de ancho de banda, ancho de banda de

visitantes, ancho de banda de spiders, promedio de ancho de banda por día, por hit y por visitante, tal como se puede ver en la siguiente tabla y detalladamente en el anexo.

Hits	
Total Hits	5,191,914
Visitor Hits	3,808,992
Spider Hits	1,382,922
Average Hits per Day	60,371
Average Hits per Visitor	22.47
Cached Requests	249,902
Failed Requests	44,949
Page Views	
Total Page Views	954,081
Average Page Views per Day	11,093
Average Page Views per Visitor	5.63
Visitors	
Total Visitors	169,535
Average Visitors per Day	1,971
Total Unique IPs	103,127
Bandwidth	
Total Bandwidth	131.21 GB
Visitor Bandwidth	87.01 GB
Spider Bandwidth	44.20 GB
Average Bandwidth per Day	1.53 GB
Average Bandwidth per Hit	26.50 KB
Average Bandwidth per Visitor	538.17 KB

Tabla 6: promedio de ancho de banda por día, por hit y por visitante

Debido a que el objetivo es analizar el comportamiento de los visitantes dentro de la plataforma, la aparición de registros que se generan por la actividad de Spider se considera indeseada (son cerca del 26 % de los registros), así como las peticiones de recursos multimedia, las peticiones fallidas y las peticiones generadas por el mismo servidor en el cual está alojada la plataforma de Oferto.

3.4.2.3 Eliminación de peticiones erróneas

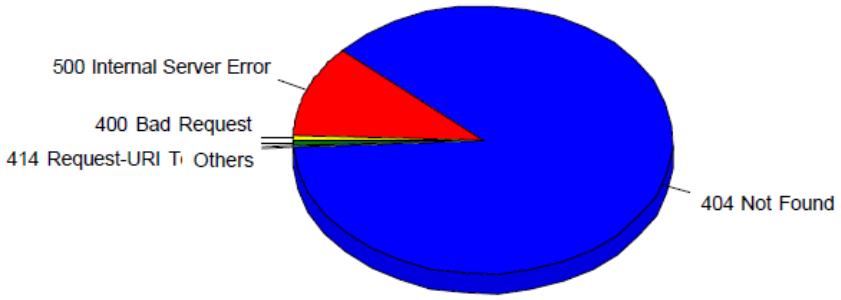
En el campo estado se eliminaron las peticiones que no son exitosas, lo cual se realizó quitando todas aquellas que inicien con los códigos de error de servidor 400 y 500. Dichos valores llamados códigos de petición http se especifican en la tabla 7.

CODIGO DE ESTADO DEL SERVIDOR	
CODIGO	**SIGNIFICADO**
1xx	**Información**
100	Continue
101	Switching Protocolo
2xx	**Éxito**
201	Creado
202	Aceptado
203	Información no autorizada
204	No hay contenido
205	Contenido reiniciado
206	Contenido Parcial
3xx	**Redirección**
300	Selecciones Multiples
301	Movido Permanentemente
302	Encontrado
303	Observar otro
304	No modificado
305	Uso de Proxy
307	Redirección temporal
4xx	**Error en el Cliente**
400	Petición errónea
401	No autorizado
402	Pago requerido
403	Prohibido
404	No encontrado
405	Método no permitido
406	No aceptado
407	Requiere autenticación del Proxy
408	Tiempo de espera agotado
409	Conflicto
5xx	**Error del Servidor**
500	Error interno del servidor

501	No implementado
502	Error del Gateway
503	Servicio no disponible
504	Tiempo agotado en el Gateway
505	Versión http no soportada

Tabla 7: código de estado del servidor

Inicialmente en el Log se pueden encontrar 44.949 registros fallidos que corresponden al 8% del total de los registros, y se encuentran distribuidos de la siguiente forma:



Error Types

	Error	Hits
1	404 Not Found	39,356
2	500 Internal Server Error	4,952
3	400 Bad Request	326
4	403 Forbidden	229
5	414 Request-URI Too Long	63
6	416 Requested Range Not Satisfiable	14
7	405 Method Not Allowed	9
	Total	44,949

Tabla 8: distribución de registros fallidos

3.4.2.4 Filtrado de imágenes y datos ruidosos

Normalmente, las páginas web cargan archivos de sonidos, videos o imágenes. Por lo que el servidor web registra las entradas que fueron solicitadas de dichos archivos, también las que fueron enviadas. Los registros generados debido a estas peticiones son 2.045.681 y corresponden al 39.4% del total de registros analizados. Los recursos con mayor número de solicitudes fueron las peticiones a imágenes .png y .gif.

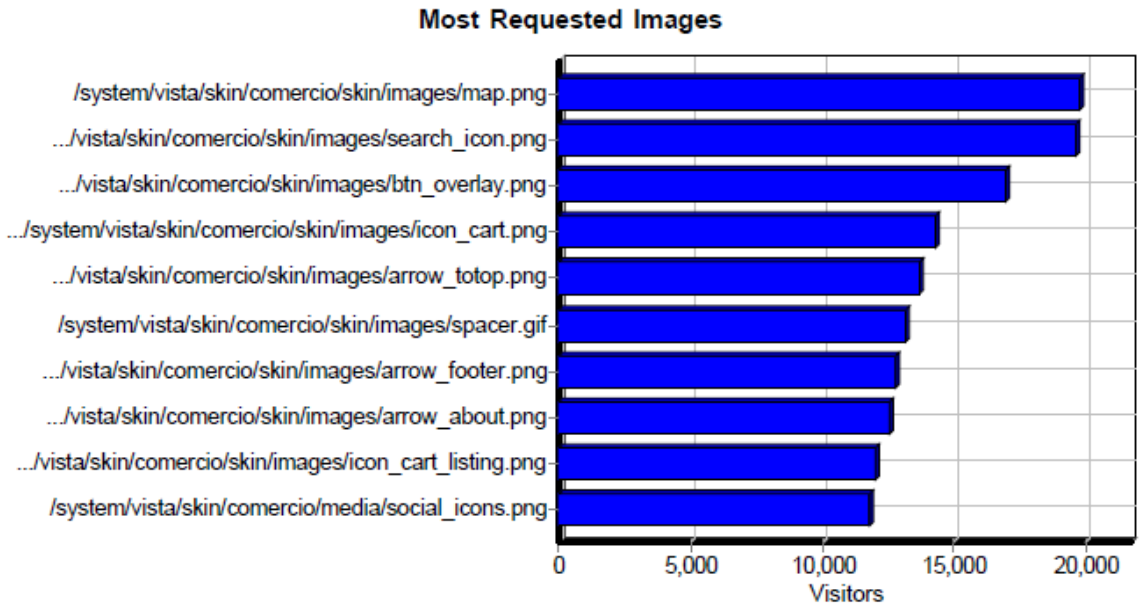


Ilustración 11: recursos con mayor número de solicitud

3.4.2.5 Eliminación de robots de acceso web

El Log, normalmente cuenta con entradas de robots web, como lo son: índices, *crawlers*, *spiders*, entre otros, que en su mayoría son generados por el administrador del sitio web para generar los permisos de acceso. Según Thurow (2003), los robots realizan tres acciones básicas:

"En primer lugar encontrar las páginas del sitio (proceso conocido como gatear o rastreo) y crear una lista de palabras y frases que se encuentran en cada página; Con esta lista lo hacen una base de datos y encontrar las páginas exactas que deben buscar introduciendo el lugar buscó la base de datos organizada por las características generales que se encuentran en sus páginas. La máquina entra en el sitio en la base de datos general se llama divisor; después de que el robot es ahora capaz de encontrar este sitio cuando el usuario final escriba una palabra o frase de búsqueda en el contenido que se encuentra en el sitio. Este paso se llama procesador de consultas."

En el reporte resultante del Weblog Expert se encuentra una lista de los Spider, Robots y Crawlers se utilizó para definir cuales registros debían separarse del proceso ya que no

corresponden a peticiones que hayan sido realizadas por un usuario. El total de registros que corresponden a peticiones realizadas por Spiders y Crawlers es de 1.382.922 y corresponde al 26.6% del total de registros.

La lista de Spiders y Crawlers encontrados en el Log fueron los siguientes:

Top Spiders

	Spider	Hits
1	Googlebot	610,776
2	MJ12bot	221,011
3	Bing Robot	171,853
4	Baidu Spider	152,190
5	AhrefsBot	82,734
6	FaceBook Crawler	35,321
7	Google Preview Robot	21,179
8	Yahoo! Slurp	14,251
9	Yandex Robot	12,714
10	Bing Preview Robot	11,407
11	DotBot	10,603
12	Mozilla/5.0 (compatible; BLEXBot/1.0; +http://webmeup-crawler.com/)	9,467
13	AddThis.com robot tech.support@clearspring.com	5,473
14	Twitterbot	4,625
15	Mozilla/5.0 (compatible; 007ac9 Crawler; http://crawler.007ac9.net/)	3,794
16	LSSRocketCrawler/1.0 LightspeedSystems	2,077
17	Sogou Spider	1,592
18	SemrushBot	1,559
19	SeznamBot	1,340
20	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/27.0.1453.116 Safari/537.36 HubSpot Webcrawler	1,242
21	Mozilla/5.0 (compatible; Linux x86_64; Mail.RU_Bot/2.0; +http://go.mail.ru/help/robots)	1,034
22	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0); 360Spider(compatible; HaosouSpider; http://www.haosou.com/help/help_3_2.html)	962
23	Google AdWords Robot	799
24	Mozilla/5.0 (Windows NT 6.1) AppleWebKit/537.1 (KHTML, like Gecko) Chrome/21.0.1180.89 Safari/537.1; 360Spider(compatible; HaosouSpider; http://www.haosou.com/help/help_3_2.html)	655
25	Mozilla/5.0 (compatible; SEOkicks-Robot; +http://www.seokicks.de/robot.html)	475
26	LinkdexBot	461
27	Exabot	322
28	meanpathbot	276
29	Rogerbot	261
30	NerdyBot	252
31	SurveyBot	240
32	Mozilla/5.0 (compatible; oBot/2.3.1; http://filterdb.iss.net/crawler/)	204
33	Mozilla/5.0 (compatible; proximic; +http://www.proximic.com/info/spider.php)	193
34	Mozilla/5.0 (Windows; Crawler; U; Windows NT 6.0; en-US; rv:1.9.0.7) Gecko/2009021910 Firefox/3.0.7 (.NET CLR 3.5.30729)	188
35	Mozilla/5.0 (compatible; GrapeshotCrawler/2.0; +http://www.grapeshot.co.uk/crawler.php)	121
36	Turnitin Robot	120
37	LinkWalker	101
38	LivelapBot/0.2 (http://site.livelap.com/crawler)	91
39	Mozilla/5.0 (compatible; NetSeer crawler/2.0; +http://www.netseer.com/crawler.html; crawler@netseer.com)	74
40	['GD CRAWLER']	68
41	Mozilla/5.0 (compatible; oBot/2.3.1; +http://filterdb.iss.net/crawler/)	62
42	Netcraft Spider	62
43	YisouSpider	56
44	ShowyouBot (http://showyou.com/crawler)	48
45	Mozilla/5.0 (compatible; DomainSigmaCrawler/0.1; +http://domainsigma.com/robot)	47
46	Mozilla/5.0 (compatible; DomainTunoCrawler/0.1; +http://www.domaintuno.com/robot)	42
47	LTX71 Robot	42
48	CRAZYWEBCRAWLER 0.9.2, http://www.crazywebcrawler.com	39
49	Google AdSense Robot	38
50	Archive.org Robot	38
	Subtotal	**1,382,579**
	Total	**1,382,922**

Tabla 9: lista de Spiders y Crawlers

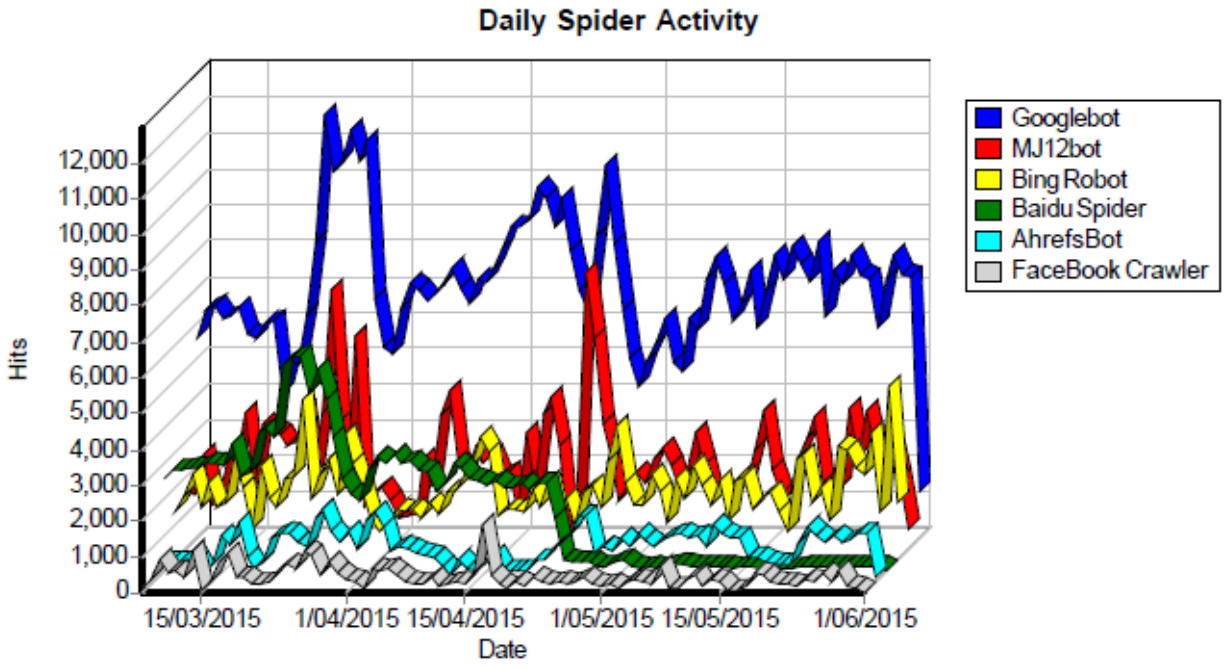


Ilustración 12: gráfico de actividad diaria de Spiders

Mediante la implementación de un algoritmo de análisis, se buscó reducir el número de este tipo de registros en la mayor cantidad posible.

3.4.2.6 Filtrado de peticiones

Teniendo en cuenta que muchos de los registros del archivo de Log no son necesarios para el proceso, se tomaron como registros válidos los que cumplan con el criterio de que el campo solicitud tenga el siguiente formato: “GET <petición>”, tomando a <petición> la solicitud realizada al servidor. Como resultado del análisis del log, se encontraron que 128.058 registros pertenecían a peticiones que no correspondían a solicitudes tipo GET y representan el 2.1% del total de los datos.

3.4.2.7 Resultados de la limpieza

Como resultado del algoritmo se obtuvieron los siguientes resultados:

a) 3.557.143 registros que contienen peticiones a recursos multimedia y documentos eliminados.

b) 128.058 registros eliminados que contenían peticiones al servidor que no eran tipo GET.

c) 630.219 registros generados por *Bots, Spiders y Crawlers* eliminados.

d) 10.454 registros con respuesta errónea de parte del servidor.

e) 7.027 registros provenientes de la MAC del servidor en la cual está alojada la aplicación.

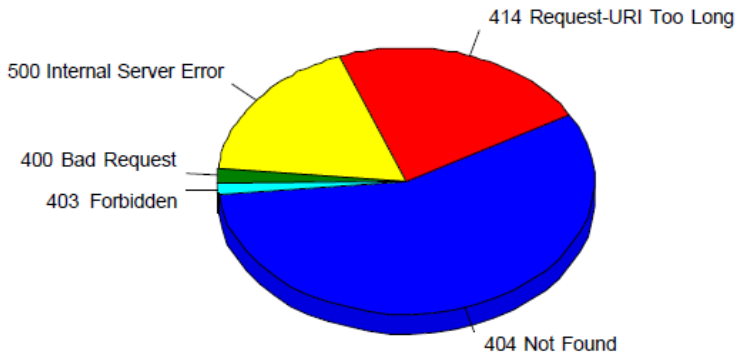
Posterior al proceso de limpieza del Log mediante un algoritmo desarrollado en C# (Ver anexo), el archivo Log resultante se analiza nuevamente con el software Weblog Expert, con el fin de validar los resultados de la limpieza en términos de peticiones realizadas por Bots, peticiones a recursos y peticiones fallidas, tal como se evidencia en la tabla número 10.

Hits	
Total Hits	505,732
Visitor Hits	501,834
Spider Hits	3,898
Average Hits per Day	5,880
Average Hits per Visitor	9.68
Cached Requests	15
Failed Requests	210
Page Views	
Total Page Views	501,395
Average Page Views per Day	5,830
Average Page Views per Visitor	9.67
Visitors	
Total Visitors	51,855
Average Visitors per Day	602
Total Unique IPs	31,537
Bandwidth	
Total Bandwidth	10.20 GB
Visitor Bandwidth	9.95 GB
Spider Bandwidth	262.55 MB
Average Bandwidth per Day	121.47 MB
Average Bandwidth per Hit	21.15 KB
Average Bandwidth per Visitor	201.11 KB

Tabla 10: validación de los resultados de la limpieza

Posterior al proceso de limpieza, el Log resultante contiene un total de hits de 505.732, de estos, 501.834 pertenecens a los visitantes de la plataforma. Se encuentra que se tienen 602 visitantes por día en promedio y 51.855 visitas durante toda la línea de tiempo.

El número de peticiones erróneas bajaron a 281 (Ver tabla 11) así como el número de peticiones a recursos que fueron 1 en total (ver tabla 12).



Error Types

	Error	Hits
1	404 Not Found	160
2	414 Request-URI Too Long	63
3	500 Internal Server Error	50
4	400 Bad Request	5
5	403 Forbidden	4
	Total	**282**

Tabla 11: número de peticiones erróneas

	Image	Hits	Incomplete Requests	Visitors	Bandwidth (KB)
1	http://www.oferto.co/files/editor/157/images/DSC00361(1).JPG	1	0	1	4,440
	Total	**1**	**0**	**N/A**	**4,440**

Tabla 12: peticiones a recursos

El número de peticiones realizadas por *Spider* después del proceso de filtración fue de 3.989 y corresponde a un 0.77% del total de registros. Se puede apreciar un gran número de peticiones realizadas por el Bot de Bing, pero después de una inspección manual sobre los registros se concluyó que estas entradas realmente corresponden a redirecciones que se realizaron desde el

buscador Bing, por tal motivo es pertinente incluirlas dentro del análisis, ya que sí corresponden con la actividad de un usuario (ver tabla 13).

