CIS1310IS02

Aplicación de Minería de Datos para la Identificación de Patrones de comportamiento en las organizaciones enfocado en Prácticas de Impresión: Caso de Estudio

Daniel Augusto Solano Oviedo

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERIA

CARRERA DE INGENIERIA DE SISTEMAS

BOGOTÁ, D.C.

2013

CIS1310IS02

Aplicación de Minería de Datos para la Identificación de Patrones de comportamiento en las organizaciones enfocado en Prácticas de Impresión: Caso de Estudio

Autor:

Daniel Augusto Solano Oviedo

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TITULO DE INGENIERO DE SISTEMAS

Director

Álvaro Fernando Quintero González

Jurados del Trabajo de Grado

Julio Ernesto Carreño Vargas

Blanca Elvira Oviedo Torres

Página web del Trabajo de Grado

http://pegasus.javeriana.edu.co/~CIS1310IS02

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
CARRERA DE INGENIERIA DE SISTEMAS
BOGOTÁ, D.C.
NOVIEMBRE, 2013

PONTIFICIA UNIVERSIDAD JAVERIANA

FACULTAD DE INGENIERIA

CARRERA DE INGENIERIA DE SISTEMAS

Rector Magnífico

Joaquín Emilio Sánchez García S.J.

Decano Académico Facultad de Ingeniería

Ingeniero Jorge Luis Sánchez Téllez

Decano del Medio Universitario Facultad de Ingeniería

Padre Sergio Bernal Restrepo S.J.

Director de la Carrera de Ingeniería de Sistemas

Ingeniero Germán Alberto Chavarro Flórez

Director Departamento de Ingeniería de Sistemas

Ingeniero Rafael Andrés González Rivera

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

Contenido

CONTENIDO **15**

INTRODUCCIÓN **24**

I - DESCRIPCION GENERAL DEL TRABAJO DE GRADO **25**

1. OPORTUNIDAD, PROBLEMÁTICA, ANTECEDENTES **25**

1.1 Descripción del contexto *25*

1.2 Formulación del problema planteado *26*

1.3 Justificación *26*

1.4 Impacto Esperado *26*

2. DESCRIPCIÓN DEL PROYECTO **27**

2.1 Visión global *27*

2.2 Objetivo general *27*

2.3 Fases Metodológicas o conjunto de objetivos específicos *27*

II - MARCO TEÓRICO **28**

1. MARCO CONTEXTUAL **28**

2. MARCO CONCEPTUAL **29**

2.1 Historia *29*

2.2 Conceptos Básicos *31*

2.3 Conceptos sobre la problemática *32*

2.4 Técnicas Minería de Datos *34*

2.4.1 Las técnicas dirigidas *35*

2.4.2 Las técnicas no dirigidas *38*

2.5 Herramientas Minería de Datos *40*

2.5.1 Software Libre *41*

2.5.2 Software licenciado *45*

3. MARCO INSTITUCIONAL **46**

III – DESARROLLO DEL TRABAJO 48

FASE 1 ENTENDIMIENTO DEL NEGOCIO 48

Determinar los objetivos del negocio 49

Evaluar la situación 50

Elaborar el plan del proyecto 51

FASE 2 ENTENDIMIENTO DE LOS DATOS 51

Recopilar los Datos iniciales 51

Descripción de los Datos 52

Revisar los Datos 58

Verificar la calidad de datos 58

FASE 3 PREPARACIÓN DE LOS DATOS 58

Seleccionar los Datos 59

Limpieza de los datos 59

Construcción de los datos 61

Aplicar formatos a los datos 63

FASE 4 MODELADO 65

Seleccionar la técnica de modelado 66

Construcción del modelo de pruebas 69

Implementación del modelo 71

Evaluación del modelo 71

FASE 5 EVALUACIÓN 80

Evaluación de los resultados 80

Revisión del proceso 81

Determinar los próximos pasos 81

FASE 6 TRANSFERENCIA 82

Plan de transferencia 82

Producción del reporte final 83

Revisión del Proyecto 83

IV - RESULTADOS Y REFLEXIÓN SOBRE LOS HALLAZGOS 84

V – CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS 85

1. CONCLUSIONES 85

2. RECOMENDACIONES 86

3. TRABAJOS FUTUROS 87

VI - REFERENCIAS Y BIBLIOGRAFÍA 87

1. REFERENCIAS 87

2. BIBLIOGRAFÍA 91

VII - ANEXOS 92

ANEXO1. GLOSARIO 92

ANEXO2. POST-MORTEM 92

ACTAS DE REUNIÓN 92

REPORTES MEGATRACK 93

- *Registros-MegaTrack 93*
- *Registros-MegaTrack-BuenasMalas 93*
- *Análisis impresión Buenas-Malas 93*

ARQUITECTURA DE LA SOLUCIÓN 93

DESCRIPCIÓN DIAGRAMAS TABLA DE HECHOS 93

MANUALES 93

- *Manual de Usuario 93*
- *Manual de instalación 93*

ARCHIVOS HERRAMIENTA WEKA 93

- *Registros-Dirigido 93*
- *Registros-NoDirigido 93*

CRONOGRAMA – PLAN DE TRABAJO PROYECTO 93

CARTA CLIENTE – PRINTER ON LINE INTEGRAL DOCUMENT SAS 93

PRESENTACIÓN TRABAJO DE GRADO 93

Tablas

Tabla 1: Descripción de los Datos 57

Tabla 2: Limpieza de los Datos 60

Tabla 3: Construcción de los Datos Impresora 61

Tabla 4: Construcción de los Datos Tiempo 63

Tabla 5: Resultados Clúster 75

Ilustraciones

Ilustración 1: Marco Contextual 28

Ilustración 2: Marco Conceptual 29

Ilustración 3: Conceptos Básicos 31

Ilustración 4 : Conceptos sobre la problemática 32

Ilustración 5 : Técnicas de Minería de Datos 34

Ilustración 6: Herramientas Minería de Datos 40

Ilustración 7 : Fase 1 Entendimiento del Negocio 49

Ilustración 8 : Fase 2 Entendimiento de los Datos 51

Ilustración 9: Versión Final Tabla de Hechos 52

Ilustración 10: Fase 3 Preparación de los Datos 59

Ilustración 11: Reportes MegaTrack 64

Ilustración 12: Archivo .arff 65

Ilustración 13: Fase 4 modelado 66

Ilustración 14: Escenario Técnicas No Dirigidas 68

Ilustración 15: Escenario Técnicas Dirigidas 68

Ilustración 16: Análisis impresión es Buenas - Malas 70

Ilustración 17: Explorador Weka 71

Ilustración 18: Primer Resultado Asociación 77

Ilustración 19: Segundo Resultado Asociación 77

Ilustración 20: Tercer Resultado Asociación 77

Ilustración 21: Fase 5 Evaluación 80

Ilustración 22: Fase 6 Transferencia 82

ABSTRACT

This document describes a data mining process, from the understanding of the business to the results analysis. The objective of this study is identifying behavior patterns related to printing practices among medium and large organizations. The purpose of the study is to reduce the consumption of resources used in daily routine work. The problematic and opportunities for this work is explained thoroughly in the document description, as well as the development process and the methodologies used.

RESUMEN

El presente documento describe el proceso de desarrollo de un estudio de minería de datos, desde el entendimiento del negocio hasta el análisis de resultados. El objetivo de este estudio es identificar si existen patrones de conducta en las medianas y grandes organizaciones relacionadas con prácticas de impresión, lo anterior con el fin de reducir el consumo excesivo e innecesario de recursos que diariamente se utilizan en las actividades rutinarias de trabajo. En la descripción del documento se explica ampliamente la problemática y oportunidad para este trabajo de grado, así como el proceso de desarrollo y las metodologías utilizadas.

RESUMEN EJECUTIVO

En los últimos años, nació la iniciativa en algunas empresas por la conservación del medio ambiente. Esta nueva forma de pensar es inculcada en los trabajadores para mejorar sus actividades rutinarias dentro de las organizaciones y consecuentemente mejorar su la calidad de vida, así como promover el cuidado del medio ambiente. Las empresas que buscan este objetivo emplean diferentes técnicas o estrategias para realizarlo, desde campañas en pro del manejo de desperdicios hasta la utilización de recursos que no sean perjudiciales para la naturaleza. Los ejemplos más sobresalientes sobre este tipo de campañas se pueden enfocar en el correcto uso de los recursos en los baños y de los ascensores, así como los recursos que se brindan en las cafeterías y salas de libre esparcimiento. No obstante, rara vez se ataca una problemática muy evidente, considerando que es un recurso esencial en las actividades laborales de cualquier organización; estamos hablando de las prácticas de impresión. Los recursos utilizados para estas prácticas, incluido el dinero, representan un valor significativo que debe ser analizado. Aun considerando los avances tecnológicos y las herramientas que permiten digitalizar cualquier tipo de documento, la impresión física de documentos es un recurso muy utilizado actualmente. Las medianas y grandes organizaciones emplean altas cantidades de dinero para brindarles esta posibilidad a sus empleados, donde su consumo es desmedido y no maneja un control ni por parte de los empleados ni de las mismas organizaciones; es precisamente del análisis de esta situación donde nace el planteamiento del presente trabajo de grado.

Algunas empresas ya empezaron a emplear sistemas para controlar este tipo de prácticas, por medio de la implementación de un software que permite llevar un control de impresiones y una estadística de consumo. Es en este punto donde nace una segunda inquietud. ¿De qué sirve llevar un control estadístico de consumo de impresión y recolectar información relevante para la organización si no se van a utilizar estos datos para mejorar la situación actual de la empresa? Es sobre estos dos cuestionamientos donde nace la propuesta de este trabajo de grado.

La propuesta consiste en realizar un estudio de minería de datos que permita identificar patrones de conducta o comportamiento de los empleados en medianas y grandes empresas en

el momento de realizar prácticas de impresión. El objetivo es muy claro, con el fin de brindarle a las empresas las herramientas necesarias para la toma de decisiones sobre estrategias y/o campañas que mejoren el consumo de recursos de impresión, resulta necesario conocer la situación que se desarrolla internamente en la empresa. Para poder terminar de aterrizar el problema y el planteamiento de la solución, se realizaron los siguientes cuestionamientos. ¿Es posible reducir el consumo de impresión en una compañía mejorando la utilización del servicio?, ¿Cómo los empleados de una compañía utilizan el servicio de impresión?, ¿Qué es una buena práctica de impresión? / ¿Qué es una mala práctica de impresión?, ¿Qué usuarios deben tener acceso al servicio de impresión?, ¿En qué casos se debe realizar una impresión en color? / ¿En qué casos de debe realizar una impresión en blanco y negro?, ¿Debe existir un límite de páginas impresas al mes por centro de costos?, ¿Todos los centros de costos deberían tener límite de páginas impresas por mes?, ¿Debe existir un límite de páginas impresas al mes por usuario?, ¿Todos los usuarios deberían tener ese límite de páginas por mes? Una vez claras las metas del proyecto y los objetivos a alcanzar se realizaron los planteamientos de la solución.

La metodología implementada para el desarrollo del estudio de minería de datos fue la metodología CRISP-DM, la cual está conformada por 6 fases: Entendimiento del negocio, entendimiento de los datos, preparaciones de los datos, modelado evaluación y finalmente transferencia. Cada una de estas fases está compuesta por un conjunto de actividades que permiten llevar a cabo el análisis de los datos y de esa forma cumplir con el propósito del proyecto. En las primeras etapas de desarrollo del proyecto fue fundamental la participación activa del cliente donde se llevó a cabo el estudio. Lo anterior, en la medida que era el encargado de brindar la información para el entendimiento del negocio y su familiarización, así como de suministrar los datos para su posterior análisis, razón por la que se realizaron varias reuniones hasta lograr el objetivo de estas primeras fases del estudio.

Una vez terminado el entendimiento del negocio y de los datos, se avanzó con la preparación de los mismos. Paralelamente se realizaron ejercicios solucionados con técnicas de minería de datos, con el objetivo de lograr una mejor preparación para el presente estudio, poder elegir la técnica más adecuada al problema y finalmente seleccionar la herramienta que será utilizada en las siguientes etapas.

Debido a todos los factores involucrados en el entendimiento del negocio y el análisis del problema se plantearon dos escenarios y consecuentemente dos técnicas de minera de datos. Se utilizó la técnica de árboles de decisión por el lado de las técnicas dirigidas y la detección automática de clúster por el lado de técnicas no dirigidas, cada escenario con su respectivo archivo de datos y la misma herramienta para el análisis. La herramienta utilizada para el estudio fue Weka 3.6 [31], debido a que es de software libre y a la sencillez en la instalación, su configuración y manejo de la misma.

Se presentó un inconveniente en la etapa de recolección de datos. Inicialmente se planteó que los datos que serían analizados corresponderían a la información obtenida del software de control de impresión (MegaTrack) y los datos de los usuarios. Debido a la confidencialidad de la información de los usuarios, respaldados por la ley de protección de datos [9] no fue posible contar con la misma, por lo que fue necesario limitar los datos a los extraídos por la herramienta MegaTrack.

Una vez superado este inconveniente se continuó con el modelado y posterior análisis de los datos. El resultado del estudio de minería de datos no fue el esperado en el planteamiento de la solución, los resultados analizados después de implementar las técnicas de minería de datos no arrojaron información trascendental que pudiera ser determinante para la organización. La razón de este resultado es posiblemente la limitante en la utilización de los datos, más que por las técnicas de minería de datos utilizadas. Para concluir, aunque los resultados no fueron del todo los esperados, se abre la ventana a posibles soluciones de un tema de gran importancia y de interés para todas las empresas. Como se describió en la problemática, no sólo concierne a una cuestión monetaria, sino a la lucha por la conservación del medio ambiente, que también juega un papel fundamental en esta situación.

INTRODUCCIÓN

En este documento se presenta toda la información del proceso de desarrollo del trabajo de grado. La primera parte consiste en la descripción del problema donde está plasmada la justificación y la razón de ser del proyecto, así como todo lo relacionado con la visión global del proyecto y las oportunidades que se encontraron después del análisis del problema. Se especifica el objetivo general y los objetivos específicos de la investigación. Una vez definida la meta del trabajo de grado, se explica la metodología que se empleará para dar solución al problema planteado, la cual está directamente relacionada a cada uno de los objetivos específicos.

El documento contiene el análisis y las actividades realizadas en cada una de las etapas que componen la metodología, desde el entendimiento del negocio hasta la fase de transferencia de la información. Una vez concluida la etapa de desarrollo, la tarea a seguir corresponde al análisis de los resultados, donde quedará plasmada toda la capacidad analítica del estudiante y la forma cómo se interpretaron los datos de acuerdo a las hipótesis planteadas en la parte inicial de la investigación. Finalmente, la última parte del documento corresponde a las conclusiones de la investigación, soportadas por las referencias bibliográficas y los anexos correspondientes.

Es importante resaltar que durante toda la investigación se realizaron diversas reuniones de seguimiento tanto con el director de trabajo de grado como con el cliente, quien es la entidad más interesada en buen desarrollo del proyecto. Cada una de estas reuniones están respaldas con una acta de reunión que podrá consultarse en los documentos anexos al trabajo de grado. Otros documentos que se encuentran plasmados en los anexos del trabajo de grado son los reportes o fuente de información brindados por el cliente para su posterior análisis, estos archivos se encuentran disponibles para su consulta, tanto los iniciales como los transformados para ser leídos por la herramienta de minería de datos.