	Spider	Hits
1	Bing Robot	2,662
2	Mozilla/5.0 (compatible; SEOkicks-Robot; +http://www.seokicks.de/robot.html)	303
3	Mozilla/5.0 (Windows; Crawler; U; Windows NT 6.0; en-US; rv:1.9.0.7) Gecko/2009021910 Firefox/3.0.7 (.NET CLR 3.5.30729)	170
4	Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/27.0.1453.116 Safari/537.36 HubSpot Webcrawler	147
5	meanpathbot	129
6	LinkWalker	100
7	Turnitin Robot	63
8	['GD CRAWLER']	62
9	Googlebot	42
10	Netcraft Spider	38
11	Mozilla/5.0 (compatible; NetSeer crawler/2.0; +http://www.netseer.com/crawler.html; crawler@netseer.com)	38
12	Google AdSense Robot	37
13	Google Feedfetcher	27
14	YisouSpider	17
15	Entireweb Robot	10
16	Alexa Robot	10
17	Mozilla/5.0 (Macintosh; Intel Mac OS X 10_9_2; ocrawler; omax7828@yahoo.com) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/36.0.1944.0 Safari/537.36	9
18	SafeSearch microdata crawler (https://safesearch.avira.com, safesearch-abuse@avira.com)	8
19	Powermarks	6
20	Mozilla/5.0 (compatible; pub-crawler; +http://wiki.github.com/bixo/bixo/bixocrawler; bixo-dev@yahoogroups.com)	6
21	AdnormCrawler www.adnorm.com/crawler	4
22	Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; Trident/5.0); 360Spider	2
23	YIJ-ASR/0.1 crawler (http://www.yahoo-help.jp/app/answers/detail/p/595/a_id/42716/)	1
24	Aboundex/0.3 (http://www.aboundex.com/crawler/)	1
25	maluuba-crawler/Nutch-1.6	1
26	Baidu Spider	1
27	Picsearch Robot	1
28	Screaming Frog SEO Spider/3.3	1
29	Xenu's Link Sleuth	1
30	Mozilla/5.0 (compatible; RSSMicro.com RSS/Atom Feed Robot)	1
	Total	3,898

Tabla 13: número de peticiones realizadas por el Bot de Bing

Con el objetivo de conocer los horarios en los cuales la mayoría de los usuarios visitan la plataforma Oferto, se analiza el grafico de “Actividad por hora del día” en donde es posible consultar las horas de conexión de manera acumulada distribuidas en las 24 horas del día. Se evidencia un incremento en las visitas a partir de las 8:00 a.m y dos picos de actividad a la 11:00 a.m y 4:00 p.m (Ver ilustración 13).

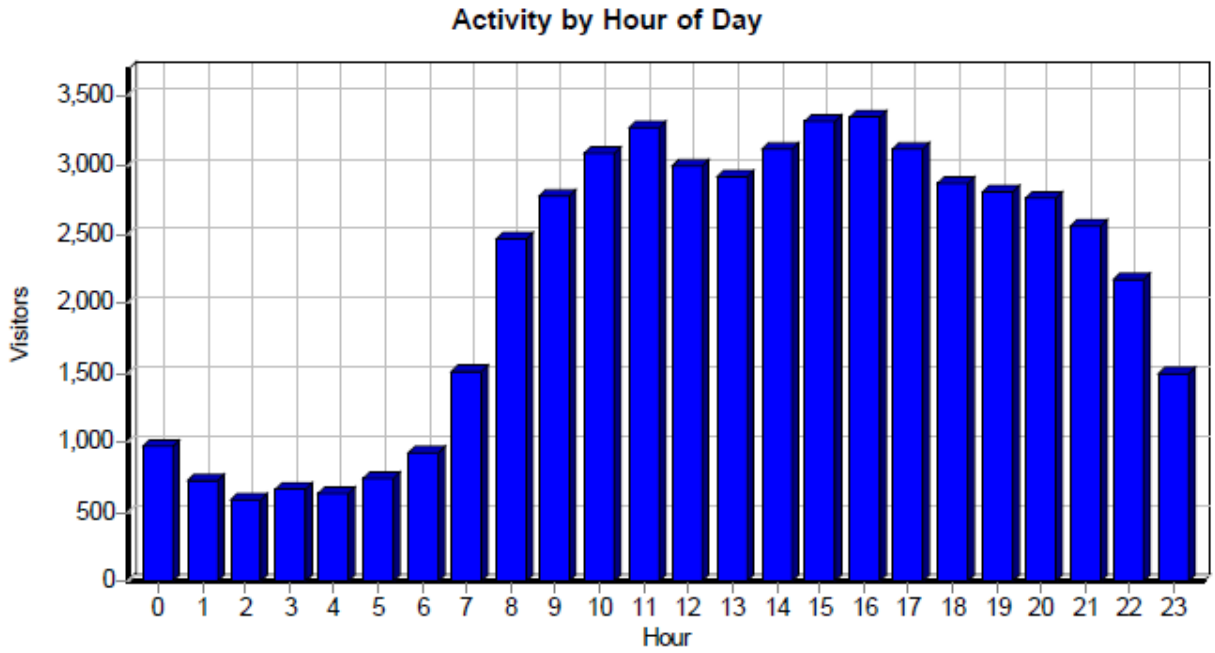
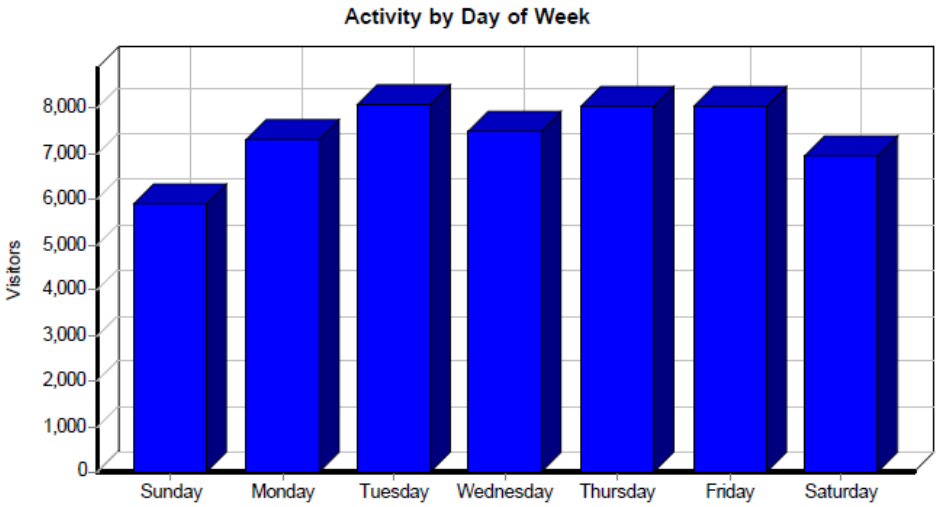


Ilustración 13: gráfico de actividad por hora del día

En cuanto a los días de la semana, se puede observar que todos los días de la semana tienen una actividad similar con excepción del día domingo en el cual se percibe una disminución en el número de visitantes.

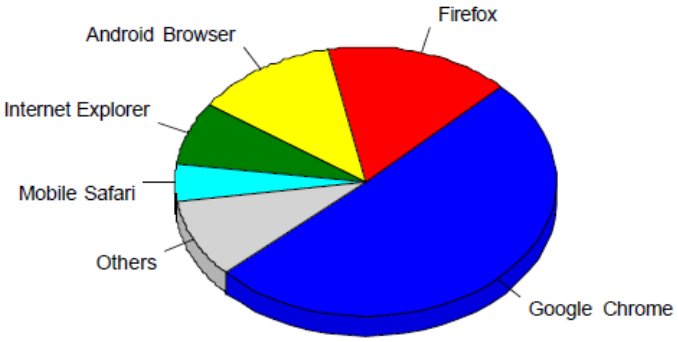


Activity by Day of Week

Day	Hits	Page Views	Visitors	Bandwidth (KB)
Sunday	44,723	43,968	5,908	1,008,309
Monday	65,865	65,405	7,314	1,559,058
Tuesday	83,098	82,497	8,075	1,658,471
Wednesday	68,298	67,409	7,506	1,642,721
Thursday	82,434	81,982	8,042	1,559,787
Friday	90,139	89,504	8,055	1,905,905
Saturday	71,175	70,630	6,955	1,362,941
Total	**505,732**	**501,395**	**51,855**	**10,697,196**

Tabla 14: actividad por día de la semana

En cuanto al uso de los navegadores por parte de los usuarios, se identifica que el más usado es Google Chrome, seguido por el navegador Firefox y Android Browser, tal como se aprecia a continuación:

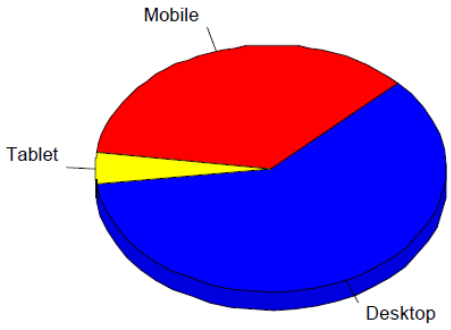


Most Used Browsers

	Browser	Hits	Visitors	% of Total Visitors
1	Google Chrome	280,119	26,949	50.52%
2	Firefox	50,421	8,357	15.67%
3	Android Browser	97,174	6,525	12.23%
4	Internet Explorer	16,155	4,011	7.52%
5	Mobile Safari	5,033	2,348	4.40%

Tabla 15: navegador más usado

En cuanto a los dispositivos de navegación, se encuentra que el más usado para ingresar en la plataforma es el PC con el 60.24% de los visitantes, seguido de los dispositivos móviles con un 35.74% (Ver tabla 16).



Device Types

	Device Type	Hits	Visitors	% of Total Visitors
1	Desktop	267,492	31,239	60.24%
2	Mobile	214,969	18,531	35.74%
3	Tablet	19,373	2,085	4.02%
	Total	**501,834**	**51,855**	**100.00%**

Tabla 16: dispositivo de navegación más usado

3.4.2.8 Normalizar horas

En el archivo de Log las horas son tomadas desde las 00:00:00 hasta las 23:59:59, sin embargo para el proceso de minería se ha decidido reemplazar los datos de las horas basados en las siguientes categorías:

Rangos horas	Categoría
00:00:00 – 05:59:59	Madrugada
06:00:00 – 11:59:59	Mañana
12:00:00 – 13:59:59	Medio día
14:00:00 – 18:59:59	Tarde
19:00:00 – 23:59:59	Noche

Tabla 17: Rangos de horas establecidos

Esto con el fin de poder realizar una agrupación de los visitantes de acuerdo a cada una de las categorías mencionadas anteriormente.

3.4.2.9 Normalización URLs

Para realizar el proceso de normalización fue necesario llevar a cabo la integración de datos como es estipulado en la metodología, ya que en la base de datos de la plataforma se encontraba información vital para determinar si la petición realizada pertenecía a una categoría, establecer las sesiones de usuario e identificar si llego a realizar la reserva de un producto perteneciente a una categoría especifica.

En algunas URL se cuenta con el id y el nombre del producto al que se accedía por ejemplo: 85.48.134.53 - - [08/Mar/2015:03:42:03 -0500] "GET /main-producto-id-595-t-lacoste_orane_3 HTTP/1.1" 200 54816 "https://www.google.es/" "Mozilla/5.0 (iPhone; CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4". Por lo que realizar la normalización es importante, ya que se busca trabajar por las categorías a las que pertenecen los productos, esto hizo que sea necesario realizar una consulta a la base de datos con el fin de obtener la categoría a la que pertenece cada

uno de los productos almacenados en ella, usando la siguiente consulta: SELECT productos.id_producto, productos.nombre, core_empresas.nombre, core_categorias.categoria FROM productos INNER JOIN core_empresas ON core_empresas.id_empresa= productos.id_empresa INNER JOIN core_categorias ON core_categorias.id_categoria= core_empresas.id_categoria.

Obteniendo con está 10.061 productos con datos como se representan en la siguiente tabla**.**

A	B	C	D
id_producto	nombre	nombre	categoria
75	Camaron Precocido Desvenado	Distribuidora Costa Azul	Viveres y Abarrotes
458	Filete de Basa	Distribuidora Costa Azul	Viveres y Abarrotes
2656	Bocachico x 500 gramos	Distribuidora Costa Azul	Viveres y Abarrotes
5315	Pollo Entero sin viscera	Distribuidora Costa Azul	Viveres y Abarrotes
101	Cafe de Origen La Nubia Vereda Morro Azul Sevilla Colombia	Cafe de Origen la Nubia	Viveres y Abarrotes
1952	Cafe de Origen La Nubia Vereda Morro Azul Sevilla Colombia	Cafe de Origen la Nubia	Viveres y Abarrotes
3228	Capuchino	Cafe de Origen la Nubia	Viveres y Abarrotes
1417	Tequila la leyenda del milagro	SUPERMERCADO LAURELES	Viveres y Abarrotes
1419	Whisky Buchanan's	SUPERMERCADO LAURELES	Viveres y Abarrotes
1420	Salmon Premium Marinus	SUPERMERCADO LAURELES	Viveres y Abarrotes
2924	productos de aseo	SUPERMERCADO LAURELES	Viveres y Abarrotes
3859	Desodorante AXE	SUPERMERCADO LAURELES	Viveres y Abarrotes
3860	Desodorante DOVE	SUPERMERCADO LAURELES	Viveres y Abarrotes
3991	whisky OLD PARR	SUPERMERCADO LAURELES	Viveres y Abarrotes
4233	Gaseosa postobon	SUPERMERCADO LAURELES	Viveres y Abarrotes
5115	Whisky Chivas Regal 12 aÃƒÂ±os	SUPERMERCADO LAURELES	Viveres y Abarrotes
6224	Whisky Chivas Regal 18 aÃƒÂ±os	SUPERMERCADO LAURELES	Viveres y Abarrotes
2667	Vino de Cereza la Casona 750 c.c.	VINOS DE BODEGA LA CASONA LTDA	Viveres y Abarrotes
3291	Vino de Cereza la Casona 1.750 c.c.	VINOS DE BODEGA LA CASONA LTDA	Viveres y Abarrotes

Tabla 18: Productos Base de datos

El paso siguiente fue determinar cómo se identificaban las categorías dentro del sitio web, posterior a un análisis de la taxonomía del mismo, se obtuvo que las categorías eran representadas en su URL de la siguiente manera. http://www.oferto.co/main-productos-c-31-t-hogar_electrodomesticos_y_oficina, donde t-hogar_electrodomesticos_y_oficina es el nombre de la categoría a la que había accedido directamente el visitante. Por lo que se realizó la siguiente consulta a la base de datos. Select id_categoria, categoria from core_categorias; con el fin de

identificar el nombre de todas las categorías que estaban creadas en el sitio web, obteniendo como resultado en la actualidad un total de 52 categorías que se mencionan a continuación:

Id_categoria	Categoría
1	Víveres y Abarrotes
5	Tecnología
7	Niños y Bebes
8	Mascotas
9	Moda
10	Música
11	Belleza y Cuidado Personal
12	Viajes y Turismo
13	Vehículos
15	Agro
18	Deportes
19	Gastronomía
22	Inmuebles
24	Servicios
26	Diseño, Arte y Decoración
27	Arquitectura y Construcción
28	Ferretería
29	Servicios Publicitarios
30	Fiestas y regalos
31	Hogar, Electrodomésticos y Oficina
33	Educación
36	Joyas y Accesorios
37	Seguros
38	Materiales Eléctricos
39	Varios
40	Juegos, Juguetes y Hobbies
41	Droguerías
42	Salud
43	Proveedores/Mayoristas
44	Servicios Empresariales
45	Institucional
46	Entretenimiento y Vida Nocturna
47	Productos Importados
48	Transporte

49	Alojamientos
50	Agencias de viaje
51	Guías de turismo
52	Casas de cambio

Tabla 19: Identificación de categorías en la base de datos

Esto permite definir que si dentro de una URL se encontraba una de las siguientes cadenas de texto:

- t-entretenimiento_y_vida_nocturna
- t-gastronomia
- t-hogar_electrodomesticos_y_oficina
- t-inmuebles
- t-joyas_y_accesorios
- t-moda
- t-salud_y_cuidado_personal
- t-servicios
- t-tecnologia
- t-vehiculos
- t-viajes_y_turismo
- t-viveres_y_abarrotes
- t_ninos_y_bebes
- t_mascotas
- t_musica
- t_agro
- t_deportes
- t_gastronomia

- t_diseno_arte_y_decoracion
- t_arquitectura_y_construccion
- t_ferreteria
- t_servicios_publicitarios
- t_fiestas_y_regalos
- t_hogar_electrodomesticos_y_oficina
- t_educacion
- t_seguros
- t_materiales_electricos
- t_varios
- t_juegos_juguetes_y_hobbies
- t_droguerias
- t_salud
- t_proveedores_y_mayoristas
- t_servicios_empresariales
- t_institucional
- t_entretenimiento_y_vida_nocturna
- t_productos_importados
- t_transporte
- t_alojamientos
- t_agencias_de_viaje
- t_guias_de_turismo
- t_casas_de_cambio

Estas pertenecían a una visita a la categoría como tal y no necesariamente a un producto específico.

Además de lo anterior en algunas URL se identifica que se accedió a uno de los mini sitios de las empresas de Oferto, por ejemplo: 66.249.73.157 - - [08/Mar/2015:03:20:51 -0500] "POST /producto-tabla_carrito HTTP/1.1" 200 1152 "http://www.dismodalc.oferto.co/main-productos-c-1095-pagina-1-orden-asc-por-nombre-cant-30" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)". Por lo que fue necesario obtener de la base de datos las empresas con su respectiva categoría, subdominio y dominio, para lo que fue necesario realizar la siguiente consulta. select id_empresa, id_categoria, subdominio, dominio, nombre from core_empresas; de la cual se obtuvo un archivo CSV, el cual cuenta con 1272 registros, de los cuales 7 de ellos no tenían nombre de empresa, por lo que fueron eliminados**.** Luego de esto se eliminaron aquellos registros que no contaban en el momento con una categoría, quedando así con 665 empresas que tenían una categoría asignada, su respectivo nombre, subdominio y/o dominio, para así realizar el cruce entre el id_categoria de la empresa y el id_categoria de la consulta hecha anteriormente, obteniendo de esta manera las categorías a las que pertenecen las empresas y poder finalmente determinar en el log a qué mini sitio se dirigía el visitante y este a cuál categoría pertenece, teniendo un resultado como el mostrado en la siguiente tabla:

A	B	C	D	E
id_empresa	nombre	dominio	subdominio	categoria
141	ARES CELL		arescell	Tecnologia
601	AVIMAR LA 7MA		avimarla7ma	Viveres y Abarrotes
112	BITCOM TIENDA TECNOLOGICA	bitcom.com.co	bitcomtecnotienda	Tecnologia
703	C.I. ENELAGRO S.A.S		heladosfrutpop	Viveres y Abarrotes
39	Cafe de Origen la Nubia		cafedeorigenlanubia	Viveres y Abarrotes
532	CAFE DONKAFE SAS		cafedonkafesas	Viveres y Abarrotes
856	COTIDIANO.ORG		cotidiano	Viveres y Abarrotes
38	Distribuidora Costa Azul		pescaderiacostaazul	Viveres y Abarrotes
769	EL PARAÃƒÂSO DEL CAFÃƒâ€°		elparaisodelcafe	Viveres y Abarrotes
886	Fama el Cordero		famaelcordero	Viveres y Abarrotes
727	FRUTAS Y VERDURAS KARINA MERCAR		frutasyverduraskarinamercar	Viveres y Abarrotes
19	Full Tecnologia		fulltecnologia	Tecnologia
785	Genocafe		genocafe	Viveres y Abarrotes
7	KIT VIRTUAL		kitvirtual	Tecnologia
714	LA GRANJA YARA-YARO S.A.S.	LaGranjaYaraYaro	lagranjayarayaro	Viveres y Abarrotes
8	Maxtech S.A.S		sutecnologia	Tecnologia
43	MULTITINTAS	multitintas.com.co	multitintas	Tecnologia
557	PLANTO SAS		plantosas	Viveres y Abarrotes
1	Rhiss.net		rhiss	Tecnologia
162	SUPERMERCADO LAURELES	supermercadolaureles.com	supermercadolaureles	Viveres y Abarrotes

Tabla 20: Información de empresas

Luego se identifica que hay muchas de las peticiones realizadas por empresarios que ingresan a su administrador del mini sitio a través de Oferto, por lo que se realiza un análisis de aquellas que cumplen con dicho criterio como por ejemplo: 181.132.14.2 - - [08/Mar/2015:03:58:33 -0500] "GET /login-usuario HTTP/1.1" 200 5620 "http://www.oferto.co/login-inicio" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36". Buscando así a través de estos registros determinar la cantidad de empresarios que pueden entrar a administrar sus sitios web. Para este proceso se tuvieron en cuenta las URL que dentro de su contenido involucraran algunas de las siguientes palabras:

- login-usuario
- login-inicio
- slide-editar
- usuario-config

- almacen-lista
- mailing-lista
- contenido-editar
- LOGIN-USUARIO
- producto-list_productos
- usuario-listar
- producto-list_ofertas
- almacen-editar
- almacen-lista
- categoria-categorias
- contenido-editar-t-1
- contenido-editar-t-2
- contenido-editar-t-3
- /mailing-editar
- /mailing-enviar
- /mailing-json_listClientes
- /mailing-lista

Esto con el fin de quitarlos de los archivos, para evitar tener información que no hace parte de los objetivos planteados, quedando solamente aquellas peticiones hechas por los visitantes de la plataforma que buscaban productos como tal.

3.4.2.10 Consolidación de datos

Luego de realizado el proceso de limpieza e integración se estructurara el data warehouse de tal manera que facilite el procesamiento de la información usando los métodos estadísticos y de máquina de aprendizaje, los cuales serán utilizados para extraer patrones de comportamiento de

los visitantes; inicialmente se cuenta con un archivo .csv que fue generado por el código de limpieza utilizado con anterioridad, dicho archivo tiene la estructura que se visualiza en la siguiente tabla.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	85.48.134.53	Usuario 1	Sesion 1	N/A	Deportes	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
2	85.48.134.53	Usuario 1	Sesion 1	N/A	Deportes	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
3	85.48.134.53	Usuario 1	Sesion 1	N/A	N/A	N/A	FALSO	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
4	85.48.134.53	Usuario 1	Sesion 1	Deportes	N/A	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
5	85.48.134.53	Usuario 1	Sesion 1	Deportes	N/A	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
6	85.48.134.53	Usuario 1	Sesion 1	Deportes	N/A	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
7	85.48.134.53	Usuario 1	Sesion 1	Deportes	N/A	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
8	85.48.134.53	Usuario 1	Sesion 1	Deportes	N/A	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
9	85.48.134.53	Usuario 1	Sesion 1	Deportes	N/A	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
10	85.48.134.53	Usuario 1	Sesion 1	Deportes	N/A	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
11	85.48.134.53	Usuario 1	Sesion 1	Deportes	N/A	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
12	85.48.134.53	Usuario 1	Sesion 1	N/A	Deportes	N/A	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.7						
13	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						
14	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						
15	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						
16	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						
17	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						
18	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						
19	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						
20	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						
21	181.132.14.2	Usuario 2	Sesion 2	N/A	N/A	N/A	FALSO	Madrugada	181.132.14.2	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.76 Safari/537.36"						

Tabla 21: estructura de archivo .csv

Este archivo tiene un total de 481.525 registros. Luego de esto se procede a realizar una normalización de la estructura con la que cuenta en esté archivo, buscando de esta manera eliminar todos aquellos registros que no cumplen con la condición de que tenga una categoría asociada a la petición, estas están identificadas en las columnas D, E, F de la anterior imagen. Quedando luego de este procedimiento un total de 63.034 registros como total, luego de esto se unifican las columnas mencionadas anteriormente, obteniendo como resultado el nombre de la categoría visitada en cada uno de los registros consolidado en la columna C. Como se muestra a continuación:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
IP	Usuario	Sesion	Categoria	Jornada	Peticion										
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
85.48.134.53	Usuario 1	Sesion 1	Deportes	Madrugada	85.48.134.53	CPU iPhone OS 8_1_3 like Mac OS X) AppleWebKit/600.1.4 (KHTML, like Gecko) CriOS/40.0.2214.73 Mobile/12B466 Safari/600.1.4"									
186.85.141.139	Usuario 3	Sesion 6	Moda	Madrugada	186.85.141.1	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.115 Safari/537.36"									
186.85.141.139	Usuario 3	Sesion 6	Moda	Madrugada	186.85.141.1	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.115 Safari/537.36"									
186.85.141.139	Usuario 3	Sesion 6	Moda	Madrugada	186.85.141.1	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.115 Safari/537.36"									
186.85.141.139	Usuario 3	Sesion 6	Moda	Madrugada	186.85.141.1	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.115 Safari/537.36"									
186.85.141.139	Usuario 3	Sesion 6	Moda	Madrugada	186.85.141.1	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.115 Safari/537.36"									
186.85.141.139	Usuario 3	Sesion 6	Moda	Madrugada	186.85.141.1	WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.115 Safari/537.36"									
186.113.157.158	Usuario 18	Sesion 24	Moda	MaÃ±ana	186.113.157.	Android 4.4. GT-I9195 Build/KOT49H) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.109 Mobile Safari/537.36"									
186.113.157.158	Usuario 18	Sesion 24	Moda	MaÃ±ana	186.113.157.	Android 4.4. GT-I9195 Build/KOT49H) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/40.0.2214.109 Mobile Safari/537.36"									

Tabla 22: Consolidado categorías visitadas

A continuación, se procede a estructurar el archivo con las siguientes características:

- Columna A: IP
- Columna B: Categoría
- Columna C: Jornada
- Columna D: Fecha
- Columna E: NumeroDia
- Columna F: NombreDia

Quedando el archivo .csv así:

A	B	C	D	E	F
IP	Categoria	Jornada	Fecha	NumeroDia	NombreDia
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
85.48.134.53	Deportes	Madrugada	08-mar-15	8	domingo
186.85.141.139	Moda	Madrugada	08-mar-15	8	domingo
186.85.141.139	Moda	Madrugada	08-mar-15	8	domingo
186.85.141.139	Moda	Madrugada	08-mar-15	8	domingo
186.85.141.139	Moda	Madrugada	08-mar-15	8	domingo
186.85.141.139	Moda	Madrugada	08-mar-15	8	domingo
186.85.141.139	Moda	Madrugada	08-mar-15	8	domingo
186.113.157.158	Moda	Manana	08-mar-15	8	domingo
186.113.157.158	Moda	Manana	08-mar-15	8	domingo

Tabla 23: archivo .csv consolidado

3.5 Fase IV. Modelado

En las fases anteriores se ha descrito cada uno de los pasos necesarios desde que se reciben los datos hasta terminar todo el proceso para dejarlos listos con el fin de aplicar modelos descriptivos y predictivos. Por lo que ahora nos enfocaremos en la Fase de Modelado en la que se utilizarán diferentes técnicas de aprendizaje automático. En esta fase puntualmente se buscará desarrollar modelos que generalicen la estructura con la cuentan dichos datos. En este caso en particular se buscará encontrar respuestas que permitan el mejoramiento de la forma en la que se muestra la información actualmente en la plataforma, basados en los días y jornadas en los que visitan las categorías. Para iniciar la etapa de modelado serán tenidas como bases las preguntas determinadas en la Fase I. Comprensión del negocio, ya que es allí donde se establecieron

aquellas incógnitas a las que se busca dar respuesta con el proceso que se está llevando a cabo, es por esto que los datos con lo que se cuenta después de realizada la fase III. Preparación de los datos, se utilizaran de acuerdo a su relación con las preguntas que se desean responder.

3.5.1 Datos disponibles

Luego de tener el archivo .csv se cuenta con 6 datos que se utilizaran para realizar el proceso de minería, estos datos se describen a continuación:

Nombre	Descripción	Tipo Atributo
IP	Identifica el visitante que realizo la petición.	String
Categoría	Identifica cuál de las categorías mencionadas en la Fase III. Preparación de los datos, fue a la que accedió el visitante en esa petición.	Nominal
Jornada	Permite determinar el momento del día, basados en las distribuciones propuestas en la Fase III. Preparación de los datos, en que el visitante accedió a esa categoría.	Nominal
Fecha	Permite conocer la fecha exacta en la que el visitante	Date

	realizó la petición.	
NumeroDia	Determina el número del día del mes en el que realizo la petición	Nominal
NombreDia	Permite identificar el día de la semana en el que se realizó la petición.	Nominal

Tabla 24: datos disponibles

3.5.2 Análisis preliminar de los datos

Inicialmente se construye el archivo .arff que permitirá usar el data Warehouse en el software Weka, en este proceso se determina eliminar el atributo fecha, debido a que no representa algún valor para dar respuestas a las preguntas del negocio establecidas, quedando su estructura como se ve en la siguiente tabla:

```
@relation relation-weka.filters.unsupervised.attribute.StringToNominal-R6-weka.filters.unsupervised.attribute.
          NumericToNominal-R5-weka.filters.unsupervised.attribute.Remove-R4

@attribute IP string

@attribute Categoria {JuegosJuguetesyHobbies,Educacion,Varios,ProductosImportados,
                      MaterialesElectricos,NinosyBebes,Seguros,Reserva,Alojamientos,
                      Deportes,registroReserva,Moda,Droguerias,ArquitecturayConstruccion,
                      DisenoArteyDecoracion,BellezayCuidadoPersonal,Salud,
                      EntretenimientoyVidaNocturna,Ferreteria,Gastronomia,
                      HogarElectrodomesticosyOficina,Inconsistencia,Inmuebles,JoyasyAccesorios,
                      JuguetesNinosyBebes,Mascotas,Musica,Servicios,ServiciosPublicitarios,
                      Tecnologia,Vehiculos,ViajesyTurismo,ViveresyAbarrotes,Agenciasdeviaje,Fiestasyregalos,Agro}

@attribute Jornada {Madrugada,Manana,MedioDia,Tarde,Noche}

@attribute NumeroDia {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 NombreDia {domingo,lunes,martes,miercoles,jueves,viernes,sabado}

@data
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
85.48.134.53,Deportes,Madrugada,8,domingo
```

Tabla 25: estructura de archivo .arff

Luego en weka se llevó a cabo un análisis básico de los atributos, como se puede evidenciar a continuación.

IP: Al ser un atributo de tipo String, no se evidencia ningún tipo de descripción, teniendo en cuenta que toma cada uno de los registros como únicos, como se muestra a continuación.

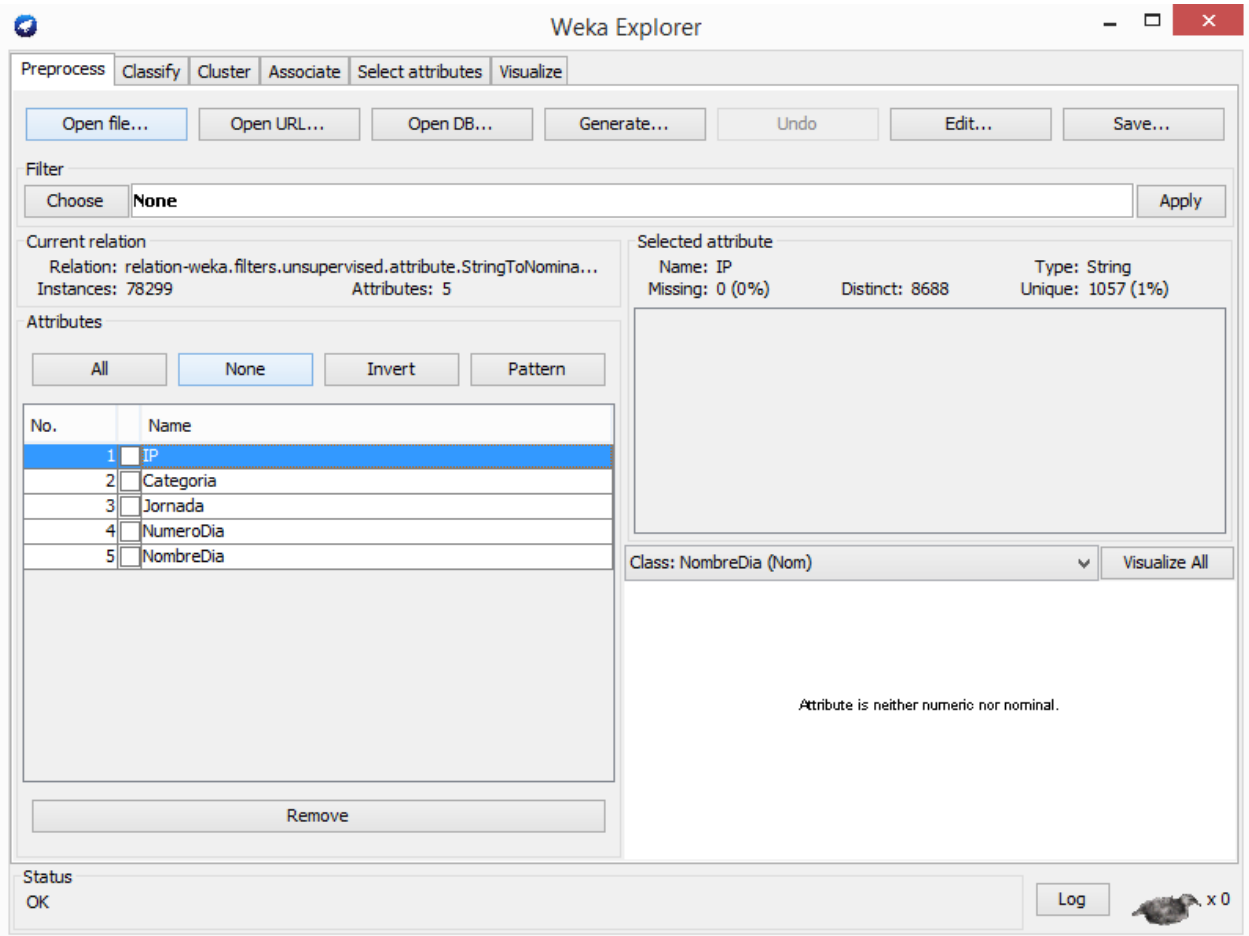


Ilustración 14: atributo tipo string

Por lo que se decide convertir a nominal a través del filtro de weka llamado StringToNominal, que permite convertir un atributo de tipo String a un tipo nominal, para lo que primero se selecciona dicho filtro como se evidencia a continuación

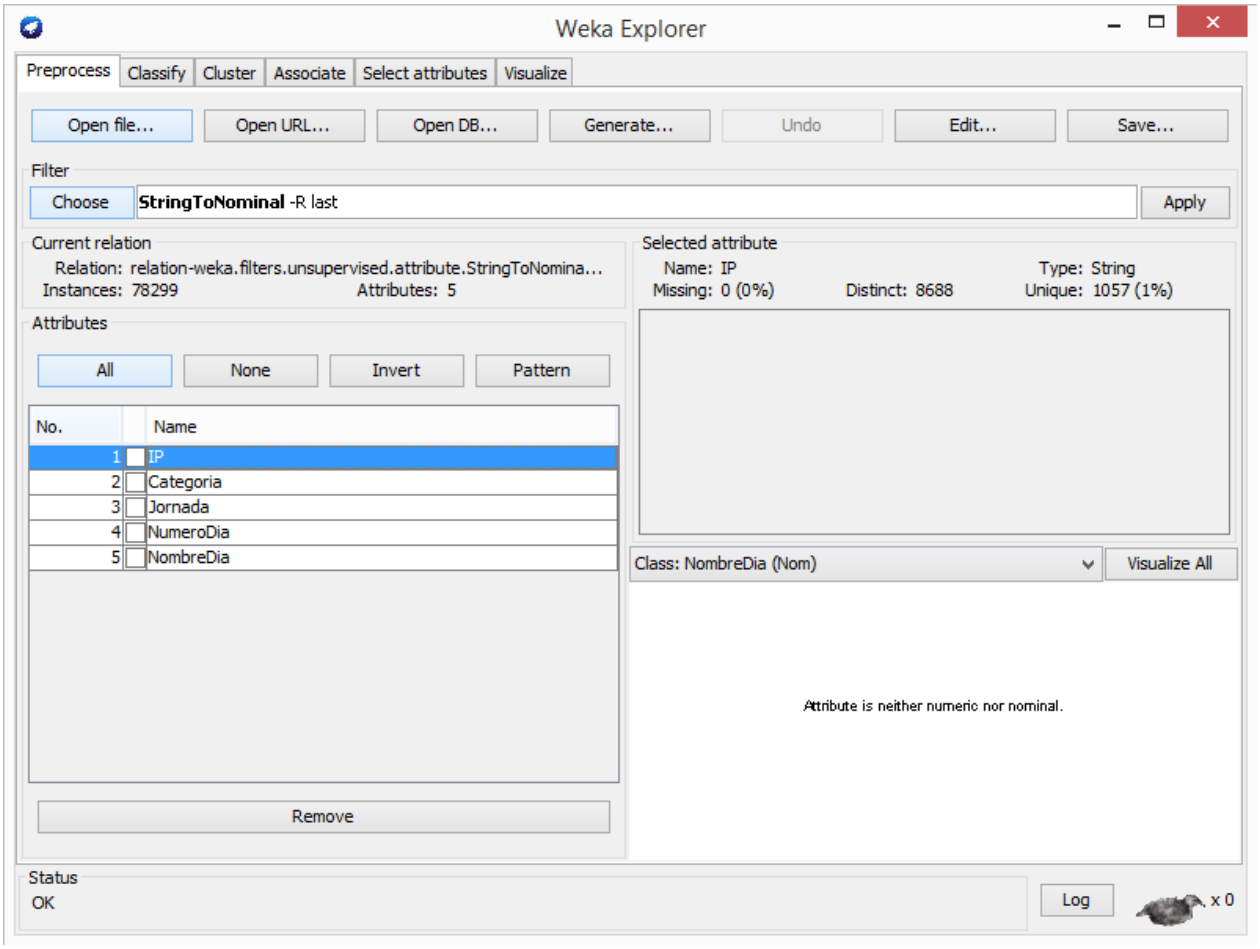


Ilustración 15: atributo tipo string to nominal

Luego, se selecciona el atributo que desea ser convertido, el que para este caso es el número 1 de la lista de atributos (IP)

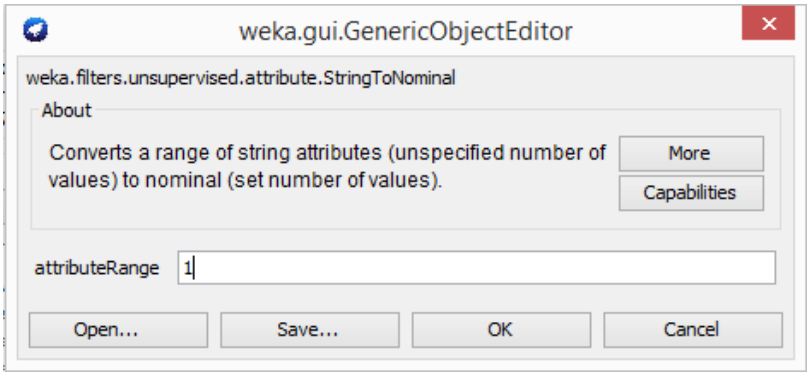


Ilustración 16: atributo a convertir

Y por último se aplica el filtro, obteniendo como resultado, la cantidad de visitas hechas por cada uno de los visitantes de la plataforma, sin embargo al ser muchos, no se visualiza gráfica para este atributo.