I - DESCRIPCION GENERAL DEL TRABAJO DE GRADO

1. Oportunidad, Problemática, Antecedentes

1.1 Descripción del contexto

La idea para el planteamiento de la presente propuesta de trabajo de grado nace de la situación actual de las medianas y grandes empresas, donde el consumo excesivo de los recursos que las empresas brindan a sus empleados es mal utilizado. Malgastar los recursos de esa forma no sólo trae pérdidas económicas a las empresas, sino que también perjudica notablemente al planeta, ya que entre más recursos se utilicen más recursos necesitarán ser generados y como todos sabemos, la materia prima se obtiene de la madre naturaleza.

Esas malas prácticas de utilización de recursos tienen varias razones, la principal es el desconocimiento de las consecuencias que estas prácticas traen y el no saber cómo utilizar de una manera correcta esos recursos; otra razón es que la mayoría de estas medianas y grandes empresas no cobran a sus empleados por la utilización de esos recursos, ya que para cumplir con sus obligaciones laborales es necesario la utilización de los mismos, en consecuencia los empleados no crean un sentido de pertenencia por el buen manejo de los recursos que reciben.

Esta propuesta de trabajo de grado pretende acotar o limitar el problema mejorando las prácticas de impresión; descubriendo, por medio de minería de datos, las malas conductas o prácticas que los empleados realizan en sus actividades diarias malgastando los recursos de impresión de las empresas (papel y tinta).

Una vez planteada la situación actual de las empresas, es necesario verificar o analizar si esta podría ser una oportunidad de trabajo, es decir, si se puede hacer algo para cambiar o mejorar esta situación. Afortunadamente, hoy en día la mayoría de las empresas se están preocupando por cambiar la forma de utilizar los recursos y cada vez son más las que implementan campañas de ahorro y reciclaje para contribuir con la conservación del medio ambiente, de esta forma nace una cultura ecológica entre las compañías.

1.2 Formulación del problema planteado

A partir de lo expuesto anteriormente, nace la pregunta ¿Es posible mejorar el consumo de recursos relacionados a prácticas de impresión identificando conductas o patrones de comportamiento en las organizaciones a través de estudios de minería de datos?

1.3 Justificación

Existen varias herramientas implementadas en las organizaciones que se encargan de llevar el control de las impresiones que se realizan a diario, estas herramientas no sólo se encargan del monitoreo de las impresoras, también pueden llevar una estadística de consumo de las impresiones. El problema radica en que es inservible recolectar unos datos importantes para la organización si no se hace nada con ellos. Es por eso que la finalidad del presente trabajo de grado será realizar un estudio de minería de datos, donde los datos a analizar son los recolectados por las herramientas mencionadas anteriormente y a partir de esto poder establecer si existen patrones de consumo y/o comportamiento dentro de las organizaciones.

Los resultados de esta investigación podrían ser supremamente útiles para medianas y grandes empresas que estén destinando presupuestos altos a prácticas de impresión y de la misma manera, consumiendo mayores recursos de la naturaleza. Si los resultados son exitosos, es decir, si se identifican patrones de consumo y/o comportamiento, las empresas a las que se apliquen estos estudios, tendrán herramientas para tomar decisiones y plantear estrategias para mejorar el consumo de sus recursos.

Es importante resaltar que no sólo las empresas serán beneficiadas económicamente, sino que, además, desde el punto de vista de la Pontificia Universidad Javeriana también se estará cumpliendo una labor en la sociedad, contribuyendo en la conservación del medio ambiente.

1.4 Impacto Esperado

Considerando el mejor de los casos, es decir, el escenario exitoso del trabajo de grado, donde una vez concluida la investigación de minería de datos se encuentren unos patrones de consumo y/o tendencias dentro de una organización, se esperaría que el impacto en la población a la que va dirigida principalmente (medianas y grandes organizaciones) sea alto.

Así mismo, se esperaría que estas empresas puedan sacar provecho de los resultados obtenidos, implementando estrategias que cambien las prácticas actuales de impresión, disminuyendo los gastos de los recursos y aumentando los beneficios de la conservación del medio ambiente.

De otro lado, que el impacto sea de corto a largo plazo dependerá únicamente de las decisiones que tome la empresa y en el momento en que lo haga, ya que el presente trabajo de grado únicamente se encargara de buscar e identificar si existen patrones de consumo, resaltando que la implementación de estrategias o cambios dentro de la organización serán decisiones exclusivas de la empresa a la que se le hará la investigación.

2. Descripción del Proyecto

2.1 Visión global

El presente trabajo de grado recolecta todo el proceso que fue llevado a cabo para realizar el estudio de minería de datos El objetivo de este estudio es identificar conductas o patrones de comportamiento en medianas y grandes organizaciones relacionados con las prácticas de impresión y de ese modo poder establecer estrategias que reduzcan el consumo de recursos o materias primas necesarias para estas prácticas.

2.2 Objetivo general

El objetivo general del presente trabajo de grado es realizar un estudio de minería de datos con el fin de identificar si existen o no patrones de consumo relacionados con prácticas de impresión dentro de medianas y grandes empresas.

2.3 Fases Metodológicas o conjunto de objetivos específicos

En esta sección se presentan los objetivos específicos que se cumplirán durante el desarrollo del trabajo de grado.

- ✓ Estudiar y analizar cuál es la naturaleza del negocio, cuál es la situación actual y qué se desea solucionar con la presente investigación.
- ✓ Realizar el proceso de abstracción, recopilación y familiarización de datos.
- ✓ Realizar actividades que filtren y organicen los datos relevantes para la investigación.

- ✓ Seleccionar una o varias técnicas de modelado para el estudio de minería de datos.
- ✓ Evaluar y verificar que el modelo construido para el análisis de los datos sea el indicado.
- ✓ Organizar y presentar los resultados obtenidos a partir del estudio de minería de datos realizado.

II - MARCO TEÓRICO

En esta sección se presenta el marco teórico del presente trabajo de grado.

1. Marco Contextual

La ***Ilustración 1*** contiene el marco contextual del marco teórico.

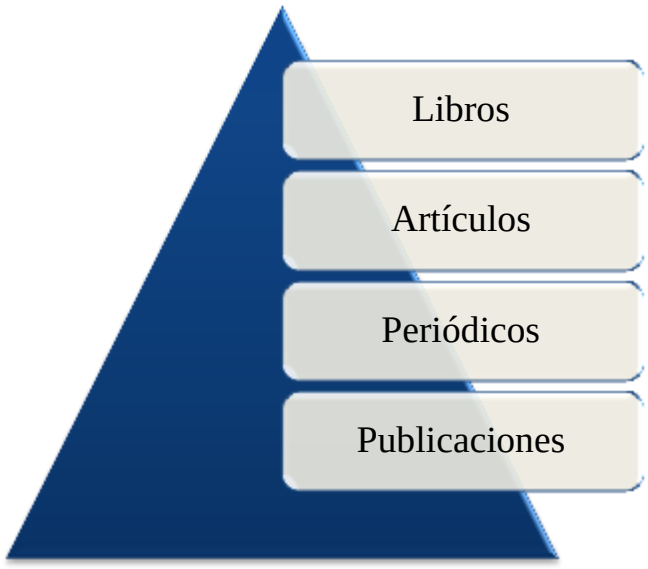


Ilustración 1: Marco Contextual

Los libros, artículos, publicaciones y demás documentos que sustentan la base teórica del estudio de minería de datos se encuentran descritos en las referencias y bibliografía del documento. Sección VI. Referencias y Bibliografía.

2. Marco Conceptual

La ***Ilustración 2*** contiene el marco conceptual del marco teórico.

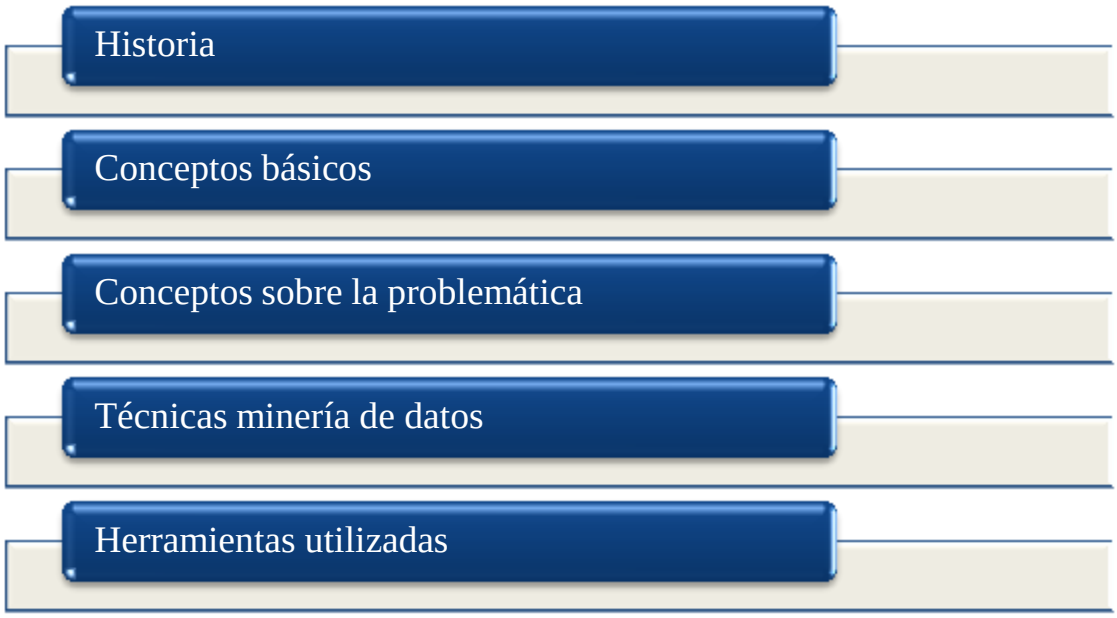


Ilustración 2: Marco Conceptual

2.1 Historia

La minería de datos es entendida como la búsqueda de patrones o comportamientos similares que se encuentren en bases de datos de una significativa amplitud. Lo anterior se logra con el soporte en áreas como modelos estadísticos y computación gráfica, que resulten en la identificación y análisis de estos patrones.

El nombre minería de datos se deriva de la relación entre el concepto "buscar" información relevante para una compañía en bases de datos de gran amplitud y "minar" una montaña de datos de manera que se encuentre información valiosa para la compañía o el negocio. Resultando en la similitud entre ambos procesos para buscar información significativa en amplias bases de datos. [1]

Sus inicios tienen lugar desde los años sesenta, donde se establece un vínculo entre la idea de identificar correlaciones que existieran entre datos pertenecientes a una base de datos sin

ruido, con el concepto de minería de datos, esto sin contar previamente con una hipótesis. Una base de datos sin ruido se refiere a documentos que son recuperados por el sistema pero que no representan mayor relevancia, como resultado de una estrategia de búsqueda superficial y aplicable a la mayoría de conceptos. [1]Así mismo, otros conceptos similares eran relacionados con esta descripción, estos son: "data fishing" y "data archaeology". Es importante mencionar, que la base de la minería de datos inicia una vez que las empresas inician el almacenamiento de datos por medio de computadores.

Fue en los inicios de los años ochenta, donde los académicos enfocados en algunos casos al área de ciencias computacionales: RakeshAgrawal, GioWiederhold, Robert Blum y Gregory Piatetsky-Shapiro, entre otros, empezaron a sentar bases más sólidas que permitieran afianzar el concepto de minería datos y KDD, por sus siglas en inglés, "Knowledge Discovery in Databases" o concepto de extracción de conocimiento. Este último concepto, KDD, se refiere al proceso de obtención de datos clave y por consiguiente al conocimiento significativo de una base de datos o repositorio de información. [2]

En la práctica, los términos Minería de datos y KDD suelen usarse como si fueran totalmente equivalentes; no obstante, es clave mencionar que la minería de datos corresponde a la etapa de descubrimiento dentro del proceso de KDD. [3]

Una vez, el término se empezó a afianzar con el concepto actual de minería de datos, la década de los ochenta finalizó únicamente con un par de empresas que ofrecieran esta tecnología. A medida que el tiempo y la sociedad avanzaron, conjuntamente la minería de datos se abrió campo, logrando ampliar el portafolio de empresas que ofrecen el servicio a más de cien, iniciando el año 2000.

Actualmente la minería de datos cuenta con soporte tecnológico fuerte que le permite tener madurez y confianza en sus usuarios, los tres soportes tecnológicos con los que cuenta son los siguientes:

- Recopilación de datos de forma masiva.
- Computadoras poderosas con multiprocesadores.
- Los algoritmos de minería de datos. [4]

2.2 Conceptos Básicos

La ***Ilustración 3*** contiene cada uno de los conceptos básicos necesarios para entender e significado de minería de datos.

Minería de datos	Datos
	Información
	Atributo
	Pronóstico
	Riesgo y probabilidad
	Tabla de hechos
	Modelo Estrella - Dimensiones
	Recomendaciones
	Búsqueda de secuencias
	Agrupamiento

Ilustración 3: Conceptos Básicos

La descripción de cada uno de los conceptos enunciados en el cuadro anterior será definida en el documento ***Anexo1. Glosario.***

2.3 Conceptos sobre la problemática

La ***Ilustración 4*** contiene los conceptos sobre la problemática planteada en el presente trabajo de grado. Estos conceptos están orientados a la lógica del negocio.

Conceptos sobre la problemática	Impresión
	Consumo de recursos
	Medianas / Grandes empresas
	Ley de protección de datos

Ilustración 4 : Conceptos sobre la problemática

Impresión

Reproducción de un texto o una ilustración en una imprenta, por medio de dispositivos de ordenadores o impresoras. [30]

Consumo de recursos

Una situación común dentro de las empresas es el consumo anormal de recursos ofrecidos por parte de sus empleados. Las empresas les permiten tener una libertad en el manejo de dichos recursos, resultando en algunos casos en un descontrol de la situación y consecuentemente afectando algunos procesos dentro de las compañías.

Un ejemplo de los recursos de las empresas está relacionado con la impresión de documentos. Esta facilidad está destinada para que los empleados de la empresa tengan la posibilidad de imprimir documentos relacionados con la operación de la misma y apoyarse en los mismos para el desarrollo de su papel dentro de la organización. No obstante, la cultura de impresión

controlada no es asumida por todos los empleados y en la mayoría de los casos se refleja un mal uso de esta facilidad.

Como resultado de un mal uso de este recurso la empresa se ve afectada en varios niveles. Los recursos de la empresa se ven afectados como el papel, la tinta, la luz, el uso de las máquinas. Lo anterior, en conjunto afecta negativamente a una organización.

Medianas/ Grandes empresas

De acuerdo con el Ministerio de Comercio, Industria y Turismo, o su sigla correspondiente "MinCIT", una mediana empresa corresponde a una cuya planta de personal se encuentre entre cincuenta y uno (51) y doscientos (200) trabajadores. De otro lado, sus activos totales deben ser entre cinco mil uno (5.001) a treinta mil (30.000) salarios mínimos mensuales legales vigentes. La clasificación del tamaño empresarial se puede realizar con al menos uno de estos criterios.