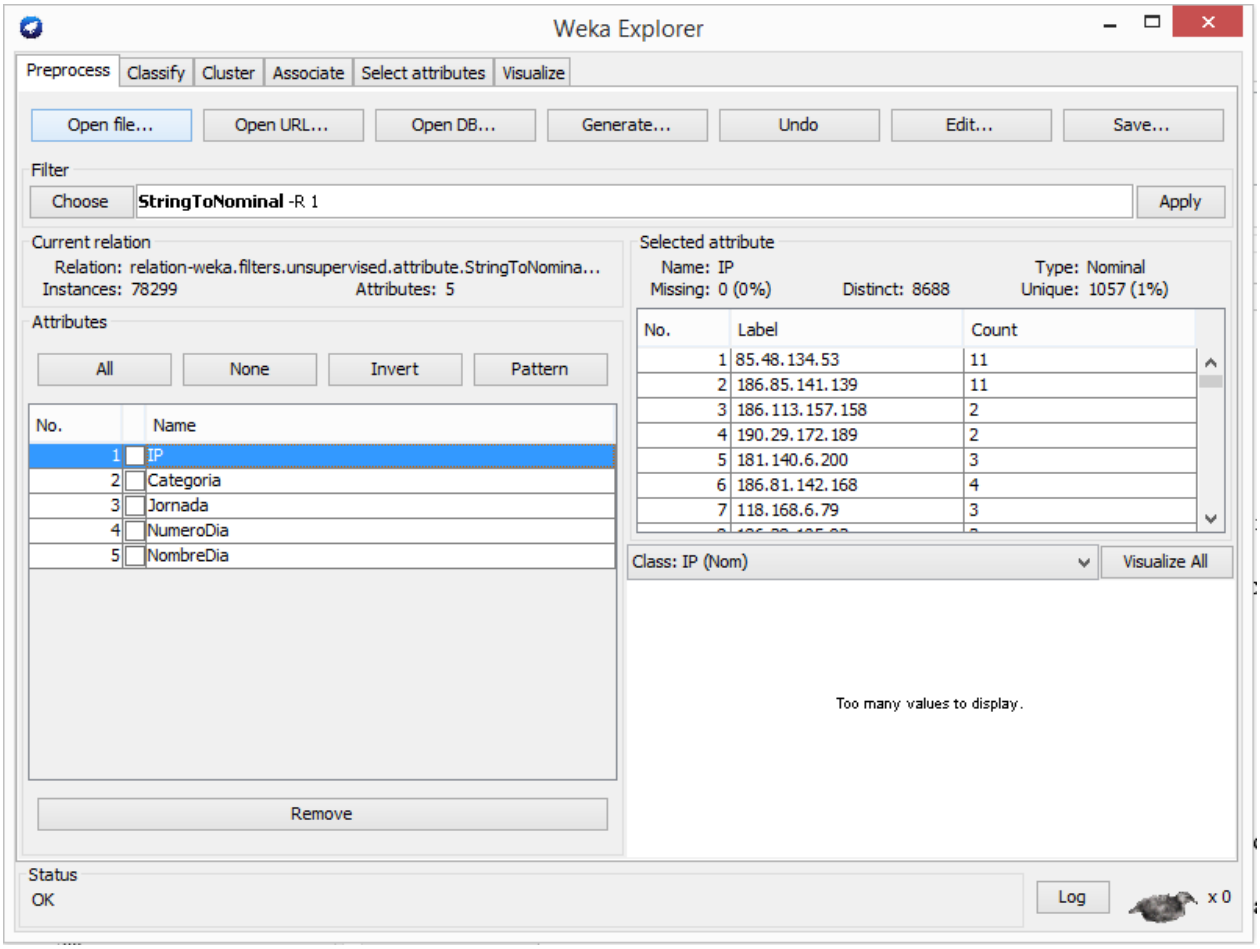


Ilustración 17: cantidad de visitas a la plataforma

Categoría: Se evidencia la sumatoria de cada una de las categorías con las que se cuenta en el archivo .arff, pudiendo realizar un análisis básico de dicho dato, como lo vemos a continuación.

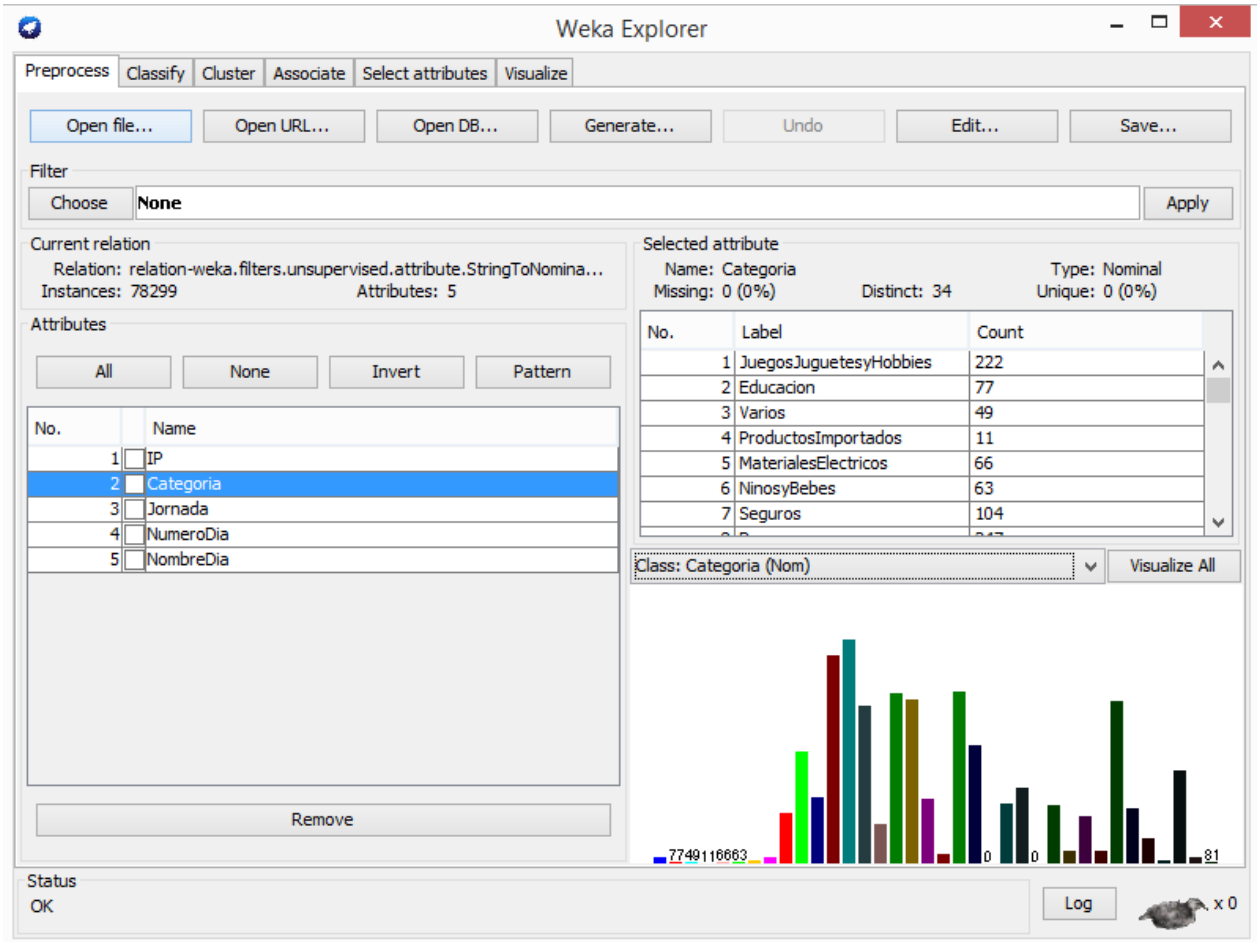


Ilustración 18: categorías

Determinando inicialmente que la categoría “Droguerías” es la que cuenta con mayor número de visitas, seguida de moda y con valores muy cercanos por gastronomía, tecnología, belleza y cuidado personal.

Jornada: similar al anterior, nos permite realizar un análisis básico acerca del comportamiento de las visitantes en cuanto a la jornada en la que ejecutan sus visitas.

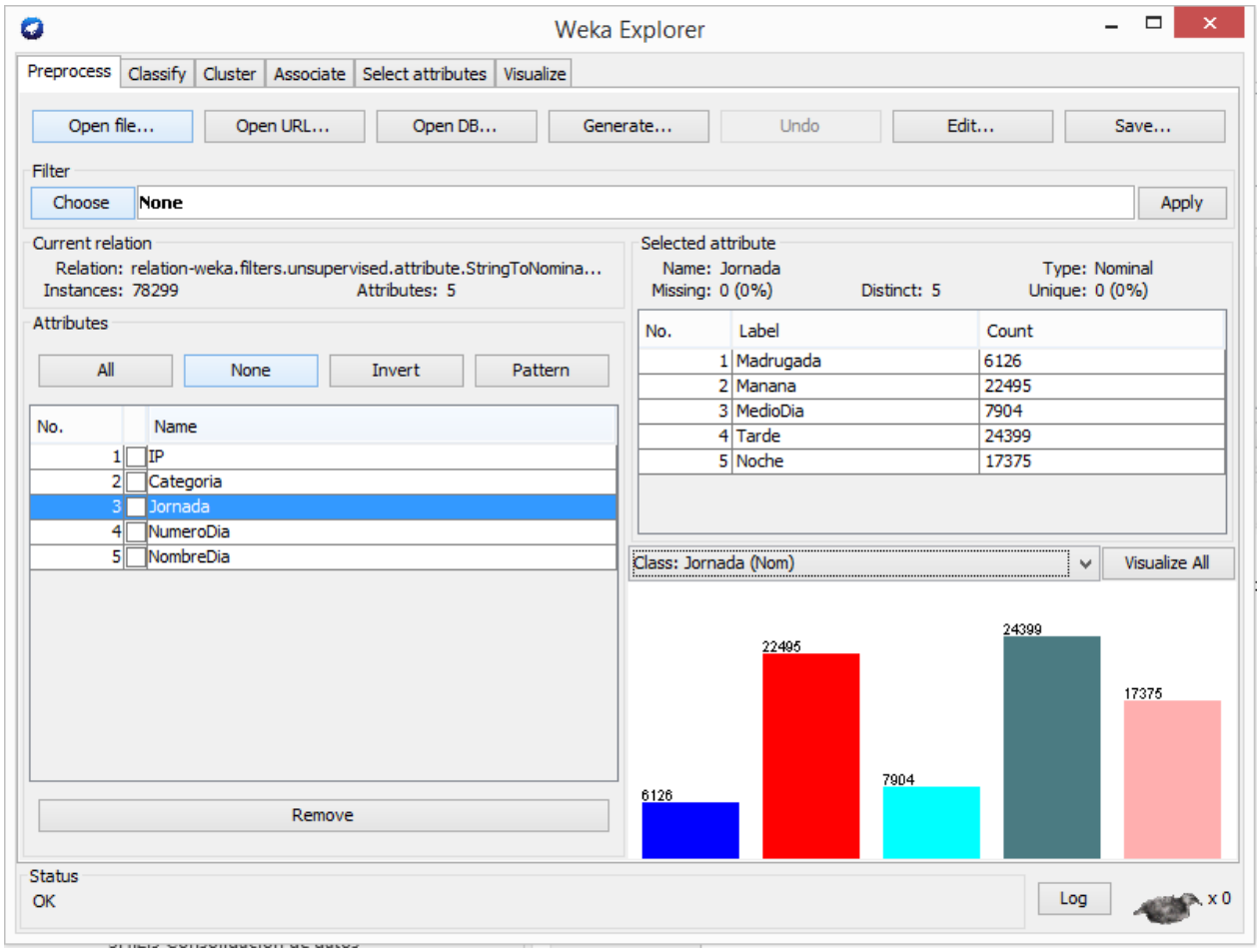


Ilustración 19: jornada

Teniendo la jornada de la tarde con el mayor número de visitas, seguido de la mañana y la noche, dejando así por último las jornadas de medio día y madrugada.

Número/Día: en este caso, se debe aclarar que el atributo NúmeroDía se decidió manejar como nominal, con el fin de manejar las visitas por día, sin importar el mes en el que fueron realizadas.

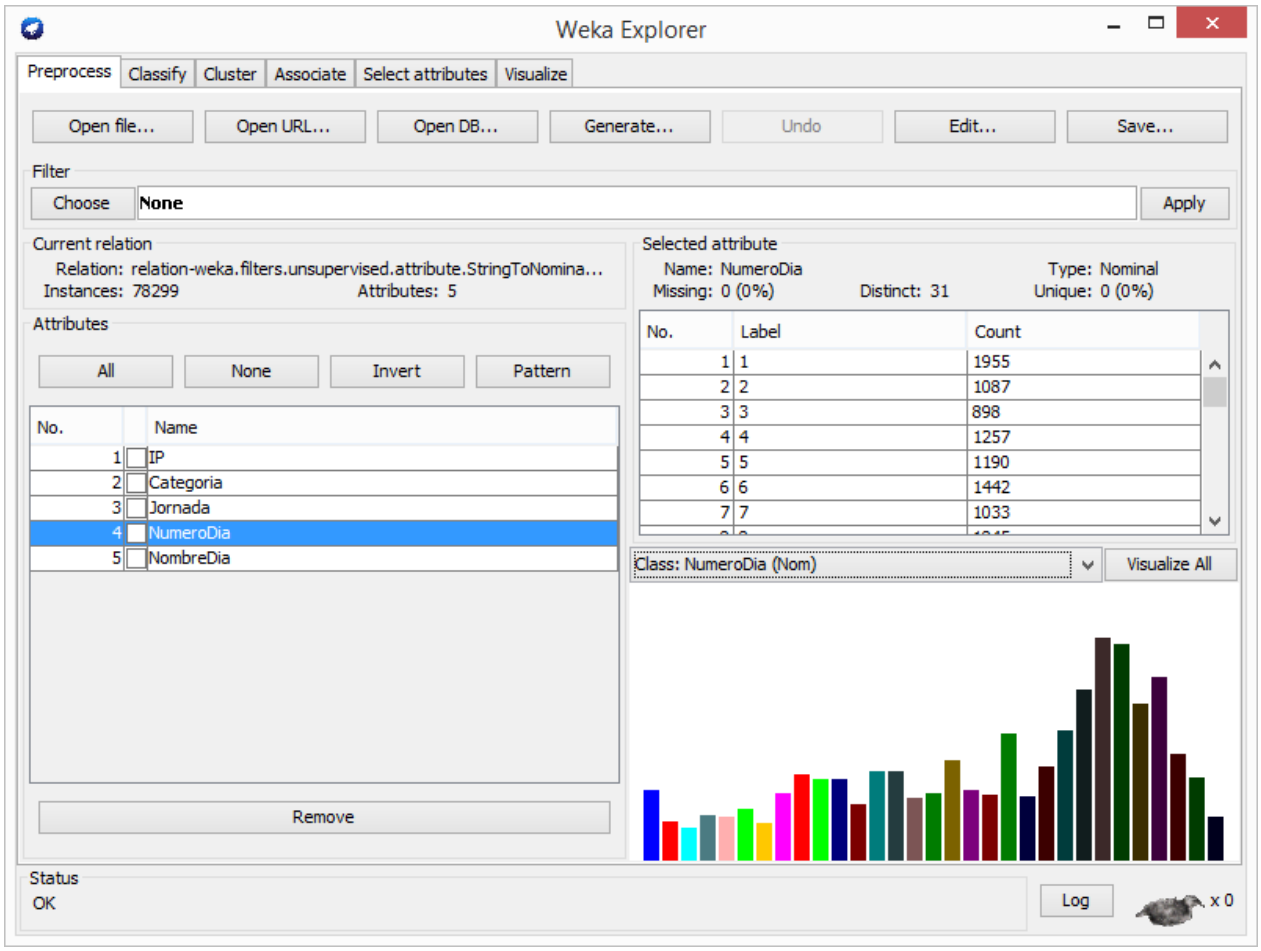


Ilustración 20: número/día

Encontrando poca variabilidad en los valores entre los días 1 y 19 y un incremento de visitas en las fechas cercanas al fin de mes.

NombreDia: este atributo también será manejado como nominal, buscando encontrar comportamientos de los visitantes de acuerdo a los días de la semana, por lo que se determinaron los días de cada una de las visitas hechas, como se observa a continuación.

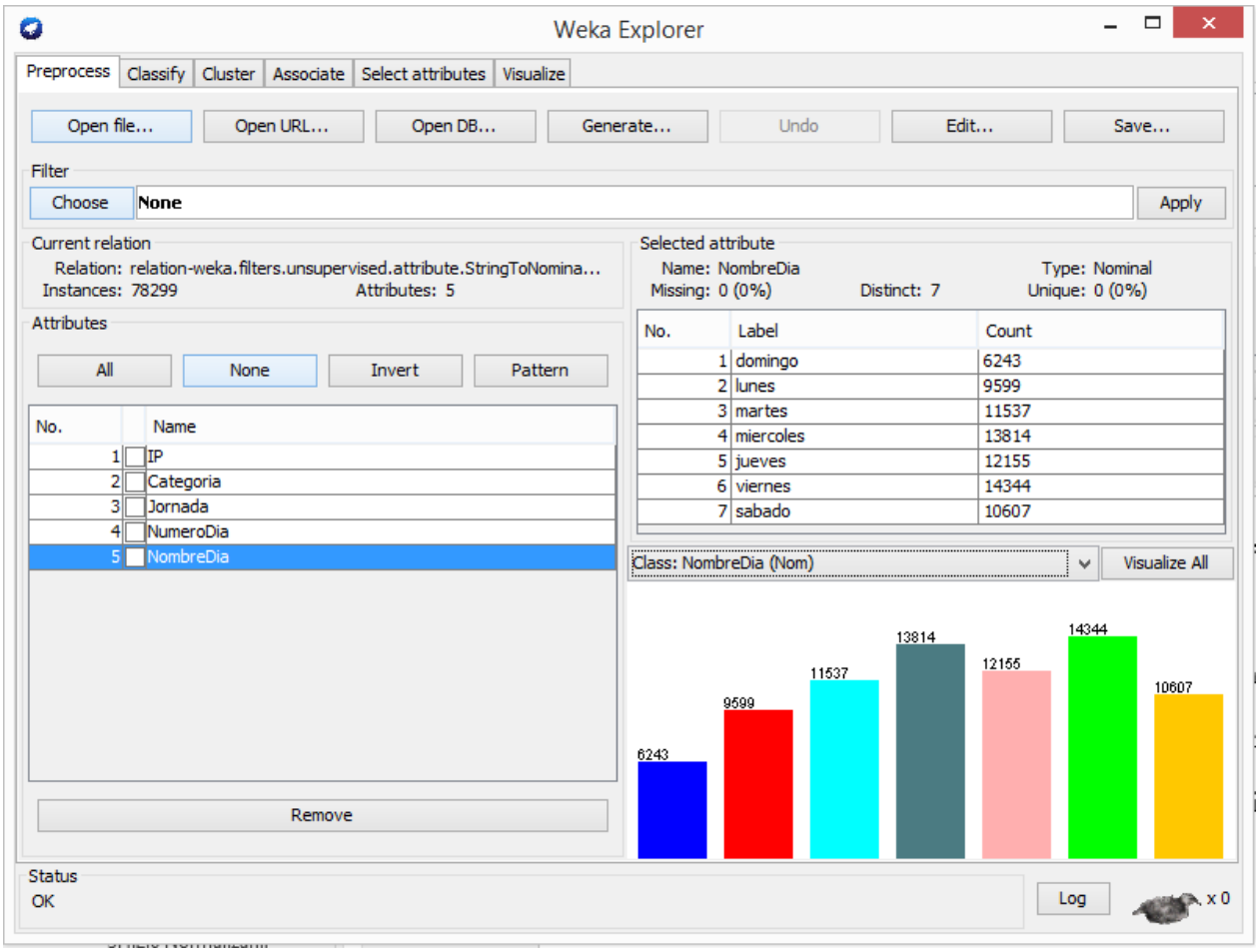


Ilustración 20: NombreDía

Encontrando inicialmente que el día de la semana que más cuenta con visitas es el viernes, seguidos por el miércoles y jueves, teniendo estos valores muy cercanos al sábado, dejando como los días de menos visitas los lunes y domingos, que son el inicio de la semana.

Para tener una mejor lectura de la información, se llevó este *dataset a Watson Analytics*, pudiendo describir puntualmente cada uno de los atributos como se muestra a continuación.

- Categorías

Ilustración 22: ambiente de categorías

En donde podemos ver que tal como se describió con anterioridad, las categorías más visitadas son aquellas que su nombre resalta sobre las demás, por ejemplificar droguerías, moda, belleza y cuidado personal, salud, tecnología y arquitectura y construcción, contrario a mascotas, viajes y turismo, vehículos e inmuebles que están dentro de las menos visitadas.

- Jornada

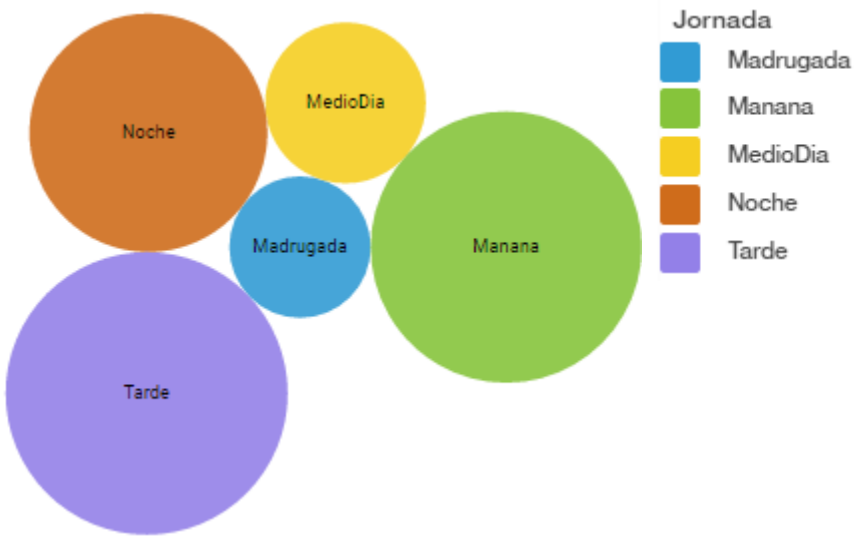


Ilustración 23: frecuencia Jornada

Se identifica que la tarde y la mañana son las jornadas que lideran el número de visitas en la plataforma, seguidas por la noche y muy lejos de ellas el medio día y la madrugada.

- NombreDia

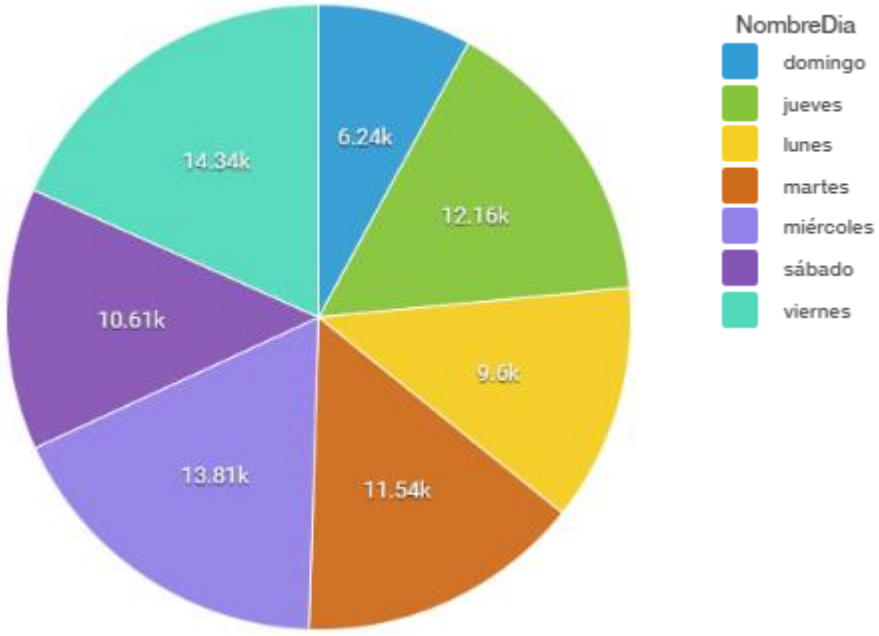


Ilustración 24: frecuencia NombreDía

Pudiendo identificar que el viernes es el día en el que mayor cantidad de visitas se tiene en la plataforma, seguido del miércoles, mientras que claramente se ve que el domingo es un día con muy pocos visitantes.

NumeroDía

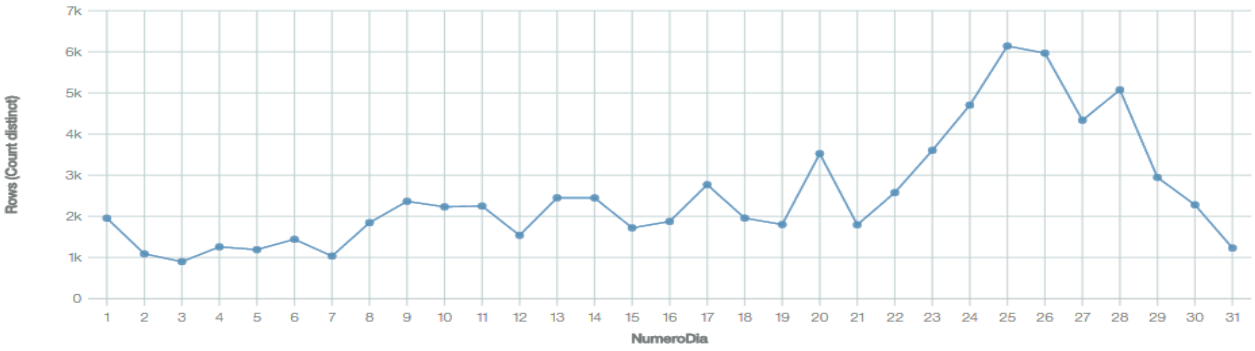


Ilustración 25: gráfico número/día

Permite visualizar que la mayor cantidad de visitas que se hacen a la plataforma son entre los días 23 y 29 del mes, teniendo en cuenta que dichos datos se trataron de manera nominal, con el fin de realizar un análisis del día sin importar el mes en el que se realizó la visita.

3.5.3 Selección, generación y ejecución de modelos

3.5.3.1 Clasificación

En la clasificación se pretende llevar a cabo un análisis de la información con la que se cuenta luego de haber realizado la limpieza de los datos y empezar a describir resultados que permitan identificar los visitantes de la plataforma y sus comportamientos. Para esta clasificación serán tenidos en cuenta los atributos Jornada, NombreDia y Categoria como se ve en la siguiente imagen.

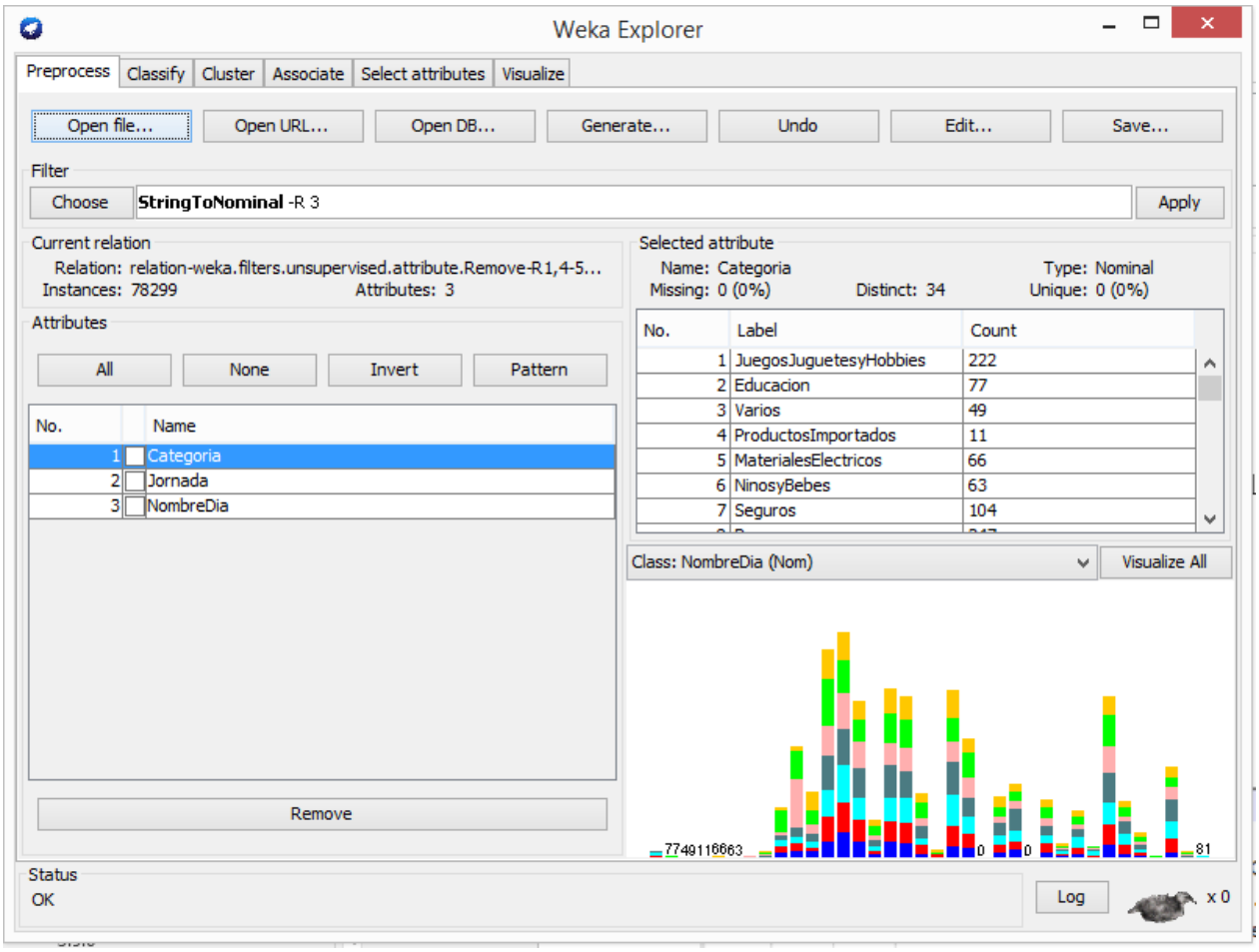


Ilustración 26: clasificación

3.5.3.1.1 Aplicación J48

El algoritmo J48 de WEKA es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos que más se ha utilizado en multitud de aplicaciones.

En este paso se llevó a cabo la aplicación del algoritmo **J48**, obteniendo los siguientes resultados:

```
=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     relation-weka.filters.unsupervised.attribute.
              Remove-R1-weka.filters.unsupervised.attribute.
              Remove-R3-weka.filters.unsupervised.attribute.
              StringToNominal-R4-weka.filters.unsupervised.
              attribute.NumericToNominal-R3-weka.filters.
              unsupervised.attribute.Remove-R3

Instances:    78299
Attributes:   3
              Categoria
              Jornada
              NombreDia
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

NombreDia = domingo
|   Jornada = Madrugada: Droguerias (528.0/456.0)
|   Jornada = Manana: BellezayCuidadoPersonal (1286.0/1154.0)
|   Jornada = MedioDia: Salud (756.0/625.0)
|   Jornada = Tarde: Droguerias (1930.0/1661.0)
|   Jornada = Noche: Droguerias (1743.0/1422.0)
NombreDia = lunes
|   Jornada = Madrugada: Tecnologia (438.0/381.0)
|   Jornada = Manana: Agenciasdeviaje (2959.0/2588.0)
|   Jornada = MedioDia: Droguerias (928.0/797.0)
|   Jornada = Tarde: Droguerias (3144.0/2755.0)
|   Jornada = Noche: Salud (2130.0/1864.0)
NombreDia = martes
|   Jornada = Madrugada: Vehiculos (623.0/495.0)
|   Jornada = Manana: Droguerias (3519.0/3183.0)
|   Jornada = MedioDia: Gastronomia (1152.0/1024.0)
|   Jornada = Tarde: Droguerias (4303.0/3716.0)
|   Jornada = Noche: Droguerias (1940.0/1665.0)
NombreDia = miercoles
|   Jornada = Madrugada: Moda (1489.0/1290.0)
|   Jornada = Manana: JoyasyAccesorios (4121.0/3513.0)
|   Jornada = MedioDia: BellezayCuidadoPersonal (1416.0/1236.0)
|   Jornada = Tarde: Gastronomia (4566.0/4113.0)
|   Jornada = Noche: Gastronomia (2222.0/1971.0)
```

```
NombreDia = jueves
|   Jornada = Madrugada: Droguerias (832.0/598.0)
|   Jornada = Manana: ArquitecturayConstruccion (3135.0/2718.0)
|   Jornada = MedioDia: BellezayCuidadoPersonal (930.0/815.0)
|   Jornada = Tarde: Droguerias (3539.0/3145.0)
|   Jornada = Noche: Deportes (3719.0/2306.0)
NombreDia = viernes
|   Jornada = Madrugada: Deportes (1518.0/642.0)
|   Jornada = Manana: Moda (4289.0/3709.0)
|   Jornada = MedioDia: Moda (1541.0/1332.0)
|   Jornada = Tarde: Moda (4192.0/3533.0)
|   Jornada = Noche: Alojamientos (2804.0/2174.0)
NombreDia = sabado
|   Jornada = Madrugada: BellezayCuidadoPersonal (698.0/593.0)
|   Jornada = Manana: Gastronomia (3186.0/2684.0)
|   Jornada = MedioDia: Salud (1181.0/1043.0)
|   Jornada = Tarde: Moda (2725.0/2436.0)
|   Jornada = Noche: registroReserva (2817.0/2447.0)

Number of Leaves  :     35

Size of the tree :  43
```