Las empresas grandes, son caracterizadas superando los límites superiores de los criterios para mediana empresa, es decir, los activos deben ser superiores a 30.000 salarios mínimos legales vigentes. [7]

El salario mínimo legal vigente para 2013 es de $589.500. [8]

Ley de Protección de datos

La ley estatutaria 1581 de 2012, por la cual se dictan disposiciones generales para la protección de datos personales y el objeto de la ley es el siguiente:" La presente ley tiene por objeto desarrollar el derecho constitucional que tienen todas las personas a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en bases de datos o archivos, y los demás derechos, libertades y garantías constitucionales a que se refiere el artículo 15 de la Constitución Política; así como el derecho a la información consagrado en el artículo 20 de la misma" [9]

2.4 Técnicas Minería de Datos

La ***Ilustración 5*** contiene las diferentes técnicas que pueden ser aplicadas en un estudio de minería a de datos.

Técnicas de minería	***Técnicas de minería de datos dirigidas***	Redes Neuronales
		Arboles de decisión
		Regresión
		Series Temporales
	Técnicas de minería de datos no dirigidas	Detección automática de Cluster
		Detección de desviaciones
		Segmentación
		Reglas de Asociación
		Patrones Secuenciales

Ilustración 5 : Técnicas de Minería de Datos

Las técnicas de minería de datos se dan como resultado de largos procesos de investigación y desarrollo de productos. Partieron cuando en un principio se inició el almacenamiento y análisis de datos por medios computacionales, permitiendo una mejora en el procesamiento de datos, en el acceso a la información y navegación por sistemas de información de manera más completa y con mayor acceso.

La minería de datos ha permitido evolucionar desde un concepto de análisis de datos dirigido al descubrimiento del conocimiento. Lo anterior partiendo de un tratamiento automatizado de los datos en evaluación y haciendo uso de los algoritmos pertinentes.

De esta manera, las técnicas de minería de datos se clasifican en dos grandes categorías: La minería de datos dirigida y la no dirigida. Las dirigidas también son conocidas como supervisadas y se caracterizan por predecir el valor de un atributo que pertenece a un

conjunto de datos. El procedimiento consiste en realizar la predicción de datos, donde su atributo es desconocido, partiendo de una relación existente entre este atributo desconocido y otros conocidos. La naturaleza de este algoritmo es predictiva.

De otro lado, se encuentra la minería de datos no dirigida, donde la base del algoritmo consiste en la identificación de patrones y tendencias de los datos actuales. Esta técnica no considera datos históricos, en la medida que no son considerados maduros. [10]

2.4.1 Las técnicas dirigidas

Redes Neuronales

Conocido en inglés como "Neural Networks", se refiere a grupos de unidades no-lineales que se encuentran interconectadas entre sí y organizadas por etapas. Este grupo de unidades pueden partir de funciones matemáticas o del almacenamiento de bases de datos en sistemas digitales como computadores. [3]

Son desarrollados bajo modelos matemáticos que permiten hacer computación inteligente. El desarrollo de modelos matemáticos permite llevar a cabo tareas y algoritmos en computadores que no se pueden usar usualmente, tales como reconocimiento de patrones, memorias y aprendizaje asociativo, predicción de series de tiempo, segmentación, entre otros. [11]

El procesamiento de los datos es distribuido a todos los procesadores o "neuronas" que realizan paralelamente el procesamiento computacional, resultando en una alta facilidad para un procesamiento masivo de datos y por lo tanto en un análisis exploratorio de los mismos. [12]

Árboles de decisión

En la mayoría de situaciones reales, ya sea en empresas o en individuos se presentan momentos donde se deben tomar decisiones que conllevaran a una serie de resultados y consecuencias finales. Estas decisiones deben partir de la derivación de una serie de eventos probabilísticos que en conjunto afectan el resultado final. No obstante, para el decisor estos

resultados de los eventos probabilísticos no se pueden conocer en el momento en que se debe tomar la decisión, de manera que el decisor debe fundamentar su decisión en las estimaciones de las probabilidades de los eventos aleatorios que están asociados al resultado final.

Con base en lo anterior, los árboles de decisión representan una herramienta de análisis para la estructuración y evaluación de situaciones que se encuentran bajo incertidumbre. La estructura principal de los árboles de decisión considera las diferentes alternativas a cada situación y asocia un evento probabilístico a las mismas, así mismo, presenta la secuencia del proceso de decisión y los resultados finales para cada ruta de acción. [13]

Los componentes de un árbol de decisión son los siguientes:

- Las alternativas de decisión o posibles cursos de acción para el decisor.
- Los eventos probabilísticos asociados al proceso de decisión, es decir, los que se encuentran bajo incertidumbre pero de alguna manera afectan el resultado final.
- La información de consecuencias relevantes, es decir, cómo afecta el resultado de esa decisión, por ejemplo económicamente.
- La secuencia del proceso de decisión o el orden en que debe ser estructurado el proceso y la relación entre esas decisiones y los eventos probabilísticos.

Una vez se estructure el árbol de decisión, considerando todos sus componentes, el decisor podrá evaluar cada curso de acción en conjunto y de esta manera soportar su decisión final a partir de dichas estimaciones y resultados.

Es clave mencionar que una de las consecuencias de aplicar esta técnica es que las variables son evaluadas independientemente una de la otra y no pueden ser descubiertas reglas de relación entre ellas. [14]

Regresión

La regresión hace referencia a una técnica estadística a partir de una ecuación matemática que establece como se relacionan las variables estudiadas. [15]

Las regresiones permiten hacer predicciones sobre una variable dependiente, usualmente denominada "y" a partir de unas o varias variables independientes usualmente denominadas "x", entre las que se existe relación. Lo anterior buscando siempre el menor error. [16]

Adicionalmente se puede afirmar que la forma aceptada y usada para predicciones continuas, es estructurarlo de manera que la salida o resultado sea una suma lineal de los valores que adaptan los atributos y cada uno ponderado de acuerdo al peso que le corresponda. Lo anterior es conocido como una regresión lineal y el proceso que permite identificar que peso le corresponde a cada atributo es conocido como el procedimiento estadístico denominado regresión. [17]

Series Temporales

Técnica referente a una secuencia de valores observados a lo largo de un periodo o tiempo, que se encuentran ordenados cronológicamente. Las series de tiempo son enfocadas en las series de datos en los que su próximo valor no puede ser definido con certeza, lo anterior es denominado una serie no determinista o aleatoria. La metodología tradicional para aplicar esta técnica descompone la serie en: tendencias, variación estacionalidad o periódica y otras fluctuaciones irregulares.

- La tendencia está relacionada con la dirección general de una variable en un periodo de observación, una forma de medir esta dirección es observando el cambio del promedio o media de la serie en un periodo largo de tiempo. Una medida para detectar y eliminar tendencias son los filtros, tales como las medias móviles.
- La estacionalidad corresponde a las fluctuaciones de una variable cada cierto periodo de tiempo, considerablemente cortos.
- Otras fluctuaciones irregulares corresponden a los valores residuales que resultan del proceso una vez sea observada la tendencia de la serie y sus variaciones por periodos de tiempo. [5]

2.4.2 Las técnicas no dirigidas

Detección automática de Clúster

También conocido como agrupamiento, consiste en la detección de grupos de individuos, es un aprendizaje no supervisado y no se conocen sus atributos, de manera que se busca determinar grupos o clústeres diferenciados del resto.

El objetivo es buscar grupos mutuamente excluyentes, buscando que cada dato dentro del grupo este lo más cercano posible a otros y por su parte, los grupos diferentes se encuentren alejados entre sí. [3]

El algoritmo divide un conjunto de datos en un número determinado de grupos, este número se conoce como "k", en la normalmente expresión para este algoritmo "k-means". Ésta técnica funciona mejor si la entrada del conjunto de datos es numérica. Es importante considerar que la técnica puede ser aplicada sin previo conocimiento del conjunto de datos ni de la estructura que va a ser descubierta y los clústeres resultantes son detectados automáticamente y podrían no representar otra interpretación natural. [14]

Detección de desviaciones

Consiste principalmente en detectar los cambios más significativos en el conjunto de datos a evaluar con respecto a valores pasados o con comportamiento normal. Su mayor uso es el filtro de altos volúmenes de datos que son menos probables de ser interesantes para el estudio. No obstante, la técnica requiere mayor concentración en determinar bajo qué punto o medida dicha desviación es significativa para aplicar dentro de los datos en consideración. [18]

Segmentación

Esta técnica consiste en la separación del conjunto total de datos en subconjunto o clases, las cuales pueden ser exhaustivas y exclusivas o jerárquicas y superpuestas. Esta técnica puede ser trabajada con otros algoritmos aplicables a cada clase considerada, tales como

"Clustering". El usuario usualmente cuenta con alta capacidad para formas estas clases, soportado en herramientas visuales. [18]

Reglas de Asociación

Reglas que implican relaciones entre un conjunto de objetos pertenecientes a una base de datos. Durante el proceso de establecer reglas de asociación se generan múltiples niveles de abstracción. [6]

Una vez se estudian y establecen dichas reglas de asociaciones dentro de una base de datos, se pueden identificar patrones de comportamiento, es decir, asociaciones entre los registros de datos. [18]

En conclusión las reglas de asociación son otra forma de generar reglas en el conjunto de datos. Sin embargo, en algunos casos las herramientas usadas generan reglas que resultan ser obvias y por lo tanto no ofrecen un valor agregado al análisis. [19]

Patrones Secuenciales

Referente al reconocimiento de patrones, basado en técnicas orientadas a evaluar la similitud y diferencia entre atributos o características de los datos evaluados. [3]

2.5 Herramientas Minería de Datos

La ***Ilustración 6*** contiene las herramientas utilizadas para la aplicación de técnicas de un estudio de minería de datos, las herramientas están organizadas en dos grupos: Software libre y software licenciado.

Herramientas minería de datos	***Software Libre***	Weka
		RapidMiner
		R
		Orange
		JHepWork
		KKIME
	Software Licenciado	DB2 INTELLIGENT MINER
		Statistica
		SPSS

Ilustración 6: Herramientas Minería de Datos

En los últimos años, dados los altos avances que se han logrado en este campo, la tecnología ha desarrollado amplias y sólidas herramientas que permiten a sus usuarios aplicar de la manera más adecuada las técnicas de minería de datos a conjuntos de datos pertenecientes a algún contexto.

Las herramientas desarrolladas pueden ser de software libre o licenciado. Dentro de las herramientas libres, se destacan las siguientes:

2.5.1 Software Libre

WEKA

El desarrollo de la herramienta Weka tuvo lugar en la Universidad de Waikato ubicada en Nueva Zelanda y el nombre corresponde a la abreviación conformada por las iniciales de "Waikato Enviroment for Knowledge Analysis". Se encuentra escrito en lenguaje Java y puede ser distribuido bajo los términos de Licencia Pública General, es decir, Software libre. Así mismo, ha sido probado para funcionar en ambiente Linux, Windows y Mancintosh.

La minería de datos es conocida como una ciencia experimental, que debe considerar diversos esquemas de aprendizaje en la medida que los conjuntos de datos varían entre sí. De esta manera, la herramienta Weka corresponde a un conjunto de algoritmos de aprendizaje y de pre y post procesamiento de datos. El diseño de la herramienta permite probar de manera ágil y flexible diversos algoritmos aplicables a esta ciencia. Adicionalmente proporciona soporte al usuario que va desde la preparación e inclusión de los datos de entrada y la evaluación de sistemas estadísticos, hasta los resultados del proceso para cualquier conjunto de datos.

Todo el proceso de interacción del usuario con la herramienta se realiza a través de una interfaz gráfica que le permite identificar y por lo tanto comparar los diferentes métodos. Consecuentemente, la herramienta brinda la posibilidad que el usuario determine los algoritmos de aprendizaje más apropiados para el tipo de datos sobre los que se está trabajando. Como resultado, el usuario puede observar todos los esquemas posibles por medio de un pre procesamiento del conjunto de datos y analizar los resultados conforme su desempeño bajo cada esquema, lo anterior sin necesidad de que el usuario desarrolle algún tipo de algoritmo.

Los métodos incluidos en la herramienta son: regresión, clasificación, "Clustering", reglas de asociación y selección de atributos. La entrada del conjunto de datos se realiza por medio de una tabla leída desde un archivo o creada a partir de una base de datos.

Existen tres formas de usar Weka:

- Aplicar al conjunto de datos un método de aprendizaje y analizar las salidas de manera que se pueda aprender más sobre los datos.
- Utilizar modelos aprendidos para generar predicciones.
- Aplicar varios "learners" y comparar su rendimiento de manera que se elija uno para la predicción.

Los esquemas de aprendizaje son el recurso más importante que ofrece la herramienta a sus usuarios.

De otro lado, el pre-procesamiento de los datos se realiza por medio de los llamados "filtros", seleccionados de un menú.

La interfaz le permite al usuario guiarse presentándole diversas opciones en un menú y desplegando las opciones aplicables a medida que se va avanzando en la selecciónLas interfaces disponibles en Weka son: "Explore", "The Knowledge Flow Interface", "Experimenter" y "Command-Line". La primera interfaz es aplicable a pequeños-medianos problemas y mantiene lo trabajado en una memoria principal, la segunda interfaz permite diseñar configuraciones para procesamiento de datos fluidos y la tercera interfaz está diseñada para ayudar a determinar qué métodos y parámetros funcionan mejor para el problema, en caso de utilizar técnicas de regresión y clasificación.

La funcionalidad básica de Weka se determina con las interfaces disponibles, así mismo, el acceso de los datos se puede hacer mediante comando de texto, dando camino a todas las funciones del sistema donde el usuario debe elegir entre las interfaces disponibles.

Por último, la herramienta se encuentra disponible en la web y puede ser descargada de una plataforma específica de instalación o un archivo ejecutable Java. [20]

RAPIDMINER

RapidMiner es otro entorno libre usado para minería de datos, que permite el análisis de datos por medio del encadenamiento de operadores en un entorno gráfico. Es considerada una herramienta líder en el mundo de código abierto y ampliamente usada en el mercado. La primera versión fue desarrollada por la universidad de Dortmund en 2001.

Dentro de sus características, se resaltan:

- Integración de datos.
- Transformación de metadatos, inspección de diseño de datos y metadatos.
- Reconocimiento de errores y propuestas de ajustes.
- Cuenta con una representación interna basada en archivos XML.
- Cuenta con una interfaz gráfica de usuario para el diseño de prototipos interactivos.
- Cuenta con una línea de comando para ser automatizado.
- Cuenta con una facilidad de Java que simplifica el uso de esta herramienta.
- Ofrece más de 500 operadores para algoritmos de aprendizaje, operadores de Weka, operadores de pre-procesamiento de datos, meta operadores, visualización y evaluación de desempeño.
- Desde 2010 posee integración con R (entorno de programación para análisis estadístico y gráfico), la cual es presentada a continuación de la descripción de la herramienta RAPIDMINER.