Ilustración 27: resultados algoritmo **J48**

Descritos en las siguientes ramas del árbol que nace del atributo NombreDia

- Domingo

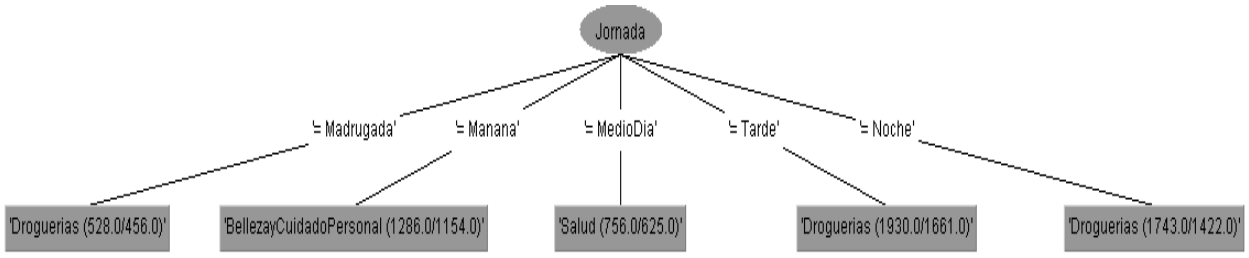


Ilustración 28: atributo nombre/día domingo

- Lunes

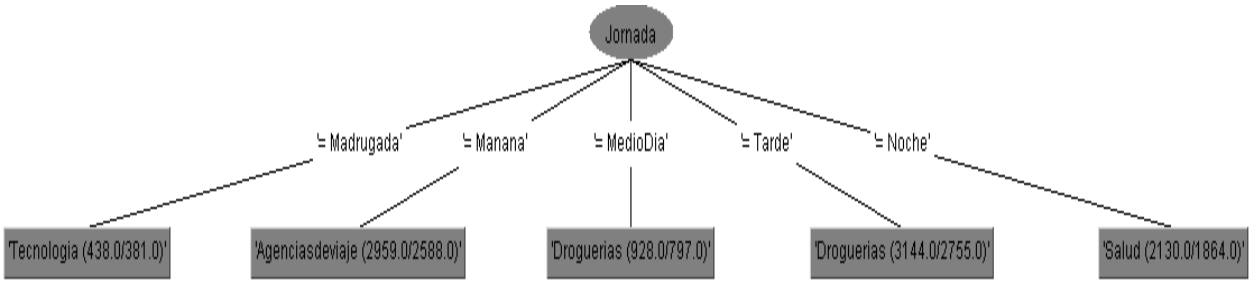


Ilustración 29: atributo nombre/día lunes

- Martes

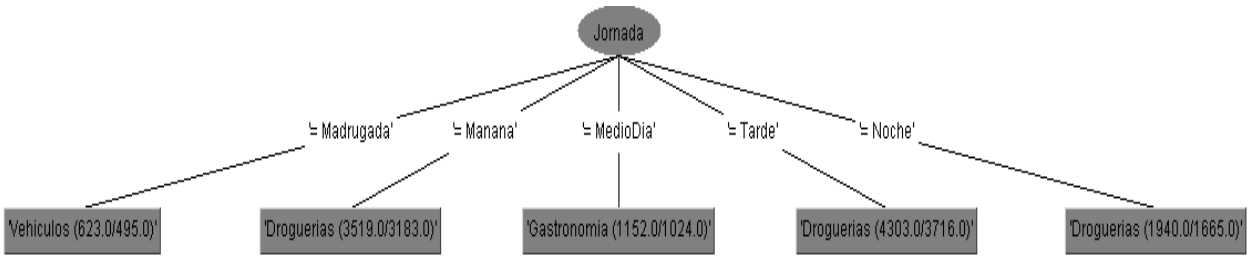


Ilustración 30: atributo nombre/día martes

- Miércoles

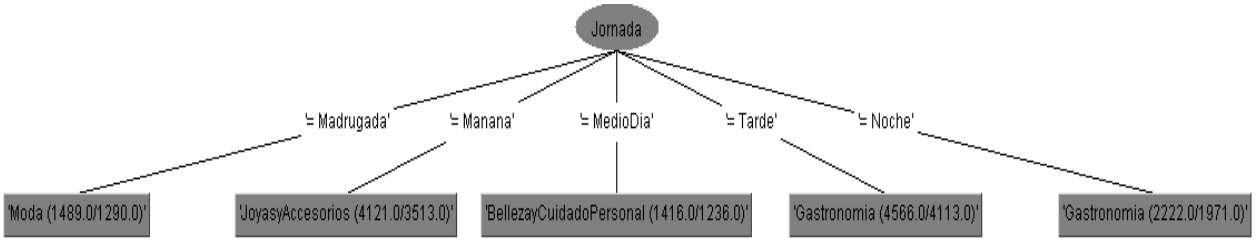


Ilustración 31: atributo nombre/día miércoles

Jueves

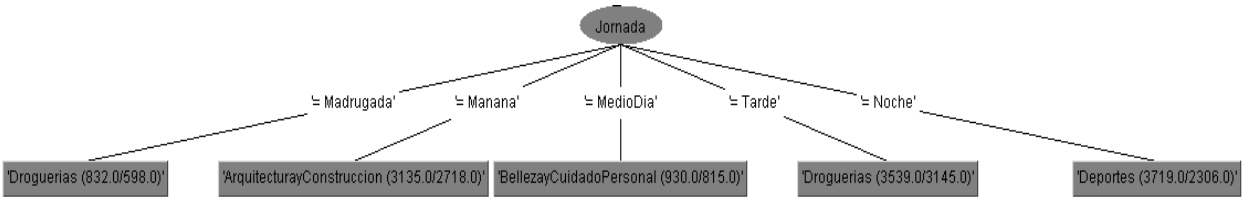


Ilustración 32: atributo nombre/día jueves

- Viernes

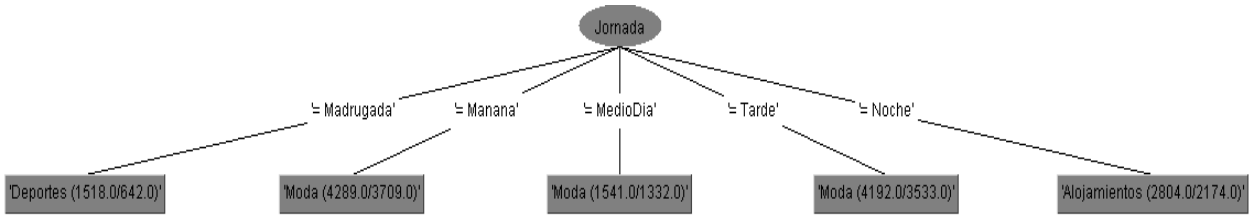


Ilustración 33: atributo nombre/día viernes

- Sábado

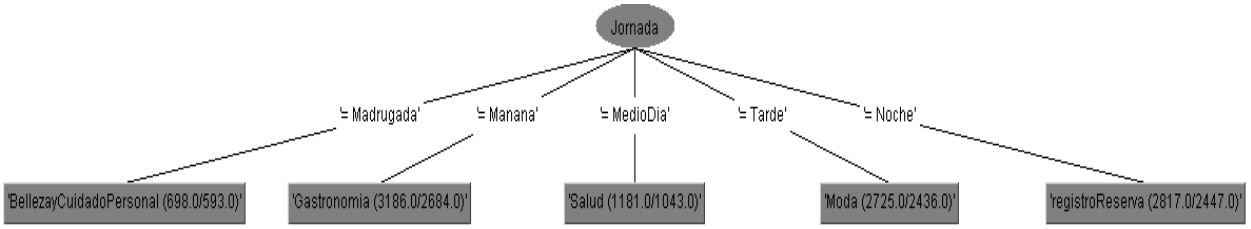


Ilustración 34: atributo nombre/día sábado

Encontrando en cada una de estas ramas del árbol una descripción de los comportamientos de los visitantes en cuanto a las categorías que visitan con respecto al día de la semana y la jornada. Luego de esto se buscó llevar la descripción de esto gráficamente a través de Watson Analytics, logrando cada uno de los siguientes resultados:

3.5.3.1.2 Visitas a Categorías por Jornada y NombreDia

En cada una de las siguientes imágenes podemos identificar el comportamiento de los visitantes en cuanto a las categorías que más visitan dependiendo del día de la semana y la jornada teniendo en cuenta que las categorías estarán determinadas por la siguiente tabla de colores.

Ilustración 35: visitas a categorías

- Domingo

Ilustración 36: visualización visitas domingo

El domingo se visualiza que la jornada en la que más visitas se obtienen es en la tarde, seguida por la noche, mañana y el medio día y la madrugada cuentan con valores similares, además de esto se identifica que la categoría droguerías es una de las más visitadas en este día.

- Lunes

Ilustración 37: visualización visitas lunes

En el lunes se ve como la tarde es la jornada más visitada con droguerías y moda como sus principales categorías, seguida por la mañana, teniendo a agencias de viajes, tecnologías y de nuevo droguerías como las categorías con más visitas, dejando a la noche en tercer lugar con salud y una vez más a droguerías.

- Martes

Ilustración 38: visualización visitas martes

En los martes se logra identificar que las jornadas se ubican en primer lugar por la tarde, seguida por la mañana y la noche, teniendo las tres como una de la categorías mas visitas a las droguerías, en el caso de la tarde teniendo a gastronomía, salud y moda, mientras que las mañanas arquitectura y construcción.

- Miércoles

Ilustración 39: visualización visitas miércoles

En este día se repiten las jornadas más importantes como lo son la tarde y la mañana, dentro de las cuales se encuentran en la tarde, la gastronomía, moda, agencias de viaje y droguerías, mientras que en la mañana las joyas, belleza y droguerías como las categorías más visitadas

- Jueves

Ilustración 40: visualización visitas jueves

Se determina que los jueves la jornadas en las que más visitas se obtienen en la plataforma es en la noche, tarde y mañana, además se ve una gran representación de la categoría deportes en la jornada de la noche, mientras que en la tarde se ven las categorías droguerías, hogar, electrodomésticos, tecnología y moda con valores muy similares; en la mañana las categorías más representativas son arquitectura, construcción y moda.

- Viernes

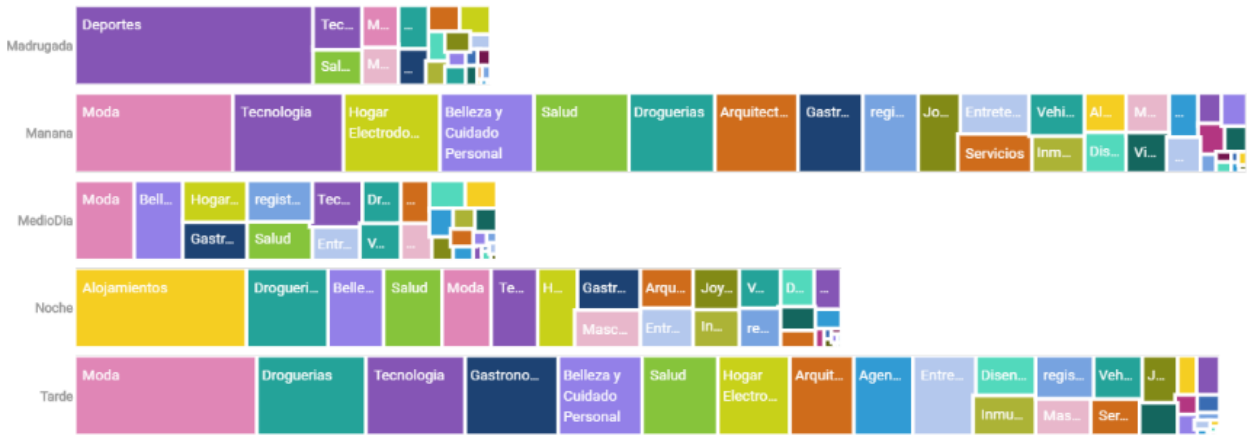


Ilustración 41: visualización visitas viernes

En este caso se puede describir que la mañana y la tarde son las jornadas más representativas del día viernes, teniendo a moda, tecnología y droguerías entre las categorías que en ambas jornadas son las más visitadas.

- Sábado

Ilustración 42: visualización visitas sábado

Por último se tienen a la mañana, la tarde y muy cerca de la noche como las jornadas más representativas, dentro de las que se repiten moda, gastronomía y droguerías como las categorías más visitadas en esos días.

3.5.3.2 Clustering (“Segmentación”)

En este proceso se busca identificar las tipologías de los grupos donde los elementos guarden gran similitudes entre sí y varias diferencias con el resto de grupos. Así se lograría realizar una segmentación de los visitantes de la plataforma.

Aplicación EM: Para el proceso de clustering se llevó a cabo la aplicación del algoritmo EM (Expectation Maximization), usando los parámetros que se visualizan a continuación:

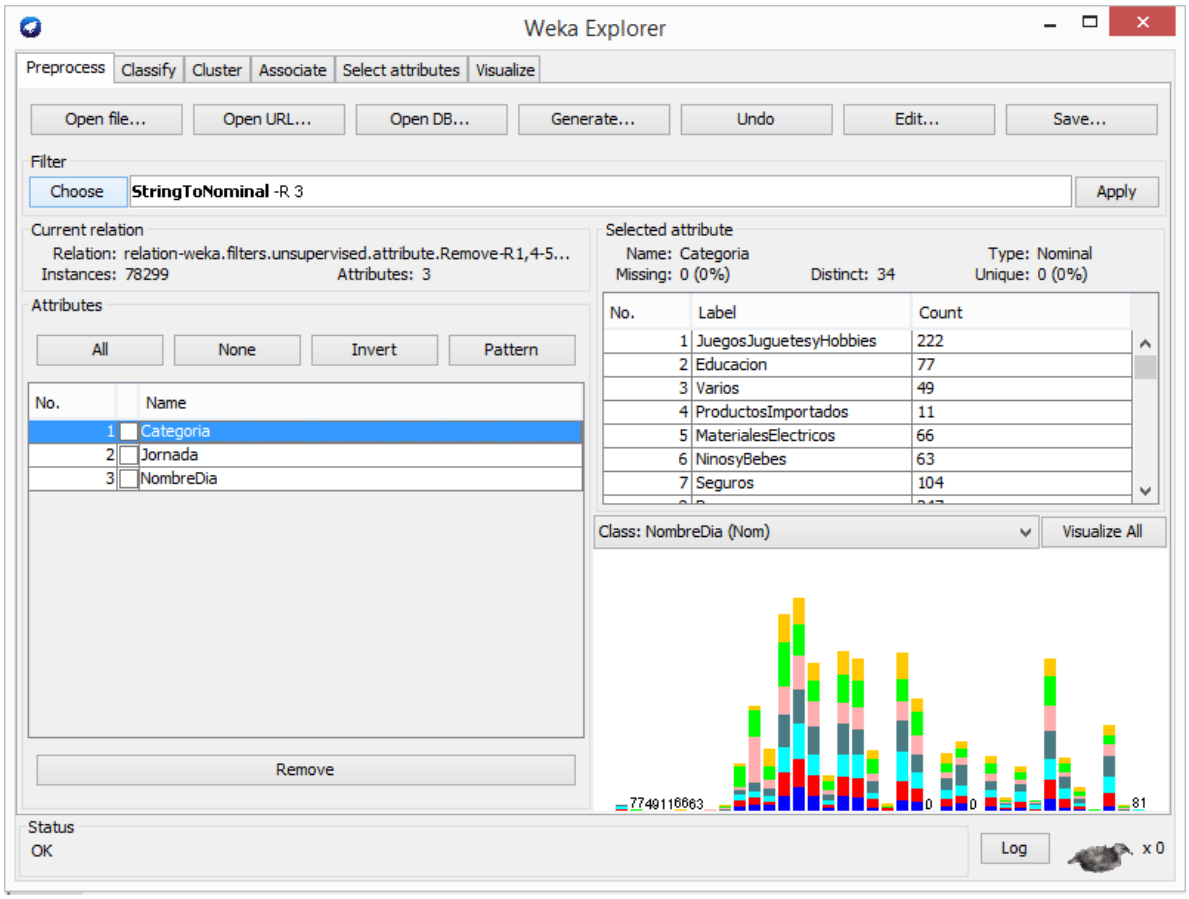


Ilustración 42: aplicación del algoritmo EM

Encontrando luego de la ejecución del algoritmo 16 *clusters* como se muestra a continuación

```
=== Run information ===

Scheme:weka.clusterers.EM -I 100 -N -1 -M 1.0E-6 -S 100
Relation:     relation-weka.filters.unsupervised.attribute.
                Remove-R1-weka.filters.unsupervised.attribute.
                Remove-R3-weka.filters.unsupervised.attribute.
                StringToNominal-R4-weka.filters.unsupervised.
                attribute.NumericToNominal-R3-weka.filters.
                unsupervised.attribute.Remove-R3

Instances:    78299
Attributes:   3
              Categoria
              Jornada
              NombreDia
Test mode:evaluate on training data

=== Model and evaluation on training set ===


EM
==

Number of clusters selected by cross validation: 17

Time taken to build model (full training data) : 8232.54 seconds

=== Model and evaluation on training set ===

Clustered Instances

 0       4437 (  6%)
 1       7342 (  9%)
 2       4846 (  6%)
 3       2997 (  4%)
 4       7072 (  9%)
 5      13436 ( 17%)
 6       2249 (  3%)
 7       4414 (  6%)
 8       3896 (  5%)
 9       1743 (  2%)
10       6648 (  8%)
11       1999 (  3%)
12       4172 (  5%)
13       1518 (  2%)
14       3298 (  4%)
15       5885 (  8%)
16       2347 (  3%)
```

Tabla 26: clusters encontrados

Aplicación SimpleKMeans

Luego de esto se ejecutó el algoritmo SimpleKMeans buscando describir los conglomerados con comportamientos similares, las variables consideradas para este proceso son Categoria, Jornada y NombreDia. Estableciendo siete (7) como la cantidad de clusters a encontrar, teniendo en cuenta que muchos de los clusters encontrados con el algoritmo EM, tenían porcentajes muy bajos y se buscaba llevar a dar resultados más concretos, obteniendo los visualizados a continuación.

```
Clustered Instances

0       9616 ( 36%)
1       7821 ( 29%)
2       3566 ( 13%)
3       1544 (  6%)
4       1892 (  7%)
5       1429 (  5%)
6        754 (  3%)
```

Tabla 27: conglomerados con comportamientos similares

Que se describen puntualmente a continuación:

- Cluster 0:

```
Attribute          0
                (18953)
==========:=========:
Categoria       Moda
Jornada        Noche
NombreDia    viernes
```

- Cluster 1:

```
Attribute           1
                (15124)
==========:==========:
Categoria  Droguerias
Jornada         Tarde
NombreDia   miercoles
```

- Cluster 2:

Attribute	2
	(6924)
Categoria	HogarElectrodomesticosyOficina
Jornada	Manana
NombreDia	lunes

- Cluster 3:

Attribute	3
	(2861)
Categoria	JoyasyAccesorios
Jornada	Manana
NombreDia	miercoles

- Cluster 4:

Attribute	4
	(3689)
Categoria	Gastronomia
Jornada	Manana
NombreDia	sabado

- Cluster 5:

Attribute	5
	(2613)
Categoria	Salud
Jornada	MedioDia
NombreDia	martes

- Cluster 6

Attribute	6
	(1513)
Categoria	Deportes
Jornada	Madrugada
NombreDia	viernes

Estos resultados obtenidos nos dan un acercamiento más profundo a las características de los visitantes de la plataforma. Teniendo la base sobre la cual se puede buscar generar conclusiones de aquí en adelante.

Para hacer una descripción más profunda de los clusters se realizó un análisis que cruzara los distintos atributos, logrando los siguientes resultados.

Visitas realizadas a las categorías por jornada: por medio de este análisis, se buscó relacionar las jornadas con las categorías, encontrando.

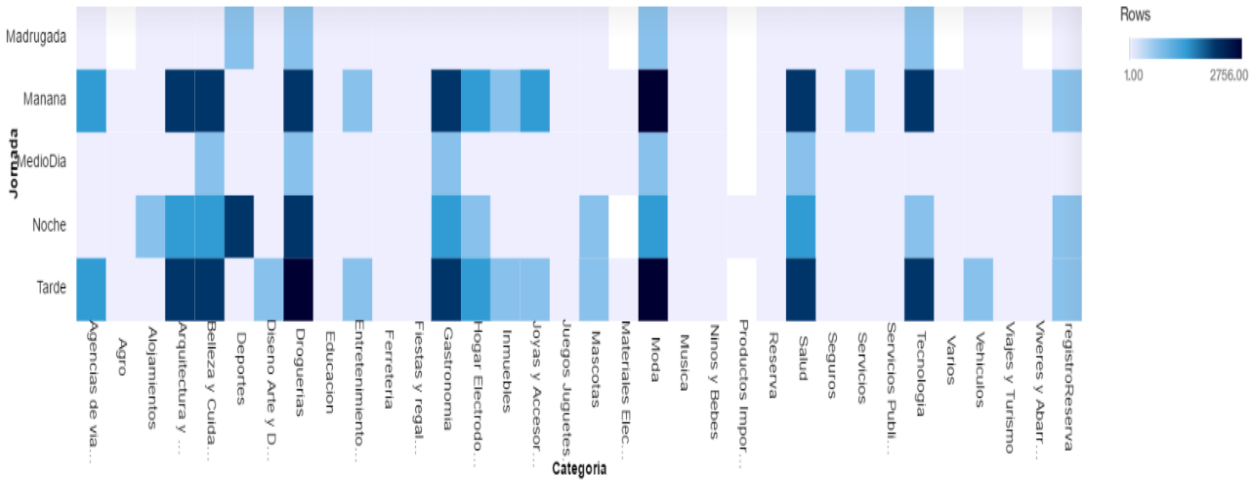


Tabla 28: categorías por jornada

Pudiendo determinar la importancia de cada una de las categorías con respecto a la jornada, pudiendo corroborar que droguerías es una categoría importante para la plataforma, siendo foco en cada una de las jornadas, pero con mayor impacto en la tarde, seguida por la noche y mañana, sin embargo estando presente en la madrugada y al medio día, pero con menos relevancia, vemos también importancia en las categorías belleza y cuidado personal, moda, salud, tecnología, arquitectura y construcción; así mismo pudiendo ver aquellas categorías que no son representativas dentro de la plataforma, como lo son productos importados, materiales eléctricos, agro, víveres y abarrotes, ferretería, fiestas y regalos.

Visitas realizadas a las categorías por Nombre Día: a través de este proceso se busca determinar las relaciones existentes entre los atributos NombreDia y Categorias, encontrando lo siguiente.

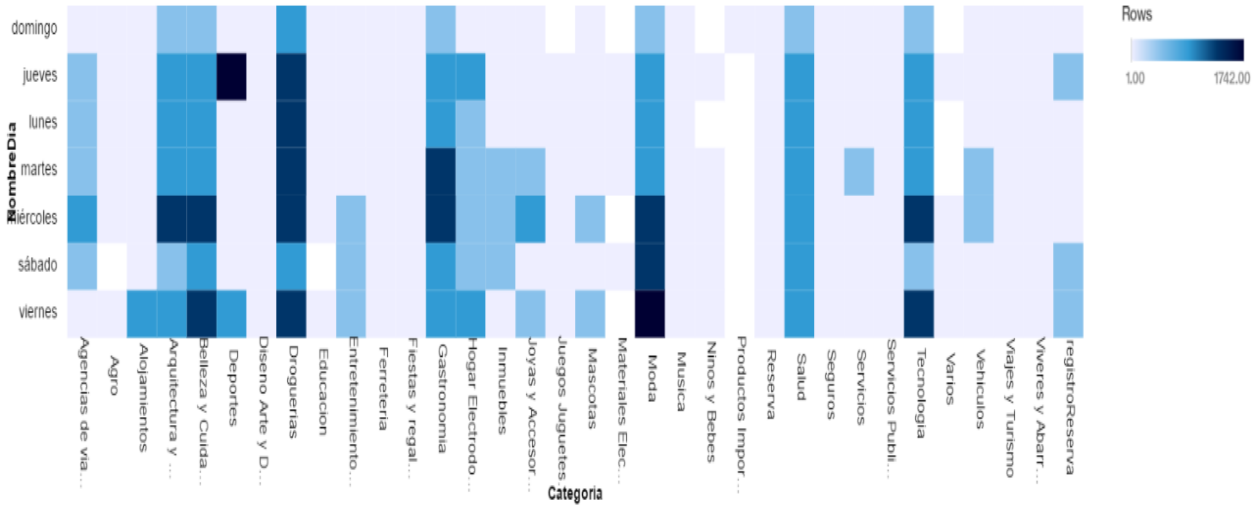


Tabla 29: visitas a las categorías por Nombre Día

Donde podemos visualizar que de nuevo droguerías es importante dentro de la plataforma, sin embargo se pueden identificar algunas relaciones que deben tenerse en cuenta como deportes los jueves, moda los viernes, tecnología los viernes y miércoles, entre lo demás que se pueden encontrar analizado de la imagen.

Visitas por Jornada y NombreDia: a continuación se busca determinar la combinación de días y jornadas importantes dentro de la plataforma, encontrando lo siguiente

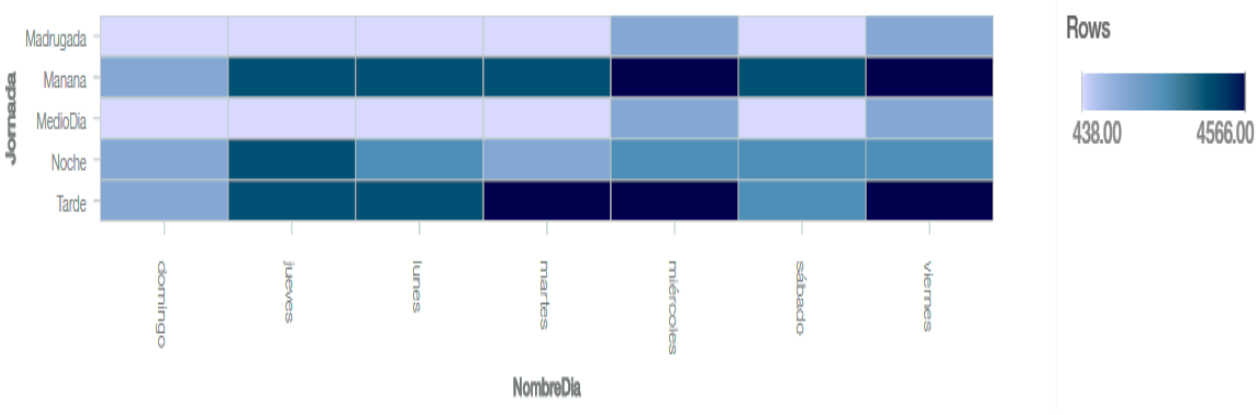


Tabla 30: visitas por Jornada y NombreDia

Las combinaciones más importantes para la plataforma son los martes en la tarde, miércoles en la mañana y la tarde; por último, los viernes en la tarde y la mañana.

Visitas por Jornada y NúmeroDía: por último se busca identificar el comportamiento de los visitantes durante la duración de un mes, encontrando lo que muestra la siguiente imagen.

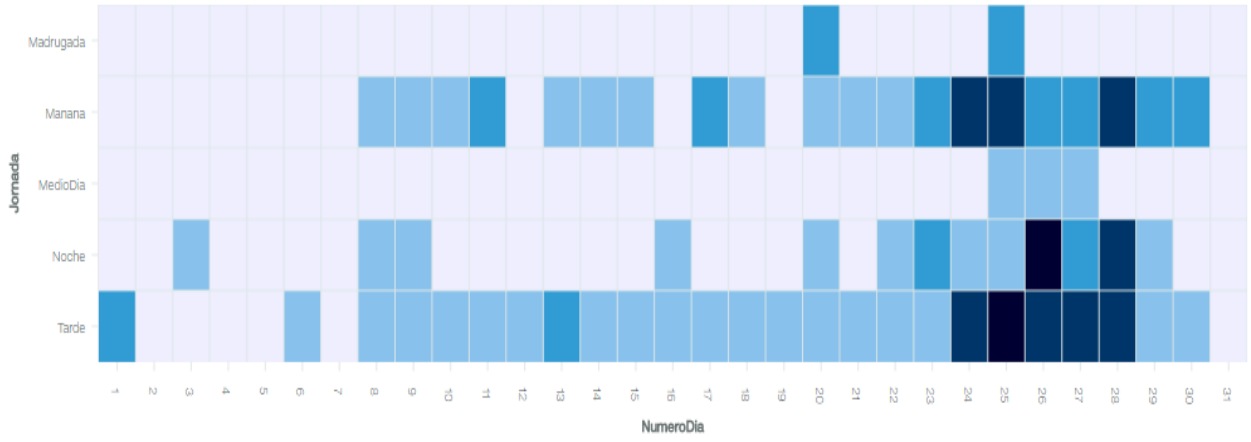


Tabla 31: visitas por Jornada y NúmeroDía

Se puede identificar que la plataforma tiene un aumento en su actividad con el pasar de los días, empezando con poca actividad en os primeros siete (7) días, luego teniendo un poco más de

visitas, sin embargo su actividad más alta se da en los últimos días del mes y en las jornadas de la tarde y la mañana, mientras que la noche se activa entre los días 20 y 29.

3.5.3.3 Reglas de asociación

El objetivo de la aplicación de esta técnica de minería es encontrar regularidades en los comportamientos de los visitantes dentro de términos de combinaciones de categorías que son visitadas muchas veces en un conjunto, quiere esto decir reglas que reflejen relaciones entre los atributos presentes en los datos.

Para llevar a cabo esta parte del proceso, se realizó un dataset que permitiera identificar las categorías accedidas por un visitante, por lo que se generó un nuevo formato que cumple con las siguientes características:

En las filas se tendrá cada una de las sesiones de un visitante y en las columnas las categorías (atributos) visitadas por él, tal como se ve en la siguiente tabla.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Id	Reserva	Alojamiento	Deportes	registroRese	Moda	Droguerias	Arquitectura	DisenoArtey	BellezayCuic	Salud	Entretenimi	Gastronomia	HogarElectrc	Inmuebles	JoyasyAcces	Mascotas	Servicios
1	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
2	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0
3	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
4	1	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	1	0	1	1	1	0	1
6	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
7	0	0	0	0	1	0	0	0	1	1	1	1	1	0	1	0	0
8	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
9	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	1	0	0	0	0	0	1	0	1	1	1	0	0
11	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
15	0	0	0	0	1	0	0	0	1	1	0	1	1	1	1	0	0
16	0	0	0	1	0	0	1	0	1	0	1	1	1	1	0	0	1
17	0	0	0	0	1	0	0	0	1	1	0	0	0	1	1	0	0
18	0	0	0	0	1	0	1	0	1	1	0	0	1	0	1	0	1
19	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0

Tabla 32: dataset

Donde los ceros significan que en esa sesión no visitaron dicha categoría y los unos que sí; este archivo permitirá crear reglas de asociación que permitan analizar más a fondo el comportamiento de los visitantes de la plataforma.

Aplicación algoritmo FP-Growth

Se llevó a cabo la ejecución del algoritmo FP-Growth en el programa RapidMiner, diseñando el proceso de la siguiente manera.

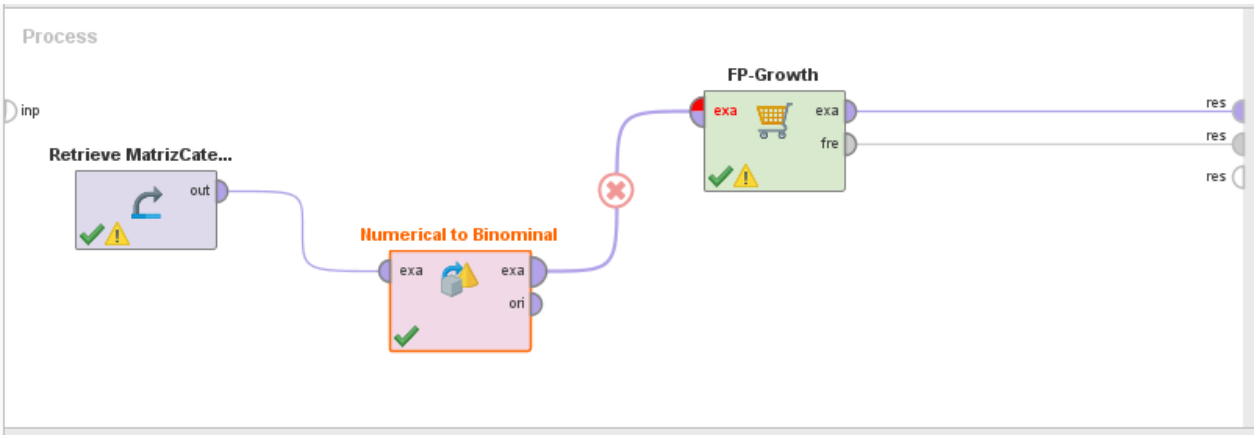


Ilustración 43: ejecución del algoritmo FP-Growth

Obteniendo sets que permitieran llevar a cabo el algoritmo apriori, para luego buscar relaciones o afinidades entre dichos sets y que se describen a continuación:

Size	Support	Item 1	Item 2	Item 3
1	0.421	Moda		
1	0.388	BellezayCuidadoPersonal		
1	0.331	Tecnologia		
1	0.329	Salud		
1	0.297	HogarElectrodomesticosy...		
1	0.265	Gastronomia		
1	0.231	EntretenimientoyVidaNoct...		
1	0.198	Inmuebles		
1	0.172	registroReserva		
1	0.162	Vehiculos		
1	0.148	ViajesyTurismo		
1	0.145	Agenciasdeviaje		
1	0.139	Servicios		
1	0.117	JoyasyAccesorios		

1	0.117	JoyasyAccesorios		
1	0.116	ArquitecturayConstruccion		
1	0.107	Mascotas		
1	0.102	Droguerias		
2	0.187	Moda	BellezayCuidadoPersonal	
2	0.182	Moda	Tecnologia	
2	0.167	Moda	Salud	
2	0.165	Moda	HogarElectrodomesticosy...	
2	0.128	Moda	Gastronomia	
2	0.134	Moda	EntretenimientoyVidaNoct...	
2	0.117	Moda	Inmuebles	
2	0.097	Moda	Vehiculos	
2	0.134	BellezayCuidadoPersonal	Tecnologia	
2	0.212	BellezayCuidadoPersonal	Salud	
2	0.122	BellezayCuidadoPersonal	HogarElectrodomesticosy...	
2	0.102	HogarElectrodomesticosy...	Inmuebles	
2	0.103	Gastronomia	EntretenimientoyVidaNoct...	
2	0.122	BellezayCuidadoPersonal	HogarElectrodomesticosy...	
2	0.132	BellezayCuidadoPersonal	Gastronomia	
2	0.119	BellezayCuidadoPersonal	EntretenimientoyVidaNoct...	
2	0.097	BellezayCuidadoPersonal	Inmuebles	
2	0.126	Tecnologia	Salud	
2	0.154	Tecnologia	HogarElectrodomesticosy...	
2	0.102	Tecnologia	Gastronomia	
2	0.099	Tecnologia	EntretenimientoyVidaNoct...	
2	0.092	Tecnologia	Inmuebles	
2	0.119	Salud	HogarElectrodomesticosy...	
2	0.107	Salud	Gastronomia	
2	0.107	Salud	EntretenimientoyVidaNoct...	
2	0.103	HogarElectrodomesticosy...	Gastronomia	
2	0.104	HogarElectrodomesticosy...	EntretenimientoyVidaNoct...	
2	0.102	HogarElectrodomesticosy...	Inmuebles	
2	0.103	Gastronomia	EntretenimientoyVidaNoct...	
3	0.094	Moda	BellezayCuidadoPersonal	Tecnologia
3	0.109	Moda	BellezayCuidadoPersonal	Salud
3	0.093	Moda	Tecnologia	Salud
3	0.101	Moda	Tecnologia	HogarElectrodomesticosy...

Tabla 33: relaciones o afinidades entresets

Estos sets serán usados para buscar reglas de asociación aplicándolas al operador *Create Association.*

Aplicación operador Create Association

A través de la aplicación de este operador de RapidMiner se busca determinar reglas de asociación, para lo que se diseñó el siguiente proceso.

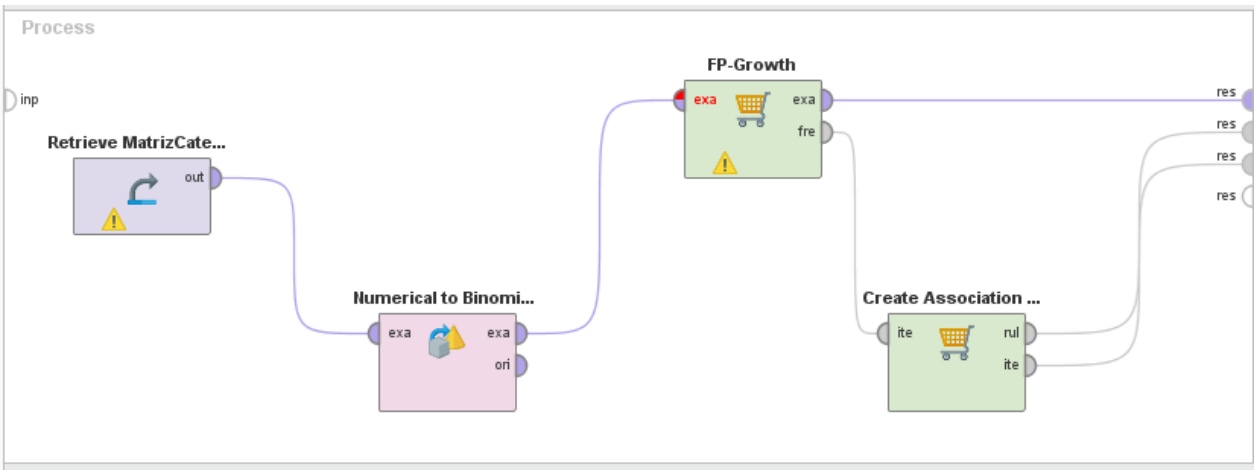


Ilustración 46: aplicación operador Create Association

Encontrando lo siguientes resultados:

No	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Conviction
3	Salud	BellezayCuidadoPersonal	0,212	0,646	0,912	-0,445	0,085	1,664	1,727
4	Tecnologia, HogarElectrodomesticosyOficina	Moda	0,101	0,654	0,954	-0,207	0,036	1,552	1,673
5	Moda, Salud	BellezayCuidadoPersonal	0,109	0,655	0,951	-0,224	0,044	1,688	1,774
6	BellezayCuidadoPersonal, Tecnologia	Moda	0,094	0,702	0,965	-0,174	0,038	1,666	1,941
7	Tecnologia, Salud	Moda	0,093	0,742	0,971	-0,158	0,040	1,760	2,241

Tabla 34: resultados aplicación operador Create Association

Una regla de asociación se forma por dos conjuntos el antecedente y el consecuente, lo que en este caso permite establecer algunas hipótesis como las siguientes:

- Visitantes que visitan la categoría salud después visitaran belleza y cuidado personal

- Aquellos que visiten tecnología, hogar, electrodomésticos, oficina, belleza y cuidado personal y salud visitaran moda.
- Los visitantes que hayan visitado moda y salud visitaran belleza y cuidado personal.

Luego de identificar las reglas de asociación mencionadas anteriormente, se busca identificar estas reglas en Watson Analytics y corroborar lo encontrado a través de RapidMiner, encontrado los siguientes resultados.

Moda:

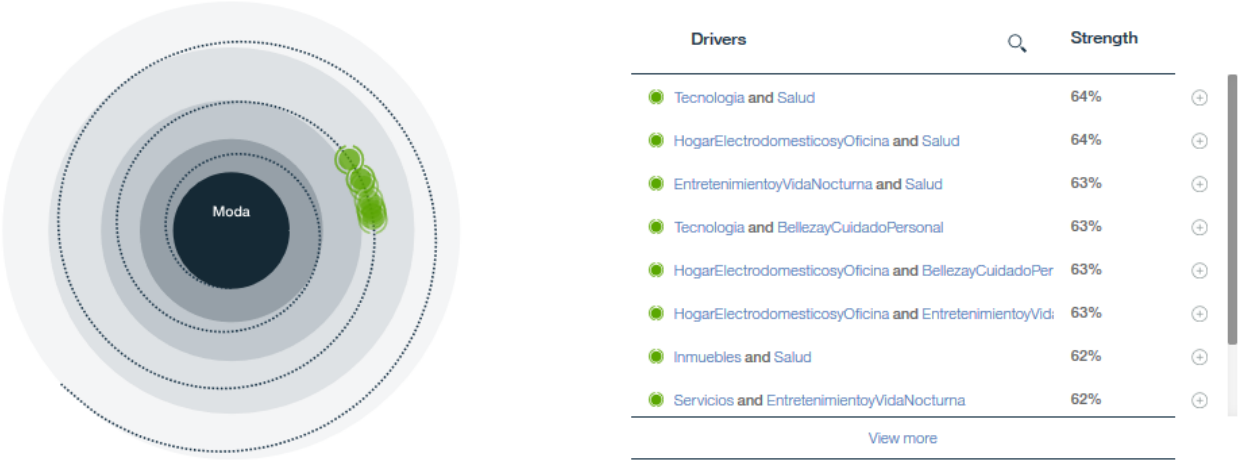
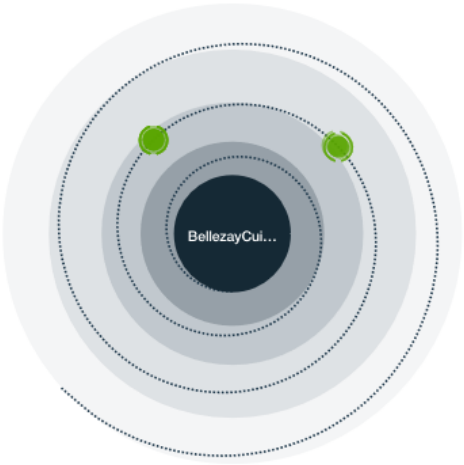


Ilustración 45: resultados de reglas de asociación: Moda

Corroborando lo que encontrábamos en las reglas de asociación, pudiendo ver que tienen una fuerte relación con tecnología, electrodomésticos, belleza y cuidado personal.

Belleza y cuidado personal:



Drivers	Strength
Salud	71%
Salud and ArquitecturayConstruccion	71%
Salud and Reserva	71%
Salud and DisenoArteyDecoracion	71%
Salud and registroReserva	71%
HogarElectrodomesticosyOficina and Salud	71%
Inmuebles and Gastronomia	65%
ViajesyTurismo and Gastronomia	65%

View more

Ilustración 46: resultados de reglas de asociación: Belleza y cuidado personal

Se vuelve a encontrar la relación que tienen belleza y cuidado personal con salud, lo que respalda lo encontrado como una regla de asociación.

3.6 Fase V. Evaluación.

3.6.1 Evaluación de los resultados

Al finalizar el proceso de modelado se llevó a cabo una reunión con el equipo de trabajo de la plataforma Oferto (Director, Profesional en mercadeo, Web master), logrando validar que los resultados obtenidos concuerdan con las hipótesis planteadas por ellos en sus informes mensuales, sin embargo logrando ahondar en algunos aspectos que para ellos no era posible detectar sin realizar el proceso de minería web, como lo son las relaciones entre dos o más atributos, además de esto se logra validar con ellos que las categorías más visitadas son aquellas que constantemente tienen una mayor cantidad de productos publicados por comerciantes, lo que

permitiría pensar que podría existir una relación entre la cantidad de productos de una categoría y las visitas que se hacen a la misma. Además de esto, se corrobora la poca reserva de productos a través de la plataforma, siendo este uno de los aspectos a fortalecer a través de la aplicación de los resultados obtenidos en el funcionamiento dinámico de la plataforma y la creación de estrategias de mercado que tengan como objetivo final aumentar la cantidad de reservas dentro de la plataforma.