El conjunto de datos de entrada a la herramienta puede ser leído en diferentes formatos. Los formatos que principalmente maneja, tanto de archivo de lectura como de escritura son:

- Formatos .arff, resultantes del uso de "Arff Eaxmple Source"
- Formatos de salida de "Data Base Example Source"
- Archivos resultantes del operador "Example Source", donde la descripción de los atributos debe guardarse en archivo XML con extensión .aml [21]

Los operadores se pueden discriminar en los siguientes tipos:

- Operadores Básicos: Permiten aplicar, agrupar, desagrupar y actualizar modelo, así como operar cadena.
- Operadores de Core: Operador de línea de comando, definición de macros, experimento, salida de archivo múltiple y recuperación de entrada/salida, limpieza de memoria y proceso.
- Operadores de entrada/salida: permite leer registros y escribir valores.

- Operadores de aprendizaje: Operadores de pre y post procesamiento de datos, de visualización y de validación de desempeño.

Por su parte, la interfaz se compone de las siguientes partes principales:

- Árbol de procesos: área de entrada de definición modelo y operadores.
- Área de resultados: salidas de la corrida del modelo y de la configuración.
- Área de compilación y ejecución: log de procesos ejecutados. [22]

R

Lenguaje de programación y entorno de software libre para computación y gráficos estadísticos. Proporciona técnicas para simulación, modelado lineal y no lineal, análisis de series temporales, pruebas estadísticas clásicas, clasificación, agrupación en clústeres, entre otros. El usuario tiene acceso a funciones como análisis de datos, manejo y almacenamiento de datos, funciones gráficas de visualización y lenguaje de programación simple. [23]

ORANGE

Herramienta ambientada para programación visual o escritura C++ y Python. La herramienta funciona en Windows, Mac OS X y en diversos sistemas operativos Linux.

ORANGE contiene diversos componentes para pre-procesamiento de datos, característica de puntuación y filtrado, modelado, evaluación del modelo, y técnicas de exploración. [24]

JHepWork

JHepWork es una herramienta libre para análisis de datos y de código abierto. Tiene por objeto crear un entorno de análisis de conjunto de datos por medio de paquetes de código abierto con una interfaz accesible a los usuarios. Así mismo, está configurada para presentar diversos paquetes numéricos implementados en lenguaje Java que le permite al usuario acceder a funciones matemáticas, números aleatorios y otros algoritmos de minería de datos. [25]

KKIME

KNIME (Konstanz Information Miner) es una plataforma libre, comprensible para integración de datos, procesamiento, análisis y exploración. Ofrece a los usuarios la facilidad de crear flujos o tuberías de datos, así como de ejecutar selectivamente algunos o todos los pasos de análisis, para finalmente analizar los resultados y modelos. El lenguaje que utiliza es Java, basado en Eclipse. Los usuarios pueden añadir texto, imágenes y procesamiento de series de tiempo. Es importante mencionar que se puede integrar con otras herramientas libres como Weka. [25]

2.5.2 Software licenciado

DB2 INTELLIGENT MINER

Herramienta distribuida según la arquitectura cliente/servidor, es comercializada por IMB. Los paquetes que ofrece permiten soportar tareas de agrupamiento, asociaciones, patrones, clasificación y orientar al descubrimiento de relaciones entre los datos. [26]

STATISTICA

Herramienta creada por StatSoft. Permite amplias aplicaciones estadísticas y es utilizada en minería de datos. Cuenta con las siguientes características:

- Permite trabajar con un alto volumen de información. Las bases de datos pueden ser importadas desde formatos Excel, Oracle o SQL.
- Permite el pre-procesamiento de datos.
- Permite crear modelos de análisis.
- Permite la visualización.[27]

SPSS

Software que cuenta con herramienta visual desarrollada por ISL con una arquitectura diseñada cliente/servidor. Cuenta con las siguientes características:

- Acceso a datos.

- Procesamiento de datos.
- Técnicas de aprendizaje (reglas de asociación y redes neuronales).
- Resultados con visualización gráfica (histogramas, diagramas, gráficos).
- Informes de resultados en texto o html. [26]

3. Marco Institucional

Empresa: Pacific Rubiales Energy.

Pacific Rubiales es una compañía publicada, listada en bolsa de valores de Colombia y de Toronto. Es la empresa independiente dedicada a la exploración y producción de petróleo y gas más grande de Colombia. Pacific Rubiales es dueña del 100% de Pacific Stratus y Meta Petroleum Limited.

La creación de la empresa inicia en 2007, cuando Petro Rubiales logró un acuerdo con los dueños de Rubiales Holdings, donde vendía el 75% de sus acciones para consolidar posteriormente AGX Resources, que a su vez cambió el nombre a Petro Rubiales Energy Corp. Durante el mismo año Petro Rubiales adquirió el 25% restante de Rubiales Holdings. De otro lado, se encontraba Pacific Stratus Energy Corp, fundada en 2004 y se dedicada a la exploración. Ambas empresas encontraron que la mejor estrategia era su unión, de manera que pudieran prestar tanto el servicio de producción y exploración y trabajar conjuntamente con petróleo pesado y con gas natural. De esta manera, en 2008 y con la unión de ambas compañías, se creó Pacific Rubiales Energy Corp.

La compañía se encuentra enfocada en identificar oportunidades de crecimiento en el sector de hidrocarburos en Colombia, Perú y Guatemala.

En 2013 Pacific Rubiales logró niveles de producción de 310.000 barriles por día y cuenta con un aproximado total de 2300 empleados. Así mismo, en 2012 la compañía invirtió US $2.500 millones en contratación de servicios y obras civiles y tuvo ingresos por US$3.880 millones. De otro lado, es preciso mencionar para el desarrollo del proyecto que Pacific Rubiales cuenta con un promedio de 80 impresores en sus sedes administrativas de Bogotá.

Para efectos del caso de estudio se tomaran los registros de los empleados que trabajan en la sede Bogotá en la torre Pacific ubicado en la carrera novena con calle ciento diez. Los registros que serán analizados corresponden en promedio a 3000 usuarios entre empleados directos y contratistas que realizan prácticas de impresión. En promedio son 80 impresoras las que son monitoreadas por el software de impresión. Finalmente el número de áreas que son monitoreadas son las siguientes: Administración, Administrativo, AIT, Áreas Nuevas, Asuntos Corporativos, Auditoria, Calle110, Campo-Rubiales, Comercialización, Compras, Comunicaciones, Coordinación Operaciones, Coordinador Operacional, Exploración, Finanzas, Geociencias, Guaduas, Hseq, Legal, Logística, Oleoductos, Operación, Perforación, Presupuesto, Producción, Proyecto Star, Proyectos, Responsabilidad Social, SAP, SCM, Seguridad, Servicios, Servicios Generales, Talento Humano, Tesorería, Tic y Transporte.

Autor: Daniel Augusto Solano Oviedo

Institución: Pontificia Universidad Javeriana – Sede Bogotá

Facultad: Facultad de Ingeniería

Departamento: Ingeniería de Sistemas

III – DESARROLLO DEL TRABAJO

En esta sección se describe cuál fue el proceso para realizar el estudio de minería de datos del presente trabajo de grado. La metodología planteada para este proyecto es la metodología CRISP-DM. [28]

La metodología CRISP-DM es un proceso organizado en seis fases, cada una de estas fases están conformadas por un conjunto de tareas generales de segundo nivel. Las tareas generales se dividen en tareas específicas, que se realizan por medio de acciones para situaciones determinadas. Un ejemplo de una tarea de segundo nivel es la "limpieza de datos"; la tarea de tercer nivel para esa tarea general seria "limpieza de datos numéricos", o "limpieza de datos categóricos". Finalmente un cuarto nivel seria recoger el conjunto de resultados sobre el proyecto de minería de datos. [29]

La metodología CRISP-MD está compuesta de seis fases como se mencionó anteriormente, cada una de estas interactúan de forma iterativa durante el ciclo de vida del proyecto de minería de datos. La primera fase se encarga del análisis del problema, desde la perspectiva de negocio y como se orienta hacia la minería de datos. La segunda fase consiste en la recolección y familiarización de los datos, para que en la fase posterior se realice la preparación de los mismos. La cuarta fase consiste en el modelado del proyecto, en esta fase se selecciona la técnica de modelado más apropiada para la situación estudiada. En la quinta fase se realiza la evaluación del modelo utilizado desde el punto de vista de cumplimiento de los criterios de éxito establecido. Finalmente, la sexta y última fase consiste en la presentación y documentación de los resultados para lograr el incremento en el conocimiento del problema planteado. [29]

Fase 1 Entendimiento del Negocio

Esta primera etapa del estudio se enfoca en el entendimiento de los objetivos del proyecto, se verifica cuáles son los requisitos desde la perspectiva del negocio y de ese modo se puede plantear cuál sería la posible solución o hipótesis desde el campo de la minería de datos. [29] La ***Ilustración 7*** refleja las actividades que componen la fase de entendimiento del negocio.

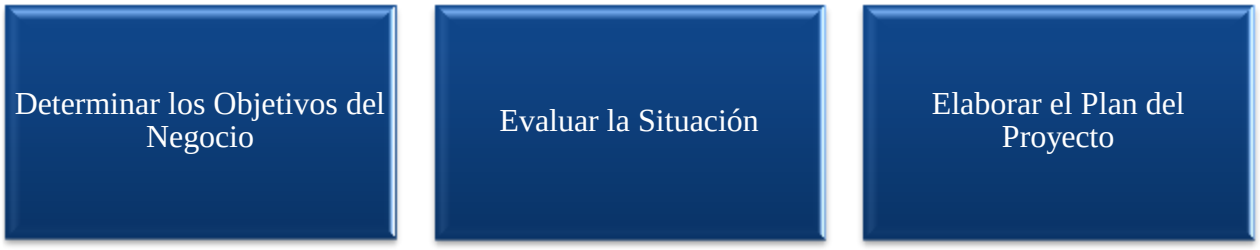


Ilustración 7 : Fase 1 Entendimiento del Negocio

Determinar los objetivos del negocio

Determinar cuál es el problema que queremos resolver

Básicamente el problema que queremos resolver con este estudio es poder disminuir el consumo excesivo de recursos que las organizaciones brindan a sus empleados para cumplir con sus labores rutinarias. Unos de los recursos peor utilizados en las empresas son los relacionados con las prácticas de impresión, por lo que este estudio se concentrará en intentar resolver el consumo innecesario de recursos como el papel, tinta y en general las partes utilizadas en una impresora.

¿Por qué usamos minería de datos para este propósito?

La minería de datos a través de sus diferentes técnicas nos permite identificar situaciones, conductas o comportamientos que no podemos identificar a simple vista. Para poder establecer una solución al problema detectado es primordial identificar qué es lo que está sucediendo dentro de las organizaciones, de esa manera poder establecer estrategias para modificar o corregir lo que sea necesario. La minería de datos es una herramienta muy útil para darle un uso práctico y adecuado a la recolección de datos estadísticos que sean relevantes para definir el comportamiento de usuarios dentro de una organización.

¿Cuáles son los criterios de éxito?

Los criterios de éxito fueron definidos a partir de la premisa de la justificación del problema, El principal criterio de éxito para el presente trabajo de grado seria poder identificar conductas o patrones de comportamiento desconocidas tanto para la empresa como para los

administradores del servicio de impresión una vez concluida la investigación. Igualmente existe la posibilidad de no llegar a este objetivo después de realizar el estudio de minería de datos. Un segundo criterio de éxito seria poder confirmar, después de realizada la investigación, las hipótesis y teorías evidentes sobre el aumento del consumo del servicio de impresión, estas podrían ser actividades de cierre de mes o la realización de impresiones para fines no laborales.

Evaluar la situación

Situación actual desde el punto de vista del negocio

La idea para el planteamiento del presente trabajo de grado nace de la situación actual de las medianas y grandes empresas, donde los recursos no están siendo utilizados adecuadamente. No utilizar los recursos adecuadamente genera pérdidas económicas a las empresas y a su vez contribuye con la contaminación del medio ambiente, ya que entre más recursos se utilicen más recursos naturales serán empleados para suplir esta necesidad.

Situación actual desde el punto de vista de la minería de datos

Un requisito fundamental para poder realizar un estudio de minería de datos es contar con la recopilación de datos asociados al tema que se está investigando; para el caso particular de la presente investigación, datos estadísticos de consumo de impresión. Una forma en la que las empresas puedan realizar este almacenamiento de información es contar con un software de control de impresión. La empresa donde se realizará el estudio de minería de datos cuenta con un software que se encarga de la administración del servicio de impresión; este software se llama MegaTrack, la función que cumple es recopilar todos los datos relacionados a las impresiones que realizan sus empleados, el objetivo es llevar una estadística de consumo y de esa forma poder llevar una contabilidad de los recursos que se están utilizando. Los datos recopilados por este software serán la fuente de información que se utilizará para el estudio de minería de datos.

Elaborar el plan del proyecto

El plan del proyecto para el presente estudio de minería de datos lo podrá encontrar en el ***Anexo cronograma – Plan de trabajo Proyecto***. En este archivo se describe todos los pasos necesarios, desde el planteamiento del problema, recolección de datos, hasta el análisis de los mismo.

Fase 2 Entendimiento de los Datos

La segunda etapa consiste en la recopilación de datos y la familiarización con los mismos, es decir se reunieron todos los datos útiles para la investigación. Este es el primer acercamiento a los datos, se revisara con detenimiento la calidad de la información. [29] La ***Ilustración 8*** refleja las actividades que componen la fase de Entendimiento de los datos.

Ilustración 8 : Fase 2 Entendimiento de los Datos

Recopilar los Datos iniciales

Antes de empezar con la recopilación de datos se realizó el ejercicio de creación de un modelo dimensional y posteriormente la creación de una tabla de hechos, lo anterior con el objetivo de contemplar todos los elementos que involucran la realización de una impresión. Una vez creada la tabla de hechos se define qué información será relevante para la investigación y cómo será el proceso de recolección de datos. El proceso para crear la tabla de hechos, con todas sus versiones, se encuentra en el ***Anexo Documento Descripción de Tacha de Hechos.***

La ***Ilustración 9*** contiene el diagrama de la versión final de la tabla de hechos.