Después de esto, se puede concluir que los resultados obtenidos en la primera experiencia del proceso de minería con la plataforma Oferto son satisfactorios, permitirán al director y web master de la plataforma proponer acciones con el fin de aplicar lo identificado en pro de la mejora de la plataforma en su parte funcional y de diseño; ya que dentro de cada uno de los resultados de las técnicas aplicadas se encuentra información que permitirá mejorar la forma en la que se distribuyen los productos, los horarios en los que se debe publicar productos de determinada categoría, horarios y días en los que se deben establecer estrategias para aumentar las visitas dentro del sitio, categorías que por distintos factores podrían parecer innecesarias dentro de la misma, entre otras más.

Finalmente, cada uno de los resultados obtenidos con los algoritmos ejecutados en Weka y RapidMiner cumplió totalmente con los requerimientos exigidos en la metodología CRISP DM y se validaron por medio de *Watson Analytics*. Lo que permitió generar conocimiento nuevo de la plataforma Oferto de la CCAQ, cuyos resultados permitirán apoyar a los directos responsables de la misma, para sus fines pertinentes, buscando finalmente generar conocimiento que apoye el objetivo de fortalecer las empresas del departamento del Quindío.

3.6.2 Proceso de revisión

Se debe tener en cuenta que durante el proceso se generaron dificultades al momento de buscar reglas de asociación, debido a que al contar con muchas categorías dentro de la plataforma y de ellas un gran porcentaje con bajo número de visitas, las reglas más representativas eran basadas en aquellas que no eran visitadas, por lo que se tuvo que buscar distintas maneras de obtener resultados adecuados para esa etapa del proceso, lo que llevo a la utilización del programa RapidMiner, facilitando esté la identificación de reglas basados en sets que se determinaban por la fuerte relación que tenían en las sesiones establecidas, por lo que para un próximo proceso, se debe buscar que el web master generalice algunas categorías con las que se cuenta en la plataforma que tienen una gran similitud con otras, como por ejemplo lo son salud, droguerías, belleza y cuidado personal, logrando esto facilitar el manejo de las mismas en cada una de las etapas del proceso de minería de datos.

Sin embargo, se puede decir que fue acertado manejar en la fase de preparación de los datos las categorías como el núcleo de la limpieza, ya que al contar con una gran cantidad de productos dentro de la plataforma, manejarlos de manera individual habría dificultado aún más la obtención de resultados claros y fáciles de interpretar; además de esto haber llevado a cabo la eliminación de todas aquellas peticiones que no fueron ejecutadas por un visitante permitió tener un resultado más preciso.

Para un próximo proceso de minería en la plataforma y con el fin de agilizar el proceso, se identificó después de varios intentos, que en el análisis para el proceso de clasificación se deben plantear hipótesis que ayuden a pensar que relaciones se buscan entre los diferentes atributos del data warehouse, para de esta manera usar aquellos necesarios y así disminuir el tiempo de ejecución al momento de aplicar cada una de las técnicas de minería, sin embargo pueden surgir

relaciones y resultados que se deberán tener en cuenta como valiosos para la conclusión del proceso de minería.

4. Conclusiones

La construcción del Data Warehouse que se realizó durante el proceso de Preparar los datos fue determinante para realizar con éxito la siguiente fase del proyecto. Cabe anotar que para el éxito de esta fase, fue indispensable la implementación de algoritmos que realizarán la limpieza de registros de forma automática basados en reglas de limpieza definidas en la implementación del código.

La metodología CRISP-DM demostró ser una metodología poderosa ya que esta abarca todo el ciclo de vida del proyecto de minería de datos, adicionalmente demostró una gran adaptabilidad a las necesidades de la CCAQ teniendo en cuenta que un proyecto de este tipo no tiene precedentes dentro de la entidad. Si bien CRISP-DM no es una fórmula mágica para el éxito de un proyecto de minería de datos, si se involucra algo de formación y algunos expertos, da un excelente punto de inicio como herramienta para dar respuesta a las preguntas que tengan los directivos sobre el negocio.

A lo largo del desarrollo de este trabajo, el objetivo ha sido identificar la existencia de patrones de comportamiento de los visitantes de la plataforma “Oferto” de la CCAQ combinando técnicas de minería web al log de acceso generado por esta. Encontrando que se puede dar por hecho que realizar un seguimiento adecuado al visitante, permite generar información interesante sobre su comportamiento dentro de la plataforma, como por ejemplo las categorías que más importantes para ellos, jornadas habituales de consulta, días de la semana con mayor importancia, el comportamiento durante el mes, relaciones entre jornadas, categorías, días de la semana día del mes. Lo que nos lleva a pensar que después de identificar dichas preferencias, se debería ofrecer

o recomendar al visitante información de manera personalizada; basados en los resultados del proceso de minería, potenciando así la forma en la que se muestra la información.

Para obtener los patrones de comportamiento de los visitantes se ha detallado un modelo de obtención de los datos basándonos en las técnicas de minería mostradas en este trabajo, teniendo como base las categorías que visitan, las jornadas y los días en los que acceden a las mismas.

Hemos planteado un modelo de obtención de clusters mediante el análisis de los atributos con los que se contaban, logrando determinar después del proceso 7 clusters, que nos permiten agrupar visitantes con características similares, logrando de esta manera identificar qué tipos de visitantes se conectan a la web en relación a las categorías más representativas.

Por otra parte, se estableció un modelo que permitiera llevar a cabo un análisis de las sesiones de los visitantes de la plataforma, logrando a partir de esté extraer reglas de asociación aplicando los operadores FP-Growth y Create Association, para así identificar reglas que faciliten el manejo dinámico de la información dentro de la categoría.

Por lo que respaldados en los resultados encontrados en este proceso se sugiere al director y web master de la empresa, establecer propuestas de mejoras en el funcionamiento lógico y diseño de la plataforma, esto con el fin de iniciar un proceso de personalización, basados en los comportamientos del visitante, buscando con esto fortalecer las relaciones con ellos y satisfacer finalmente sus necesidades de consumo.

Como punto de inicio se sugiere manejar la publicidad de los productos en los banners de la plataforma según el día y la jornada, teniendo como base los resultados obtenidos en la aplicación de cada una de las técnicas de minería, como lo son realizar campañas dirigidas enfatizando de acuerdo a los siguientes ítems:

- En la madrugada deportes, tecnología, moda y droguerías.

- En la mañana moda, después entre salud, gastronomía, droguerías, tecnología, belleza y cuidado personal, arquitectura y construcción.
- Al medio día en mayor proporción productos de gastronomía, salud, moda, droguerías, belleza y cuidado personal.
- En la tarde en mayor proporción productos de las categorías moda, droguerías en el index de la plataforma.
- En la noche en mayor proporción productos de las categorías deportes y droguerías.

Además de implementar un módulo inicial de recomendaciones automáticas, esto para las categorías identificadas en las reglas de asociación como lo son que aquellos que hayan visitado salud, probablemente querrán conocer acerca de los productos de belleza y cuidado personal, los que hayan visitado tecnología, hogar, electrodomésticos y oficina les podrían interesar los productos de la categoría moda, igual que aquellos que visitaron tecnología y salud, buscando así medir el impacto de estas recomendaciones en los visitantes. Todo lo anterior permite brindar al web master información relevante para el mejoramiento de la plataforma basado en la realidad de la misma y buscando así generar un impacto positivo en los visitantes.

Por último vale la pena resaltar una vez más la importancia del proceso de limpieza de los datos ya que gran cantidad de los datos iniciales tienen una gran cantidad de ruido que no permitirá obtener los resultados que se obtuvieron.

5. Recomendaciones

Luego de terminado el proceso se sugiere integrar categorías y normalizar las URL desde la parte lógica de la plataforma, con el fin de facilitar un futuro proceso de minería, ya que al tener un alto número de categorías con pocas visitas y productos asociados dificulta el proceso durante la limpieza y el modelamiento; proponiendo en su reemplazo generar mejores estrategias de búsqueda dentro de la plataforma.

También se debería pensar por parte del web master integrar en la plataforma el algoritmo desarrollado en la Fase III. Preparación de los datos, con el fin de agilizar el proceso y construir de manera automática un data warehouse con la estructura propuesta en la Fase IV. Modelado, lo que permitiría disminuir el tiempo invertido en cada una de estas etapas y pudiendo así invertir el tiempo en la aplicación de diferentes técnicas que permitan hallar diferentes o mejores resultados.

Por otra parte es importante generar copias de seguridad de los log de acceso, para poder contar en un futuro con una muestra de por lo menos un año, lo que permitiría encontrar comportamientos en relación a los meses y posiblemente desarrollar un módulo dentro de la plataforma, que realice todo lo desarrollado en este trabajo de manera automatizada y enfocar dicho módulo a la inteligencia de negocios.

Luego de realizado el proceso de minería web y basándose en la limitación de información que se encontraba almacenada en el servidor de la plataforma, se recomienda realizar una copia de seguridad de los Log de acceso mensualmente para poder replicar el proceso hecho en este proyecto a el total de la información.

Referencias bibliográficas

- Adam, D. (2015). *Information Technology*. En http://datamining.itsc.uah.edu/adam/documentation.html. Extraído el 28 de Octubre del 2015.

- Baeza-Yates, R. Ribeiro-Neto, B. (1999). *Modern Information retrieval,* Addison – Wesley.

- Baeza-Yates, R., Poblete, B. (2007), *Un modelo de minería de consultas para el diseño del contenido y la estructura de un sitio Web*, Universitat Pompeu Fabra & Centro de Investigación de la Web.

- Calderon,A. (2007). *Proyecto TariyKDD, lo que tus datos dicen*. *Essentia Libre* en

- http://www.mclibre.org/descargar/docs/revista-essentia-libre/essentia-libre-06-200703.pdf Extraído el 28 de Octubre del 2015.

- Chen, M. Han, J. Yu, P. (1996) *Data mining: An overview from database perspective*. En IEEE Transactions on Knowledgeand Data Engineering. En http://www.nyu.edu/classes/jcf/g22.3033-002/handouts/chen96data.pdf Extraído el 22 de Octubre del 2015

- Demsar, J., Zupan, B. Leban, G. Curk, T. (2014) *Orange: From experimental machine learning to interactive data mining*. Technical report, Faculty of Computer and Information Science, University of Liubliana, Slovenia, Dep. of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA.

- Evangelos, S., J. Han, (1996). *Proceeding of the Second International Conference on Knowledge Discovery and Data Mining*. Portland, EEUU.

- Facca, F., Lanzi, P., (2005) *Mining interesting knowledge from weblogs: a survey*. Dipartimento di Elettronica e Informazione, Artificial Intelligence and Robotics Laboratory, Politecnico di Milano, Italy.

- Fayyad, U.M, Piatetsky-Shapiro, G. Smyth, P., (1996) *From Data Mining to Knowledge Discovery: An Overview*, En Advances in Knowledge Discovery and Data Mining, AAAI Pres/ The MIT Pres. En https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131 Extraído el 22 de Octubre del 2015

- Fayyad U. (1998) Mining Databases: Towards Algorithms for Knowledge Discovery, *Bulletin of the IEEE Computer Society*, Vol.21, No. 1, March.

- Fayyad, U. Piatetskiy-Shapiro, G.; Smith. P.; Ramasasmy, U. (1996) *Advances in Knowledge Discovery and Data Mining*. AAAI Press / MIT Press.

- Fayyad, U., Piatetsky-Shapiro, Gregory; Smyth, Padhraic. 1996. From Data Mining to Knowledge Discovery in Databases.

- González, J., (2011) S*istema de apoyo para la acreditación de la calidad de programas académicos de la Universidad de Caldas, aplicando técnicas en minería de datos*. Tesis para optar al título de magister en Gestión y Desarrollo de Proyectos de Software, Universidad Autónoma de Manizales, Manizales, Colombia.

- Han J. Fu Y. Wang W. et al (1996), *DBMiner: A System for Mining Knowledge in Large Relational Databases*. The second International Conference on Knowledge Discovery & Data Mining, Portland, Oregon.

- Han, J., Fu, Y., Wang, W. et al (1996) *DBMiner: Interactive Mining of Multiple-Level Knowledge in Relational Databases*. ACM SIGMOD, Montreal, Canada.

- Han J., Chiang J., Chee S., Chen J., Chen q. et al (1997) DBMiner: A System for Data Mining in Relational Databases and Data Warehouses. *CASCON: Meeting of Minds*, Toronto, Canadá.

- Han, J., Kamber, M. (2006) *Data Mining: Concepts and Techniques* (2nd edition). Morgan Kaufmann Publishers, San Francisco, EEUU.

- Hernández, O. J. Ramírez, J, Ferri Ramírez, C. (2005). I*ntroducción a la minería de datos.* Editorial Pearson. Madrid España.

- Hernández, O. J. (2015) *Curso de Análisis y Extracción de Conocimiento en Sistemas de Información: Datawarehouse y Datamining En* http://users.dsic.upv.es/~jorallo/cursoDWDM/ extraído el 14 Agosto del 2015.

- IBM Corporation (2011). Manual del usuario del sistema básico de IBM SPSS. En ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/es/client/Manuals/IBM_SPSS_Statistics_Core_System_Users_Guide.pdf extraído el 22 de Octubre del 2015

- Kosala, R., Blockeel, H (2000). Web Mining Research: A Survey. *ACM SIGKDD Explorations, Newsletter of the Special Interest Group on Knowledge Discovery and Data Mining*, 1-9.

- Lopes, P., Roy, B., (2015), *Dynamic Recommendation System Using Web Usage Minning for E-Commerce*. Deparment of Computer Enginnering St. Francis Institute of Technology.

- Manning, C. D.; Shütze, H., (1999). Foundations of Statistical Natural Language Processing, MIT Pres.

- Moine, J. (2013). *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo*. Recuperado de http://sedici.unlp.edu.ar/bitstream/handle/10915/29582/Documento_completo.pdf?sequence=1 Extraído del 22 de Agosto del 2015.

- Molina, L.C. (1998). *Data mining no processo de extração de conhecimento de bases de dados.* Tesis de máster. São Carlos (Brasil): Instituto de Ciências Matemáticas e Computação. Universidad de São Paulo.

- Molina, L.C., (2002). *Data mining: torturando a los datos hasta que confiesen*, Recuperado de http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html extraído el 28 de Octubre del 2015

- Michalski R., (1983). *A Theory and Methodology of Inductive Learning*. Morgan-Kauffman, EEUU.

- Michalski R., I. Bratko, M. Kubat, (1998). *Machine Learning and data mining: Methods and Applications*. Wiley & Sons Ltd., EE.UU.

- Perichinsky, G. y R. Garcia Martinez, (2000). A Data Mining Approach to Computational Taxonomy. *Proceedings del Workshop de Investigadores en Ciencias de la Computación*. Págs.107-110. Departamento de Publicaciones de la Facultad de Informática. Universidad Nacional de La Plata, Buenos Aires, Argentina.

- Perichinsky,G., R. García-Martínez, A. Proto, (2000). *Knowledge Discovery Based on Computational Taxonomy And Intelligent Data Mining*. CD del VI Congreso Argentino de Ciencias de la Computación. Ushuaia, Argentina.

- Perichinsky, G., R. Garcia-Martínez, A. Proto, A. Sevetto, et al (2001). *Integrated Environment of Systems Automated Engineering. Proceedings del II Workshop* de Investigadores en Ciencias de la Computación. Editado por Universidad Nacional de San Luis en el CD Wicc.

- Perichinsky, G. Servente, M., Servetto, A. et al (2003). Taxonomic Evidence and Robustness of the Classification Appying Intelligent Data Mining. Proceedings del VIII Congreso Argentino de Ciencias de la Computación. Pág. 1797-1808.

- Piatetski-Shapiro, G., U. Fayyad, P. Smith, (1996). *From data mining to Knowledge discovery*. AAAI Press/MIT Press, California, EEUU.

- Piatetski-Shapiro, G.; W. Frawley, C. Matheus, (1991). *Knowledge discovery in databases: an overview*. AAAI-MIP Press, Menlo Park, California, EEUU.

- RapidMiner Studio (Octubre, 2015). RapidMiner En https://rapidminer.com/products/studio/ extraído el 16 de Agosto del 2015.
- Reyes, J., García, R. (2005) El proceso de descubrimiento de conocimiento en bases de datos. *Revista de la Facultad de Ingeniería Mecánica y Eléctrica.* Universidad Autónoma de Nuevo León, Vol. VIII, No. 26, México.
- Reyes, R. y Salgueiro, Y. (2010) Herramienta para realizar la Minería de Datos en el Data Warehouse de un Cuadro de Mando Integral. *Innovación Tecnológica*. Vol. 16. Núm. 2.
- Rodríguez, O (2010). *Metodología para el Desarrollo de Proyectos en Minería de Datos* CRISP-DM
- Robert, C., Mobasher, B., Srivastava, J., (Febrero, 1999). Data Preparation for Mining World Wide Web. *Knowledge and Information Systems.* Volume 1, Issue 1, 5-32.
- Russell, S.; Norvig, P.,(2002). *Artificial Intelligence: A Modern Approach.* Pearson Education Limited, Harlow, Inglaterra.
- Scotto, M. Sillitti, A. Succi, G. Vernazza, T., (2004) *Managing Web-Based Information*, International Conference on Enterprise Information Systems (ICEIS 2004), Porto, Portugal, Page 1-3.
- Salton, G.; McGill, M. J., (1983). *Introduction to modern information retrieval.* McGraw Hill, New York.
- Servente, M., R. García-Martínez, (2002). *Algoritmos TDIDT aplicados a la minería de datos inteligente*. Tesis Doctoral Universidad de Buenos Aires, Argentina.
- Song, Q., Shepperd, M., (2005) *Mining web browsing patterns for E-commerce*. Xi'an Jiaotong University, China, Brunel University, UK.
- Sosa, M. O., & Sosa Bruchmann, E. C. (2014). *Estudio de técnicas de Data Mining aplicadas al análisis de datos generados con la metodología Blended Learning*. In XVI Workshop de Investigadores en Ciencias de la Computación.

- The Webalizer Features (2015). Webalizer. En http://www.mrunix.net/webalizer/ Extraído el 30 de Octubre del 2015

- Torres, M. (2015). *Análisis de algoritmos de aprendizaje automático para la caracterización de usuario de la Web*. *Depto. Ingeniería de Sistemas Telemáticos* – ETSIT – Universidad Politécnica de Madrid, Madrid, España.

- Villena, J., Barceló, E., Velasco, J., (2002), *Minería de uso de la web mediante huellas y sesiones*, *Depto. Ingeniería de Sistemas Telemáticos* – ETSIT – Universidad Politécnica de Madrid, Madrid, España.

- Waikato ML Group. *Attributerelation file format* (arff).En http://www.cs.waikato.ac.nz/ml/weka/arff.html. Extraído el 15 de Septiembre del 2015

- Waikato ML Group. *The waikato environment for knowledge analisys* en http://www.cs.waikato.ac.nz/ml/weka. Extraído el 30 de Octubre del 2015

- Waitten, I., Eibe F., (2001) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* En http://www.sigmod.org/sigmod/record/issues/0203/bookreview2-geller.pdf Extraído el 3 de Diciembre del 2015.

- Dürsteler, J.C.: Minería Web. Revista digital de InfoVis.net, 2005. http://www.infovis.net/printMag.php?num=172&lang=1