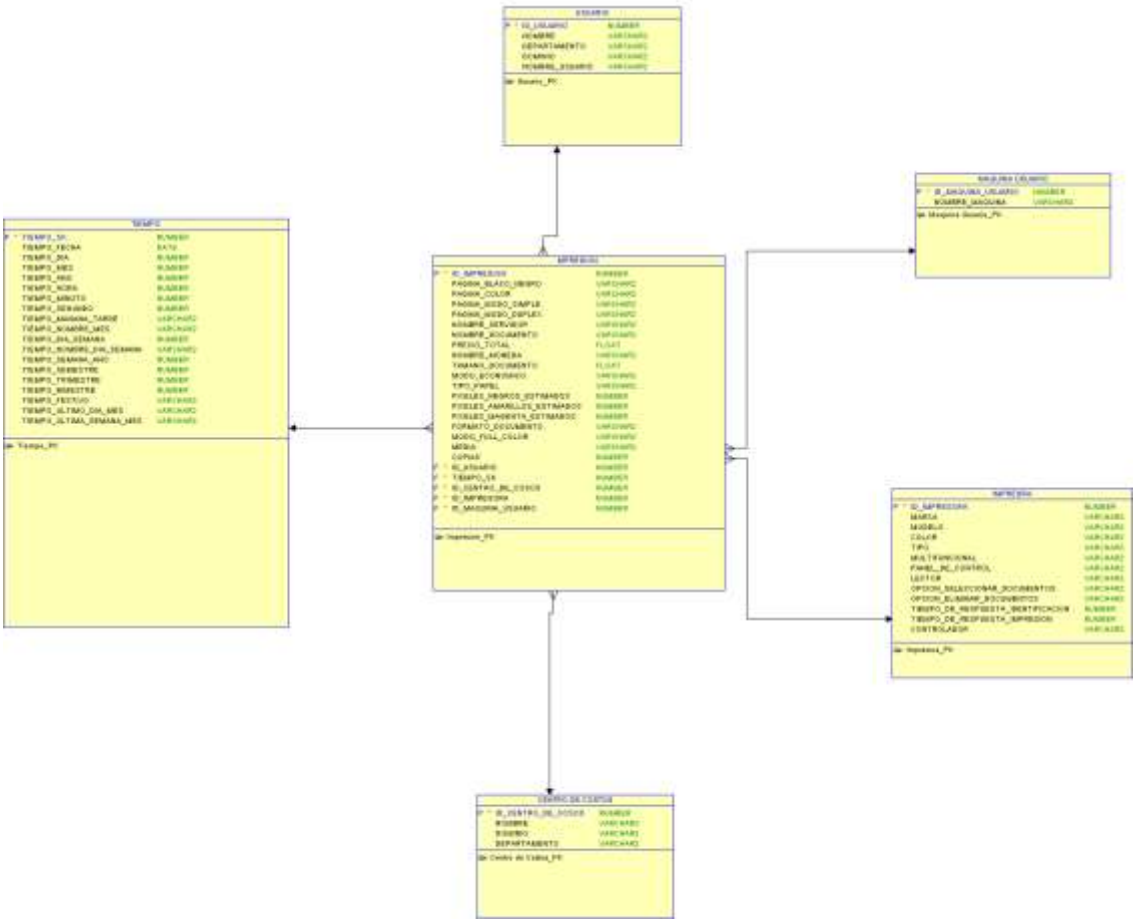


Ilustración 9: Versión Final Tabla de Hechos

Descripción de los Datos

En esta sección vamos a describir los datos iniciales obtenidos, tales como número de registros y campos por registros, su identificación, el significado de cada campo y la descripción del formato inicial.

La ***Tabla 1*** contiene la descripción de los datos recopilados para el presente estudio de minería de datos.

Nombre	Descripción
Nombre del Servidor	Nombre del servidor de impresión desde donde se obtiene la impresión.
Marca	La marca de la impresora (Para este cliente todas la impresoras son HP).
Modelo	Modelo de la impresora.
Color	Color de la superficie de la impresora.
Tipo	Tipo de funcionamiento de la impresora. Puede ser tipo Color o Monocromática.
Multifuncional	Especifica si la impresora es multifuncional.
Panel de Control	Especifica si la impresora cuenta con un panel de control.
Lector	Especifica si la impresora cuenta o no con lector. Este atributo hace referencia a la accesibilidad de la impresora.
Opción Seleccionar Documento	Específica si por medio de la impresora se puede seleccionar los documentos que desea imprimir; esto solo es posible si la impresora cuenta con un panel de control.
Opción Eliminar Documento	Específica si por medio de la impresora se pueden eliminar los documentos; lo anterior sólo es posible si la impresora cuenta con un panel de control.

Tiempo de Respuesta de identificación	Tiempo que la impresora necesita para reconocer la tarjeta del usuario desde el directorio activo.
Tiempo de Respuesta Impresión	Tiempo que la impresora se toma para realizar la impresión.
Controlador	Hace referencia al driver de la impresora utilizado por la compañía.
Copias	Especifica el número de impresiones que se realizaron por documento.
Número de páginas impresas	Número de páginas que tiene el documento.
Número de páginas impresas en modo simple	Número de páginas que tiene el documento en modo simple.
Número de páginas impresas dúplex	Número de páginas que tiene el documento en modo dúplex.
Número de páginas blanco y negro impresas	Número de páginas que tiene el documento en blanco y negro.
Número de páginas impresas en color	Número de páginas que tiene el documento en color.
Precio Total	El precio total del documento.

Tamaño del documento	El tamaño total del documento en MB.
Modo económico	Especifica si la impresión se realizó en modo económico o full color.
Nombre de la Moneda	El nombre de la moneda que se utilizó para calcular el valor total de la impresión.
Tipo de Papel	El tipo de papel que se utilizó para realizar la impresión. Puede ser las siguientes opciones: {'plain','unknownmedia','cardstock','colored','preprinted','letterhead','recycled','transparency','bond','labels','usertype4','usertype3','rough','usertype5'}
Media	El tamaño del papel que se utilizó para realizar la impresión. Pueden ser las siguientes opciones:{'carta (8.5x11 in)','a4 (8.27x11.7 in)','na','legal (8.5x14 in)','custom (105 x 241 in)','custom (197 x 273 in)','custom (216 x 330 in)','b4 (jis) (10.5x14.3 in)','custom (110 x 220 in)','c5 (6.385x9.02 in)','b5 (jis) (7.17x10.5 in)','executive (7.25x10.5 in)','a3 (11.7x16.5 in)','custom (100 x 148 in)','b (11x17 in)','b5 (6.93x9.84 in)'}
Dominio	Dominio de la compañía desde donde se realizó la impresión.
Nombre del documento	Nombre del documento impreso.
Nombre de grupo	Nombre del área donde se realizó la impresión.

Usuario	El nombre del usuario de la compañía.
Nombre	Nombre completo del usuario.
Departamento	Departamento dentro de la compañía donde se realizó la impresión.
Nombre de la Maquia	Nombre de la máquina del usuario donde se realizó la impresión.
Píxeles negros estimados	Número de pixeles negros utilizados en la impresión.
Píxeles amarillos estimados	Número de pixeles amarillos utilizados en la impresión.
Píxeles magenta estimados	Número de pixeles magenta utilizados en la impresión.
Fecha de impresión	Fecha en la que se realizó la impresión.
Día	Día del mes en que se realizó la impresión.
Mes	Mes en el que se realizó la impresión.
Año	Año en el que se realizó la impresión.
Hora	Hora en la que se realizó la impresión.
Minuto	Minuto en la que se realizó la impresión.

Segundo	Segundo en la que se realizó la impresión.
Mañana-Tarde	Especifica si la impresión se realizó en horas de la mañana o en horas de la tarde. Puede ser AM o PM.
Nombre Mes	Nombre del mes donde se realizó la impresión.
Día Semana	Día de la semana donde se realizó la impresión. Puede ser de 1 a 7 según corresponde.
Nombre día Semana	Nombre de día de la semana. Puede ser {Lunes, Martes, Miércoles, Jueves, Viernes, Sábado, Domingo}
Semana de año	Número de la semana del año donde se realizó la impresión.
Semestre	Número de semestre del año donde se realizó la impresión.
Trimestre	Número de Trimestre del año donde se realizó la impresión.
Bimestre	Número de Bimestre del año donde se realizó la impresión.
Festivo	Especifica si la impresión se realizó un día festivo o no.
Último día del mes	Especifica si la impresión fue realizada el último día del mes.
Última semana del mes	Especifica si la impresión fue realizada la última semana del mes.

Tabla 1: Descripción de los Datos

Revisar los Datos

Fue necesario realizar un cambio en el planteamiento inicial de la solución. En un principio se había considerado utilizar la información personal de los usuarios tales como: edad, cargo, número de hijos, edad de los hijos, funciones, estudios realizados, lugar de residencia, etc. Lo anterior para poder recolectar toda la información que pudiera ser relevante e influyente en las prácticas de impresión. Después de mantener varias reuniones con el cliente y por la ley de protección de datos se definió que definitivamente el uso de esta información personal de los usuarios estaba restringido.

El otro inconveniente que se detectó en la revisión de datos, es que algunos atributos no podrían ser recolectados ya que no era posible que la herramienta MegaTrack los proporcionara. Ejemplo como este son los datos de las máquinas de los usuarios donde se realizan las impresiones, se tenía contemplado poder obtener la información de marca, modelo, sistema operativo y si el equipo era propio o de la empresa. Por esta razón tuvimos que limitar los datos a la información que podía ser extraída por MegaTrack, esta información está explicada en la ***Tabla 1: Descripción de Datos*.**

Verificar la calidad de datos

En esta sección se realizó la verificación de los datos para determinar la consistencia de los valores de los campos, la cantidad y distribución de los valores nulos y para encontrar valores fuera de rangos que pueden generar ruido para el proceso.

Este proceso de verificación se realizó en todos los reportes extraídos por la herramienta. En los campos donde no se encontraban registros se cambió los campos vacíos por un valor null.

Fase 3 Preparación de los datos

En esta tercera fase se realizan todas las actividades que filtren y organicen los datos relevantes para la investigación, es decir seleccionar los datos que realmente sirven y desechar los que no. Un ejemplo de un dato que no es relevante para la investigación es el nombre del servidor, la razón es que este dato es constante para todos los registros por lo que

no causara ninguna diferencia o efecto en estudio [29] La ***Ilustración 10*** refleja las actividades que componen la fase de preparación de los datos.

Ilustración 10: Fase 3 Preparación de los Datos

Seleccionar los Datos

Se seleccionó una lista de 53 atributos como se especificó en la *Tabla 1: Descripción de los datos.* La cantidad de registros tomados fueron los correspondientes a un año de impresión en la compañía. Ya que el mes de noviembre del 2013 no ha terminado, el año empezará desde Noviembre del 2012 hasta Octubre de 2013 (12 meses). La cantidad de registros de impresión del rango de fecha definida fue de 611334.

Limpieza de los datos

Después de realizar la depuración de los datos mencionada en la sección anterior, donde se justificó la razón por la cual no se tendrán en cuenta atributos relacionados con la información personal de los usuarios y atributos que no es posible extraer utilizando la herramienta, fue necesario realizar la eliminación o limpieza de atributos que sólo generan ruido para la investigación. La ***Tabla 2*** muestra los atributos que no serán tomados en cuenta en la investigación con su respectiva justificación.

Nombre	Razón
Nombre del Servidor	Todos los registros de impresión fueron realizados desde el mismo servidor de impresión, por lo que este valor es el mismo en todos los casos. Lo que convierte en un atributo no relevante.
Tiempo de Respuesta de identificación	El tiempo de identificación siempre son 8 segundos. Lo que convierte en un atributo no relevante.
Copias	Para todos los casos el número de copias de cada documento fue 1. Lo que convierte en un atributo no relevante.
Nombre de la moneda	Ya que el estudio se realizó en la misma empresa el nombre de la moneda siempre fue PSC. Lo que convierte en un atributo no relevante.
Nombre del documento	El tipo de dato del documento en un string, y son muy pocos los registros que se repiten, por lo que este atributo no es significante para la investigación.
Nombre de la Máquina	El nombre de la máquina varia constantemente, por lo que este parámetro puede convertirse en un distractor para la investigación

Tabla 2: Limpieza de los Datos

Construcción de los datos

Los datos extraídos por la herramienta MegaTrack están en un formato determinado, fue necesario realizar una serie de tareas para construir los atributos requeridos por la solución planteada en la tabla de hechos. Por ejemplo en la caso particular de los atributos relacionados con la entidad impresora; a partir del dato impresora que es extraído de la herramienta MegaTrack, se construyen los datos como: Marca, Modelo, Color, Tipo, Multifuncional, Panel de control, Opción seleccionar documento, Opción eliminar documento, Tiempo de respuesta identificación y Tiempo de respuesta impresión. La ***Tabla 3*** muestra como fue el proceso de construcción de los atributos relacionados a la impresora.

Reporte MegaTrack	Atributo Generado	
HP Color LaserJet 4700 **HP Color LaserJet CM4730 MFP** **HP Color LaserJet CP4520 Series** **HP LaserJet M4345 MFP**	Marca	N.A.
	Modelo	Color
		Tipo
		Multifuncional
		Panel de Control
		Opción Seleccionar Documento
		Opción Eliminar Documento
		Tiempo de Respuesta de identificación
		Tiempo de Respuesta Impresión

Tabla 3: Construcción de los Datos Impresora

Para la construcción de los datos de la entidad tiempo se realizaron varias operaciones y formulas desde Excel, lo anterior debido a que el reporte de MegaTrack generaba únicamente un atributo (fecha), este atributo se debía transformar en todos los atributos relacionados el mismo. La ***Tabla 4*** muestra la forma en la que se realizó la construcción de los atributos relacionados a la entidad tiempo.

Reporte MegaTrack	Atributo Generado	Ejemplo
Fecha de impresión **20/09/2013 7:58:00**	*Día*	20
	Mes	9
	Año	2013
	Hora	7
	Minuto	58
	Segundo	0
	Mañana-Tarde	AM
	Nombre Mes	Septiembre
	Día Semana	5
	Nombre día Semana	Viernes
	Semana de año	38

	Semestre	2
	Trimestre	3
	Bimestre	5
	Festivo	No
	Último día del mes	No
	Última semana del mes	No

Tabla 4: Construcción de los Datos Tiempo

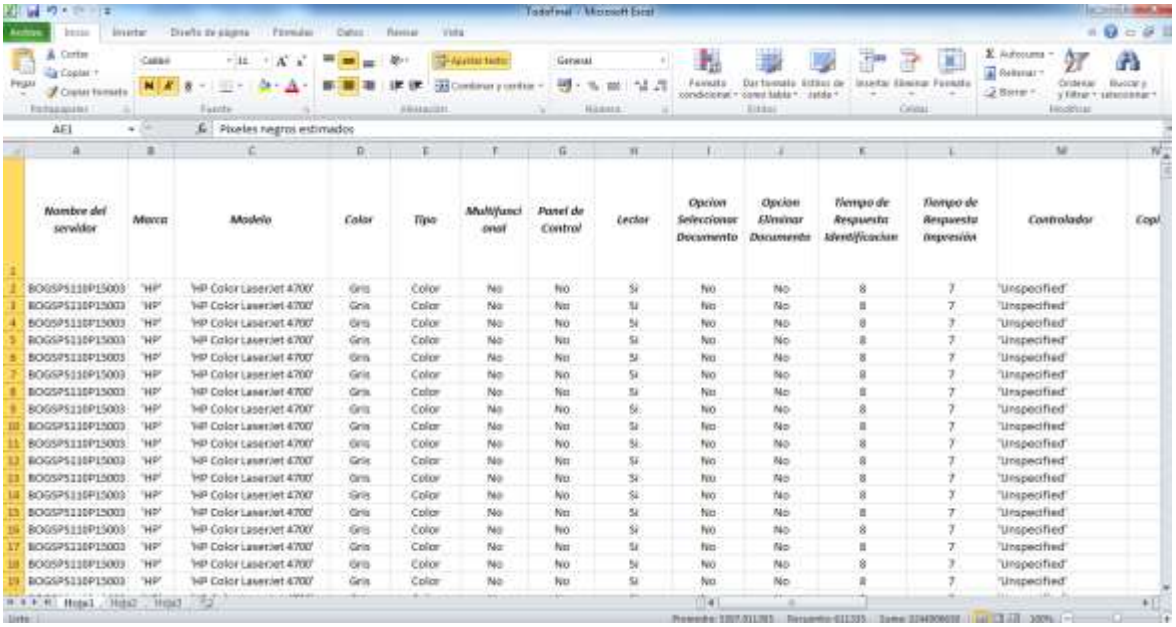
Para consultar las fórmulas de Excel empleadas para la integración de los datos consultar el documento ***Analisis_impresión_Buenas_Malas.xlsx*** en los anexos del trabajo de grado.

Aplicar formatos a los datos

En esta sección se explicará la forma en la que se realizó la transformación sintáctica de los datos sin modificar su significado. Los reportes que se extraen de MegaTrack son entregados por el cliente a través de archivos de Excel, por lo que es necesario realizar el cambio en su formato y transformarlos en archivos .arff como se muestra en las siguientes imágenes.

Una vez tenemos la información en el archivo de Excel se guarda con la extensión CSV (delimitado con comas) lo anterior para poder abrir el archivo con un editor de texto.

La ***Ilustración 11*** muestra el reporte de MegaTrack en formato de Excel.



	Nombre del servidor	Marca	Modelo	Color	Tipo	Multifuncional	Panel de Control	Lector	Opcion Seleccionar Documento	Opcion Eliminar Documento	Tiempo de Respuesta Identificacion	Tiempo de Respuesta Impresión	Controlador	Copi
1														
2	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
3	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
4	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
5	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
6	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
7	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
8	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
9	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
10	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
11	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
12	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
13	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
14	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
15	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
16	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
17	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
18	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	
19	BOGSPS110P15003	'HP'	'HP Color LaserJet 4700'	Gris	Color	No	No	Si	No	No	8	7	'Unspecified'	

Ilustración 11: Reportes MegaTrack

Después de tener el archivo con un editor de texto y separados los valores con coma, se procede a guardar los archivo con la extensión .arff. Estos archivos pueden ser modificados pero únicamente con un editor de texto como notepad o notepad ++.

La ***Ilustración 12*** muestra cómo quedan los archivos listos para ser leídos por la herramienta Weka.

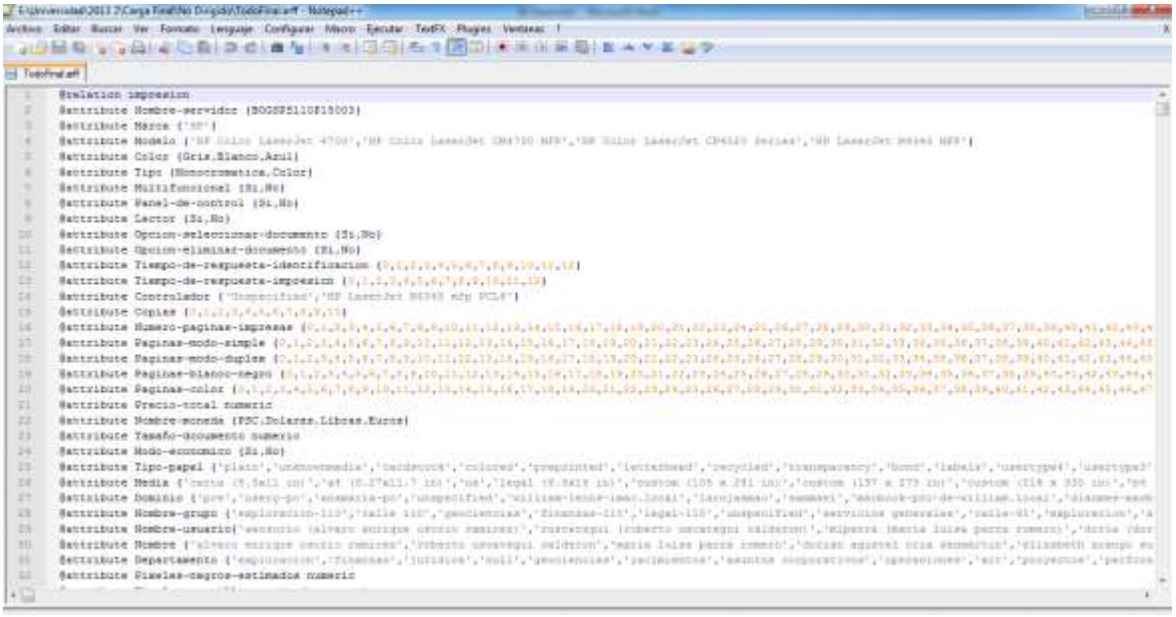


Ilustración 12: Archivo .arff

Se presentaron unos problemas cuando se estaban cargando los datos a la herramienta Weka debido al manejo de mayúsculas y minúsculas. Para dar solución a este error se aplicó una formula en Excel que convertía todos los valores en minúsculas, de ese modo no tendríamos problema de incompatibilidad. La fórmula que se utilizó para este procediendo fue =MINUSC ( ).

Fase 4 Modelado

Esta fase de modelado es una de las más importantes y fundamentales en la investigación ya que es en este punto donde se selecciona una o varias técnicas de modelamiento para el estudio de minería de datos. Existen varias técnicas para la solución un mismo problema, la idea es poder seleccionar la que sea más adecuada para dar resolver el problema planteado. [29]

La ***Ilustración 13*** refleja las actividades que componen la fase de modelado.

Ilustración 13: Fase 4 modelado

Seleccionar la técnica de modelado

Para poder seleccionar la técnica más apropiada para el problema que se está desarrollando es fundamental definir el objetivo del mismo, es decir tener claro que es lo que queremos. A partir de este cuestionamiento se realizar el siguiente análisis.

¿Qué queremos?

1. Determinar si es posible reducir el consumo de impresión en una compañía mejorando la utilización del servicio.
2. Identificar como los empleados de una compañía utiliza el servicio de impresión.
3. Determinar que es una buena práctica de impresión.
4. Determinar que es una mala práctica de impresión.
5. Definir que usuarios deben tener acceso al servicio de impresión.
6. Determinar en qué casos se debe realizar una impresión en Color y en qué caso se debe realizar en Blanco y Negro.
7. Determinar si debe existir un límite de páginas impresas al mes por centro de costos.
8. Definir si todos los centros de costos deben tener límite de páginas impresas por mes.
9. Determinar si debe existir un límite de páginas impresas al mes por usuario.
10. Definir si todos los usuarios deben tener un límite de páginas por mes.

Variables

1. **La impresión debe ser:**
    - Color - Blanco / negro

2. **Debe tener acceso al servicio de impresión:**
    - Si - No

Para poder continuar con el planteamiento de la solución se estableció que era necesario categorizar y definir lo que es una buena o mala impresión. Esta información fue brindada por el cliente.

1. Cantidad de páginas color
2. Cantidad de páginas blanco/negro
3. Modo simple
4. Modo dúplex
5. Precio total
6. Tamaño documento
7. Modo económico
8. Modo full color
9. Tipo papel
10. Media: tamaño del papel
11. Pixeles negros
12. Pixeles amarillos
13. Pixeles magenta

Para el estudio que vamos a realizar no se empleará una única técnica de minería de datos, para esto se plantearan dos escenarios diferentes.

Escenario 1

Identificar patrones de comportamiento desconocidos. Para esto utilizaremos todos los datos mencionados anteriormente y aplicaremos técnicas no dirigidas o no supervisadas.

La ***Ilustración 14*** contiene el escenario de técnicas no dirigidas.

Ilustración 14: Escenario Técnicas No Dirigidas

Escenario 2

Para este escenario surgen dos cuestionamientos fundamentales, cuando una impresión es buena y cuando es mala. La segunda es a que perfiles de usuarios se le debe asignar el servicio de impresión. La ***Ilustración 15*** contiene el escenario de técnicas dirigidas.

Ilustración 15: Escenario Técnicas Dirigidas

En conclusión se utilizará la técnica de árboles de decisión por el lado de las técnicas dirigidas y la detección automática de clúster por el lado de técnicas no dirigidas, cada escenario con su respectivo archivo de datos y la misma herramienta para el análisis.

Las razones por las cuales se seleccionó la técnica de árboles de decisión para analizar el escenario donde se plantearon los cuestionamientos, es decir donde se utilizara técnicas de minería de datos dirigidas son las siguientes:

- Ya que se plantearon dos problemas específicos estos podrán ser analizados independientemente y a medida, es decir orientando y alimentando el árbol para encontrar una solución esperada.

- Por medio de los árboles de decisión es posible cuantificar el costo y a probabilidad que suceda un evento.
- Por su estructura jerárquica en forma de árbol, permite estimar cuales son las opciones para investigar y cuál podría ser su resultado.
- Los arboles de decisión ayudan a tomar mejores decisiones sobre los datos analizados.
- Los problemas planteados no tienen más de dos soluciones, por lo que un árbol de decisión es indicado para validar y analizar estas soluciones con eficiencia y mayor certeza.

Para el escenario donde el objetivo es identificar relaciones entre las variables y de ese modo encontrar patrones de comportamiento, se seleccionó la técnica de detección automática de clúster. La razón es que esta técnica permite determinar grupos diferenciados del resto de los datos, por medio de esta técnica se puede realizar un procedimiento de aprendizaje cuando no se parte una premisa específica, lo anterior quiere decir que esta técnica es ideal para el escenario donde no es evidente ningún tipo de relación entre las variables de los datos analizados, y se busca que por medio de un conjunto de reglas llegar a una solución eficaz y adecuada al problema planteado.

La herramienta que se empleará para el estudio fue Weka 3.6 [43], debido a que es de software libre y a la sencillez en la instalación, su configuración y manejo de la misma.

Construcción del modelo de pruebas

Para la construcción del modelo de prueba se generó un plan para probar la calidad y validez del modelo que se construirá. Se realizó un ejercicio tomando los valores y atributos que definió el cliente para determinar si una impresión era buena o mala.

La ***Ilustración 16*** muestra el análisis de impresiones buenas, regulares y malas.

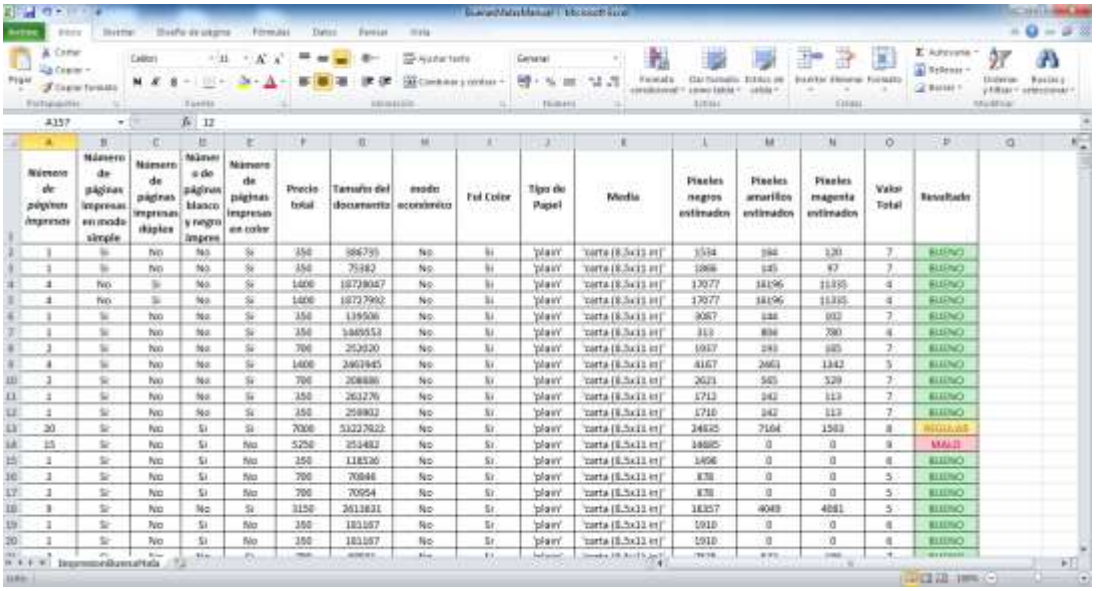


Ilustración 16: Análisis impresión es Buenas - Malas

Siguiendo las indicaciones y parámetros que definió el cliente se realizó un ejercicio manual, para validar en que rango de categoría podría estar la impresión, buena, regular o mal. Cada atributo se asigna un valor dependiendo del tamaño del mismo, al final se realiza una sumatoria para definir su categoría. Si el resultado de la sumatoria es igual a 7 o menor, la impresión se considera buena, si el resultado es igual a 8 se considera una impresión regular, finamente si la impresión es mayor a 7 se define como una impresión mala. A continuación se muestra las formulas utilizada en Excel para realizar el cálculo descrito.

```
=SUMA(SI(A3>10,1,0),SI(A3>25,1,0),SI(A3>35,2,0),SI(B3="Si",1,0),
SI(E3="Si",1,0),SI(F3>3500,2,0),SI(G3>10000,1,0),
SI(I3="Si",1,0),SI(J3<>"'plain"',1,0),SI(K3<>"'carta (8.5x11
in)"',1,0),SI(L3>1000,1,0),SI(M3<700,1,0),SI(N3<700,1,0))
```

```
=SI(O2>7,SI(O2>8,"MALO","REGULAR"),"BUENO")
```

Implementación del modelo

Para la mostrar la forma en que se realizó la implementación del modelo definido utilizando la herramienta Weka diríjase al ***Anexo Manual de usuario***. En este documento encontrará cada una de las funcionalidades que brinda la herramienta, su descripción y la forma en que fue aplicada al problema.

Evaluación del modelo

En esta sección se interpreta los modelos de acuerdo al conocimiento del dominio de los criterios de éxitos preestablecidos. La ***Ilustración 17*** muestra todas las funcionalidades de la opción Explorer en la herramienta Weka.

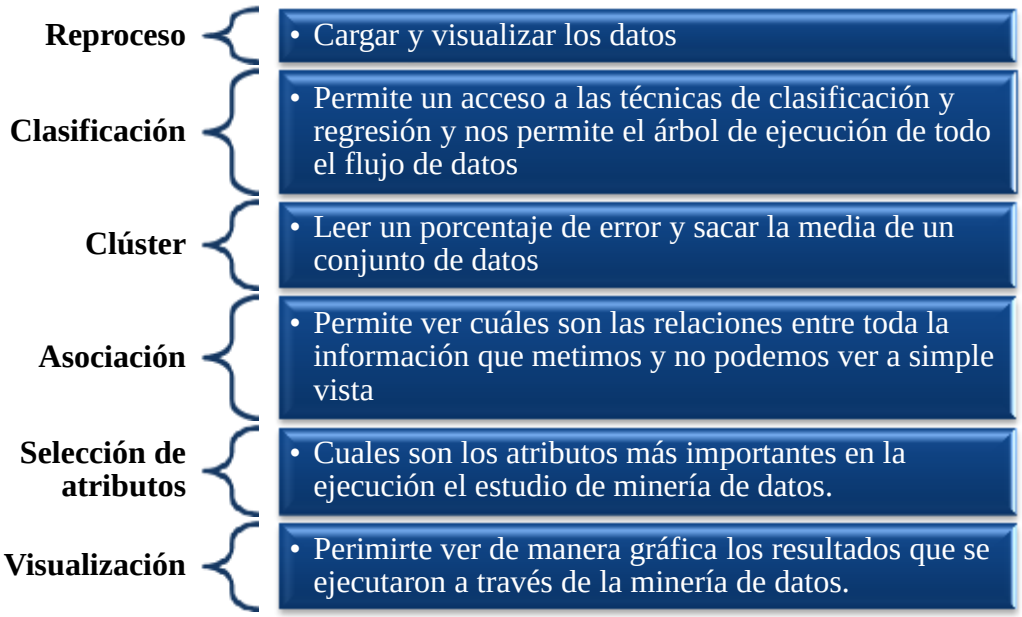


Ilustración 17: Explorador Weka

Filtro: Algoritmo Discretize

- Se realizó sobre el atributo y no sobre la instancia.
- Se utilizó un algoritmo que ya fue supervisado y aprobado por la herramienta Weka
- Es un algoritmo ágil y confiable.

Clúster

Seleccionamos el algoritmo SimpleKMeans para la funcionalidad de clúster de la herramienta Weka. El algoritmo SimpleKMeans consiste en el análisis de datos por grupos. Este algoritmo funciona dividendo los datos recolectados en conjuntos o bloques llamados Clúster y esta separación se realiza por agrupaciones con características similares.

Este algoritmo permite separar los datos en subconjuntos de datos para realizar análisis de manera independiente. Las ventajas que se analizaron para la utilización de este algoritmo fueron las siguientes:

- Es un algoritmo eficiente
- Tiene un nivel de precisión es alto
- No están en proceso de revisión
- Es un algoritmo supervisado y aceptado por la herramienta utilizada para el estudio de minería de datos.

La siguiente Tabla contiene los resultados de los datos predominantes en cada uno de los registros. La ***Tabla 5*** contiene los resultados del Clúster.

Atributo / Clúster	***Full Data: 611334***	***Clúster 0: 388488***	***Clúster 1: 222886***
Modelo	HP Color CP4520	HP Color CP4520	HP M4345
Color	Gris	Azul	Gris
Tipo	Color	Color	Monocromática
Multifuncional	No	No	Si
Panel de Control	No	No	Si

Lector	Si	Si	Si
Opción Seleccionar Documento	No	No	Si
Opción Eliminar Documento	No	No	Si
Tiempo de Respuesta Impresión	5	5	4
Número de páginas impresas	1	1	1
Número de páginas impresas en modo simple	1	1	1
Número de páginas impresas dúplex	0	0	0
Número de páginas blanco y negro impresas	0	0	1
Número de páginas impresas en color	0	1	0
Precio Total	(25-27775)	(25-27775)	(25-27775)
Tamaño del documento	(125484.5-inf)	(125484.5-inf)	(125484.5-inf)

modo económico	No	No	No
Tipo de Papel	plain	plain	Plain
Media	carta (8.5x11 in)	carta (8.5x11 in)	carta (8.5x11 in)
Dominio	pre	pre	Pre
nombre de grupo	Finanzas -110	Finanzas -110	Unspecified
Usuario	unspecified ()	unspecified ()	unspecified ()
Nombre	null	null	Null
departamento	Finanzas	Finanzas	Finanzas
Píxeles negros estimados	(1951.5-inf)	(1951.5-inf)	(1951.5-inf)
Píxeles amarillos estimados	(-inf-0.5)	(228.5-9942)	(-inf-0.5)
Píxeles magenta estimados	(-inf-17.5)	(171.5-1481222)	(-inf-17.5)
Fecha de impresión	(1353862800000-1359306000000)	(1353862800000-1359306000000)	(1353862800000-1359306000000)
Día	11	15	11
Mes	9	10	9

Ano	2013	2013	2013
Hora	11	11	9
Minuto	39	37	47
Segundo	0	0	0
Mañana-Tarde	AM	AM	AM
Nombre Mes	Septiembre	Octubre	Septiembre
Día Semana	2	2	2
Nombre día Semana	Martes	Martes	Martes
Semana de ano	38	40	38
Semestre	2	2	2
Trimestre	3	3	3
Bimestre	5	5	5
Festivo	No	No	No
Ultimo día del mes	Si	Si	Si
Última semana del mes	No	No	No

Tabla 5: Resultados Clúster

Conclusión

Podemos ver que los datos no son constantes, los registros están distribuidos de manera no uniforme por lo que los resultados en diferentes clúster pueden ser totalmente diferentes.

Asociación

Nos permite ver la información oculta detrás de los registros ingresados.

- Seleccionamos el algoritmo de asociación: **Algoritmo Apriori**.

El algoritmo Apriori es utilizado para encontrar reglas de asociación entre variables dentro de un conjunto de datos. Como su nombre lo indica "A priori" quiere decir previo, y en el contexto del problema consiste en el conocimiento previo de los conjuntos frecuentes de datos. La razón principal para la utilización de este algoritmo en el estudio de minería de datos y la detección de asociaciones es para reducir el espacio de búsqueda y de esa manera aumentar la eficiencia en dicha búsqueda.

Otra ventaja significativa que posee el algoritmo Apriori es que controla los datos de transacciones. Un ejemplo para el caso específico, es que en un conjunto de registros de impresión, el algoritmo podría controlar que hora del día suelen realizar las impresiones los empleados de la compañía.

- Soporte mínimo de todas las instancias: 0.95 (580767 instancias)
- Métricas mínimas (coincidencias): 0.9
- Numero de ciclos: 1

Mejores reglas encontradas

En este campo se encuentra las relaciones o recomendaciones que encontró la herramienta según los datos ingresados:

La ***Ilustración 18*** contiene el resultado de la primera asociación.

```
 1. Media=carta (8.5x11 in) tiempo-es-festivo=No 582948 ==> Tipo-papel=plain 580771    conf:(1)
 2. Modo-economico=No Media=carta (8.5x11 in) 605976 ==> Tipo-papel=plain 603694    conf:(1)
 3. Media=carta (8.5x11 in) 608330 ==> Tipo-papel=plain 606031    conf:(1)
 4. Modo-economico=No Media=carta (8.5x11 in) tiempo-es-ultimo-dia-mes=Si 603517 ==> Tipo-papel=plain 601236
 5. Media=carta (8.5x11 in) tiempo-es-ultimo-dia-mes=Si 605871 ==> Tipo-papel=plain 603573    conf:(1)
 6. Tipo-papel=plain 608891 ==> Modo-economico=No 606551    conf:(1)
 7. Tipo-papel=plain Media=carta (8.5x11 in) 606031 ==> Modo-economico=No 603694    conf:(1)
 8. Tipo-papel=plain tiempo-es-ultimo-dia-mes=Si 606426 ==> Modo-economico=No 604086    conf:(1)
 9. Media=carta (8.5x11 in) 608330 ==> Modo-economico=No 605976    conf:(1)
10. tiempo-es-ultimo-dia-mes=Si 608868 ==> Modo-economico=No 606511    conf:(1)
```

Ilustración 18: Primer Resultado Asociación

La ***Ilustración 19*** contiene el resultado de la segunda asociación.

```
 1. tiempo-manana-tarde=AM 579447 ==> tiempo-es-ultimo-dia-mes=Si 577069    conf:(1)
 2. tiempo-es-festivo=No 585900 ==> tiempo-es-ultimo-dia-mes=Si 583448    conf:(1)
 3. Paginas-modo-duplex=0 562842 ==> tiempo-es-ultimo-dia-mes=Si 560461    conf:(1)
 4. tiempo-manana-tarde=AM tiempo-es-festivo=No 555210 ==> tiempo-es-ultimo-dia-mes=Si 552840    conf:(1)
 5. tiempo-es-ultimo-dia-mes=Si 608868 ==> tiempo-es-festivo=No 583448    conf:(0.96)
 6. tiempo-manana-tarde=AM 579447 ==> tiempo-es-festivo=No 555210    conf:(0.96)
 7. tiempo-manana-tarde=AM tiempo-es-ultimo-dia-mes=Si 577069 ==> tiempo-es-festivo=No 552840    conf:(0.96)
 8. tiempo-manana-tarde=AM 579447 ==> tiempo-es-festivo=No tiempo-es-ultimo-dia-mes=Si 552840    conf:(0.95)
 9. tiempo-es-ultimo-dia-mes=Si 608868 ==> tiempo-manana-tarde=AM 577069    conf:(0.95)
10. tiempo-es-festivo=No 585900 ==> tiempo-manana-tarde=AM 555210    conf:(0.95)
```

Ilustración 19: Segundo Resultado Asociación

La ***Ilustración 20*** contiene el resultado de la tercera asociación.

```
 1. tiempo-ano=2013 525796 ==> tiempo-es-ultimo-dia-mes=Si 525796    conf:(1)
 2. tiempo-manana-tarde=AM 579447 ==> tiempo-es-ultimo-dia-mes=Si 577069    conf:(1)
 3. Paginas-modo-duplex=0 562842 ==> tiempo-es-ultimo-dia-mes=Si 560461    conf:(1)
 4. Paginas-modo-duplex=0 tiempo-manana-tarde=AM 533317 ==> tiempo-es-ultimo-dia-mes=Si 531019    conf:(1)
 5. Lector=Si 530585 ==> tiempo-es-ultimo-dia-mes=Si 528209    conf:(1)
 6. tiempo-es-ultimo-dia-mes=Si 608868 ==> tiempo-manana-tarde=AM 577069    conf:(0.95)
 7. Paginas-modo-duplex=0 562842 ==> tiempo-manana-tarde=AM 533317    conf:(0.95)
 8. Paginas-modo-duplex=0 tiempo-es-ultimo-dia-mes=Si 560461 ==> tiempo-manana-tarde=AM 531019    conf:(0.95)
 9. Paginas-modo-duplex=0 562842 ==> tiempo-manana-tarde=AM tiempo-es-ultimo-dia-mes=Si 531019    conf:(0.94)
10. tiempo-es-ultimo-dia-mes=Si 608868 ==> Paginas-modo-duplex=0 560461    conf:(0.92)
```

Ilustración 20: Tercer Resultado Asociación

Análisis de resultado

Según los resultados anteriores podemos evidenciar que en las dos primera iteraciones los atributos que están en todas las reglas con el de tipo de papel, tamaño de papel y si el día es festivo o no. La razón de estos resultados es que el porcentaje de repetición de valores es muy alto, más del 90% lo que ocasionó que no se encuentre ninguna relación relevante para la investigación.

En la tercera iteración se realizó el ejercicio de remover esos atributos que no era relevantes y que estaban ocasionando ruido para el estudio, el resultado de esa iteración evidenció un aumento en la impresión a final de mes y con mayor porcentaje de actividad en horas de la mañana. Siendo consecuentes con el análisis en la opción de clúster, nos dice que el área o centro de costos que más utiliza el servicio de impresión a finales del mes es el departamento de Finanzas.

Selección de atributos

Nos permite identificar cuáles son los atributos más relevantes para el desarrollo del estudio de minería de datos:

- El método que se seleccionó fue el mejor primero. **(BestFirst)**

Este algoritmo fue utilizado dentro de la funcionalidad selección de atributo, el objetivo es determinar que atributos son los más relevantes para la investigación y de ese modo identificar la variable más influyente. Entre las opciones de algoritmo que se contemplaron para identificación de atributos fueron: Búsqueda exhaustiva, Búsqueda genérica y el algoritmo BestFirst.

El algoritmo BestFirst tiene tres formas de buscar:

- Comienza con el conjunto vacío de atributos y busca hacia adelante
- Comienza con el conjunto completo de atributos y busca hacia atrás
- Comienza desde cualquier punto y realiza una búsqueda en ambas direcciones

El algoritmo explora en el grafo y expande el nodo más prometedor seleccionado por una regla específica, intentando predecir un posible camino a la solución final. Este algoritmo es útil para la selección eficiente del mejor candidato para una solución, implementando una cola de prioridad.

Después de realizar el primer análisis la herramienta nos recomienda dos variables:

- Tiempo fecha
- Tiempo es festivo

Entonces seleccionamos las variables que nos recomendó para realizar un nuevo análisis. En primer lugar utilizamos la variable tiempo fecha. Este atributo nos muestra los datos más relevantes y cuál es el más importante entre ellos.

Resultado:

- Dominio
- Tiempo mes
- Tiempo nombre mes
- Tiempo semana ano
- **Tiempo es última semana mes – Es el más importante para la herramienta.**

Con la segunda recomendación, utilizamos la variable tiempo es festivo.

Resultado:

- Tiempo día mes
- Tiempo día semana
- Tiempo día semana nombre
- Tiempo semana ano
- **Tiempo es última semana mes – Es el más importante para la herramienta.**

Visualización:

Muestra gráficamente la distribución de todos los atributos, mostrando graficas en dos dimensiones en las que representa en los ejes todos los posibles pares de combinación de los atributos. Nos permite ver correlaciones y asociaciones entre los atributos en una forma gráfica.

Fase 5 Evaluación

En esta sección se realiza la evaluación del modelo construido para posteriormente realizar el análisis de los datos de la mejor manera. Es importante hacer esta evaluación antes de pasar a las conclusiones de los resultados obtenidos a partir del presente modelo. Al finalizar esta etapa se debe tener la certeza que los objetivos de negocio fueron alcanzados. [29] La ***Ilustración 21*** refleja las actividades que componen la fase de Evaluación.

Ilustración 21: Fase 5 Evaluación

Evaluación de los resultados

Según la evaluación realizada en la sección anterior se puede evidenciar que se encontraron patrones de comportamiento que son evidentes para la organización, es normal que el volumen de impresión aumento a finalizar el mes laboral debido a las actividades de cierre y con mayor razón en el área financiera. El resultado esperado era encontrar patrones desconocidos, debido al problema de Habeas Data que se presentó en el desarrollo del estudio de minería de datos no fue posible cumplir este objetivo, eso no significa que no se cumpliera con el objetivo del trabajo de grado, ya que como se explicó en el planteamiento, el objetivo consistía en realizar un estudio para identificar si existía o no un patrón de comportamiento, para un estudio de minería de datos es posible el no hallar un resultado especifico. La

principal recomendación para el caso planteado es continuar con la investigación, cambiando la empresa de casos de estudio. Una empresa que permita el manejo de la información correspondiente a los usuarios, aumentaría la posibilidad de encontrar resultados significativos.

Revisión del proceso

Esta sección consiste en calificar el proceso que se realizó en el estudio de minera de datos, el objetivo es identificar que elementos podrían mejorarse.

En general el proceso ejecutado durante el presente estudio fue satisfactorio, se presentaron inconvenientes en la recolección de datos por el caso mencionado en las secciones anteriores (ley de protección de datos), esto generó un retraso y un replanteamiento del estudio de minera de datos.

Para la elección del modelo y las técnicas utilizadas, fue necesario un estudio riguroso y la realización de varios ejercicios de práctica. La elección de la herramienta por cual se realizó el estudio también se considera como una buena decisión ya que la utilización de la misma fue muy sencilla y práctica.

Determinar los próximos pasos

Para los objetivos establecidos en un principio los resultados no fueron exitosos o satisfactorios ya que no se encontraron asociaciones que indicaran algún tipo de patrón de conducta entre los datos analizados. La razón de no encontrar la información esperada se considera que no radica en las técnicas seleccionadas si no en los datos recolectados, por lo que no se recomienda la realización de una nueva iteración.

Como se mencionó anteriormente, para próximos estudios similares a este es importante contar con toda la información relevante para la problemática, por lo que se sugiere implementar este modelo en una empresa donde se permitido la captura y manipulación de datos. Existes ejercicios como el cambio de identidad de usuarios para resolver los inconvenientes y restricciones establecidas por la ley de protección de datos, sin tener que poner en riesgo la seguridad de la información de los empleados de la organización estudiada.

Fase 6 Transferencia

Esta etapa final consiste en la organización y presentación de los resultados obtenidos a partir del estudio de minería de datos realizado de forma que el cliente pueda utilizarlos de la mejor manera posible. Es importante interpretar y analizar los datos que arroja la investigación para determinar si la hipótesis planteada en un inicio se cumple o si por lo contrario después del estudio no se llega a una conclusión exitosa. [29] La ***Ilustración 22*** refleja las actividades que componen la fase de Transferencia.

Ilustración 22: Fase 6 Transferencia

Plan de transferencia

Debido al cambio en el planteamiento de la investigación, en esta sección de plan de transferencia se expone como quedaron resueltos o pendientes los cuestionamientos realizados para la elaboración del modelo.

Después de finalizar el estudio se puede afirmar que es posible reducir el consumo de impresión mejorando la utilización del servicio, para dar un ejemplo relacionado con los hallazgos encontrados tenemos lo siguiente: En la investigación se encontró que existe un incremento del volumen de impresión a finales del mes y se supone que es debido a las actividades de cierre mensuales, si tenemos este conocimiento podemos atacar al problema directamente, estableciendo campañas y capacitaciones a usuarios de como imprimir en estos días del mes de tal forma que su trabajo no se vea afectado. Lo anterior quiere decir que si conocemos la causa del problema, es más sencillo solucionarlo.

Para los cuestionamientos sobre si debe asignarse o no el servicio de impresión a los empleados, y si debe existir una restricción o limitante en el volumen de impresión, estas

decisiones son únicamente tomadas por la organización, ya que son las que proveen el servicio y lo administran a su parecer. Lo que es posible es realizar un acompañamiento y una asesoría con argumentos para que se tomen las mejores decisiones.

En conclusión para dar respuesta a la pregunta **¿Qué queremos?** Podemos definir que aunque el resultado no fue del todo el esperado, se definió un claro punto de partida para dar solución a la problemática planteada. Quedan cuestionamientos pendientes para próximas investigaciones tal como **¿En realidad es posible identificar un patrón de comportamiento desconocido o relevante utilizando todos los datos necesarios?**

Producción del reporte final

La producción del reporte final se realizara por medio de una presentación frente a los jurados establecidos por la Pontificia Universidad Javeriana y el director de trabajo de grado. Para ver la presentación diríjase al anexo ***Presentación Trabajo de Grado***. Los resultados entregados en dicha exposición estarán resalados por el documento de memoria de trabajo de grado y los correspondientes anexos. Todos estos documentos podrán ser consultados en el siguiente enlace:

http://pegasus.javeriana.edu.co/~CIS1310IS02/

Revisión del Proyecto

La revisión del proyecto se realizó mediante las correcciones sugeridas por los jurados y el director de trabajo de grado en el momento de la sustentación.

IV - RESULTADOS Y REFLEXIÓN SOBRE LOS HALLAZGOS

Después de realizar el análisis de los resultados con los datos obtenidos en el modelo construido se puede evidenciar que no se identificaron patrones ocultos de conducta relacionados a prácticas de impresión. Se realizaron varias iteraciones en las dos técnicas definidas en la fase de modelado realizando varios ejercicios donde se removían y adicionaban atributos del modelo para intentar encontrar alguna asociación entre ellos que pudiera definir alguna conducta significativa para la investigación. Lo que se puede determinar después de la revisión de los datos es que en la gran mayoría de estos, sus valores se repetían constantemente lo que no permitía un correcto desarrollo para el estudio.

La limitación en la recolección y manipulación de los datos relacionados con la información personal de los usuarios obstruyó el desarrollo del estudio de minería de datos ya que la mayor cantidad de información que podría ser relevante para el estudio provenía de esos datos. Debido a este inconveniente fue necesario acotar los datos de estudio a los suministrados por el software de control de impresión llamado MegaTrack.

Según la información recolectada se puede deducir que existe un incremento en las prácticas de impresión en los últimos días del mes, y con un gran porcentaje de actividad en las horas de la mañana, los picos más altos de impresión son por parte del departamento de finanzas según la información extraída en las diferentes iteraciones del clúster. Es evidente que este incremento se debe a los cierre del mes y corresponde a todas las actividades empleadas para esta labor.

El 97% de los registros de impresión se realizan en modo full color, en los estándares de papel establecidas por la compañía; lo que podemos deducir de esta información es que no existe un cultura por el ahorro y reciclaje. Probablemente los usuarios no son conscientes del gasto incensario cuando se realizan impresión de documentos sin utilizar las opciones acordes a la necesidad de cada trabajo, la otra hipótesis que surge a partir de este análisis es que los usuarios desconocen la forma de manipular los dispositivos de impresión en el momento de seleccionar el tipo económico o full color, se podría implementar una campaña de capacitación a los usuarios para corregir y mejorar las prácticas de impresión en la compañía.

El concepto final por parte del autor de este trabajo de grado es que los resultados extraídos y posteriormente analizados son evidentes y predecibles, es muy común que en las organizaciones existan picos de impresión al final del mes por asuntos relacionados a cierre de mes, el objetivo de este estudio consistía en identificar situaciones desconocidas que no fuera posible percibir a simple vista, lastimosamente para la investigación el inconveniente con los datos no permitió cumplir el objetivo, seguramente mejorar la calidad en la recolección de datos mejoraría el hallazgo de los resultados esperados.

V – CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

1. Conclusiones

El siguiente listado corresponde a las conclusiones obtenidas una vez finalizado el estudio de minería de datos por parte del estudiante:

- El objetivo del presente trabajo de grado consistió en realizar un estudio de minería de datos que permitirá identificar si existe o no un patrón de comportamiento relacionado a prácticas de impresión. Después de realizar el estudio se puede concluir que si existe un patrón de comportamiento, aunque por el inconveniente en la recolección de los datos este patrón es demasiado evidente como se reportó en la fase de evaluación del proyecto. Es preciso mencionar que el hallazgo podría haber tenido una mayor significancia en caso de haber contado con toda la información.
- Existe una gran oportunidad de negocio en el campo planteado por la problemática, cada vez crece más el interés de las organizaciones en implementar estrategias que mejoren el consumo de recursos y disminuyan los gastos operativos. También es muy importante el hecho que las empresas buscan contribuir con la conservación del medio ambiente por medio de estas estrategias.

- Se cumplió con el objetivo de aportar a la solución de la problemática, aunque no se cumplió en su totalidad, se creó un punto de partida para futuros trabajos de investigación relacionados con el tema.
- La herramienta de minería de datos (Weka) fue una excelente elección; por una parte es una herramienta de software libre lo que facilita su adquisición, y por otra parte la instalación, configuración y administración de la misma facilita al estudiante para realizar el desarrollo de la investigación. Para futuros estudios se debe tener en cuenta la posibilidad de utilizar más de una herramienta para poder realizar comparaciones entre ellas.
- Fue de gran utilidad la realización de reuniones de seguimiento con el cliente y con el director del trabajo de grado durante el desarrollo del proyecto, esta estrategia garantizó la buena interpretación en el análisis de la situación, el planteamiento para su desarrollo y finalmente su resultado.

2. Recomendaciones

A continuación se muestran las recomendaciones, a partir de la experiencia conseguida en el desarrollo del presente trabajo de grado:

- La primera recomendación hace referencia al problema de Habeas Data que se presentó en la recolección de datos para las primeras fases del estudio. Para realizar un estudio de este tipo es necesario contar con todos los datos posibles, por lo que se recomienda utilizar una empresa donde los datos personales no sean tan críticos y sea posible gestionar la adquisición y manipulación de los mismos. Al escoger una multinacional como caso de estudio disminuyo la posibilidad de utilización de los datos necesarios para la investigación.
- Realizar una mejor gestión en las primeras fases de un estudio de minería de datos, lo más importante en un trabajo como este, es la información que se puede recolectar. No existirá el caso en que demasiada información es contraproducente.
- Definir claramente las metas, alcances y objetivos del trabajo de grado, así como las limitaciones y fronteras del mismo.

3. Trabajos Futuros

A continuación se presentan algunas ideas para trabajos futuros que el estudiante recomienda basándose en la experiencia después de realizar el presente proyecto:

- Inicialmente la recomendación es continuar con este estudio de minería de datos pero solucionado el inconveniente de la limitación de los datos correspondientes a los usuarios, el problema de Habeas Data se puede solventar utilizando nombres ficticios y no publicando la información personal de cada empleado. Seguramente se podrán identificar muchas situaciones y casos que no podemos imaginarnos.
- Existe una gran oportunidad para realizar una propuesta de negocio, creando un módulo adicional al software de control de impresión como MegaTrack con una aplicación de minería de datos.
- Plantear una propuesta de trabajo de grado que no solo incluya el estudio y la identificación de conductas, si no que vaya más allá, que pueda llegar hasta el planteamiento de soluciones por medio de un estudio como el realizado.

VI - REFERENCIAS Y BIBLIOGRAFÍA

Esta sección presenta las referencias y la bibliografía utilizada para el desarrollo del trabajo de grado.

1. Referencias

[1] Enrique Valenzuela Toro, Marcelo Lemunao Gutiérrez. Una Herramienta Estratégica para el Mundo de los Negocios. 29 de Junio 2010. Páginas [3-6].

[2] WebMining Consultores. KDD: Proceso de Extracción de conocimiento [Web Mining]; 2011 [Citado 2013 Septiembre 14] Disponible en: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

[3] Maneiro, Mariela Yanina. Minería de Datos; 2008 [Citado 2013 Septiembre 14] Disponible en:

http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MineriaDatosYany2008.pdf

[4] Neftalí de Jesús Calderón Méndez. Minería de datos una herramienta para la toma de decisiones; 2006 [Citado 2013 Septiembre 17] Disponible en: http://biblioteca.usac.edu.gt/tesis/08/08_0307_CS.pdf

[5] Luis M. Molinero. Análisis de series Temporales; 2004 [Citado 2013 Septiembre 28] Disponible en: http://www.seh-lelha.org/tseries.htm

[6] Peter Rob, Carlos Coronel. Sistemas de Bases de datos. 5ta Ed.

[7] Definición Tamaño Empresarial Micro, Pequeña, Mediana o Grande; 2012 [Citado 2013 Octubre 2] Disponible en: http://www.mipymes.gov.co/publicaciones.php?id=2761

[8] Clasificación de empresas en Colombia [Bancoldex]; 2013 [Citado 2013 Octubre 2] Disponible en: http://www.bancoldex.com/contenido/contenido.aspx?catID=112&conID=315

[9] Ley 1581 de 2012 Nivel Nacional [Régimen Legal de Bogotá D.C.]; 2012 [Citado 2013 Octubre 3] Disponible en: http://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=49981

[10] María N. Moreno García∗, Luis A. Miguel Quintales, Francisco J. García Peñalvo y M. José Polo Martín. APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN LA CONSTRUCCIÓN Y VALIDACIÓN DE MODELOS PREDICTIVOS Y ASOCIATIVOS A PARTIR DE ESPECIFICACIONES DE REQUISITOS DE SOFTWARE; [Citado 2013 Octubre 3] Disponible en: http://ceur-ws.org/Vol-84/paper4.pdf

[11] M.V. Guzmán, H. Carrillo, E. Villaseñor, E. Valencia, R. Calero, L. E. Morán y A. Acosta. Minería de Datos con Redes Neuronales Artificiales: Aplicación en Vacunas – Tuberculosis; [Citado 2013 Octubre 3] Disponible en: http://www.dynamics.unam.edu/DinamicaNoLineal/Articulos/MineriaRedesNVacunas.pdf

[12]J. Bigus, McGrawHill .Data Mining with neural networks; 1996.

[13] Mario Castillo Hernández. Toma de Decisiones en las empresas. 1ra Ed; 2008. Páginas [149-151].

[14] Michael J. A. Berry, Gordon S. Linoff. Mastering data mining the art and science of customer relationship management; 2da Ed; 2004.Capituo [5].

[15] Pronostico correlación y regresión; [Citado 2013 Octubre 5] Disponible en: http://metodoscuantitativo2.galeon.com/enlaces2217022.html

[16] Regresión; [Citado 2013 Octubre 5] Disponible en: http://www.bioestadistica.uma.es/libro/node40.htm

[17] Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining Practical Machine Learning Tools and Techniques.3ra Ed; 2011. Páginas [2-38].

[18] Verónica S. Bogado y Mariana C. Arruzabla. Sistemas Operativos; 2003 [Citado 2013 Octubre 5] Disponible en: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/MonografiaMD.PDF

[19] Michael J. A. Berry, Gordon S. Linoff. Mastering data mining the art and science of customer relationship management; 2da Ed; 2004. Capítulo [14].

[20] Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining Practical Machine Learning Tools and Techniques. 3ra Ed; 2011. Páginas [403-406].

[21] Wilfredy Santamaría Ruíz. MODELO DE DETECCION DE FRAUDE BASADO EN EL DESCUBRIMIENTO SIMBOLICO DE REGLAS DE CLASIFICACION EXTRIDAS DE UNA RED NEURONAL; 2010 [Citado 2013 Octubre 6] Disponible en: http://www.bdigital.unal.edu.co/3086/1/299742.2010.pdf

[22] Juan Carlos Gonzales Cardona. SISTEMA DE APOYO PARA LA ACREDITACIÓN DE LA CALIDAD DE PROGRAMAS ACADEMICOS DE LA UNIVERSIDAD DE CALDAS, APLICANDO TÉCNICAS EN MINERÍA DE DATOS; 2011 [Citado 2013Octubre 18] Disponible

en:http://repositorio.autonoma.edu.co/jspui/bitstream/11182/350/1/Msc.GyDlloSoft_Informe Final_JuanCarlosGonzalez.pdf

[23] Datametrics. Minería de datos con R. Su mejor amigo para los Grandes datos; 2013 [Citado 2013 Octubre 18] Disponible en: http://www.idata.com.co/index.php/blog-page/37-mineria-de-datos-con-r-su-mejor-amigo-para-los-grandes-datos

[24] Orange; [Citado 2013 Octubre 18] Disponible en: http://orange.biolab.si/features/

[25] Álvaro Alejandro Alcántara Mori. Formulación de Minería de Datos para la Empresa Distribuidora de Productos Espinoza Aguilar S.A.2012; Páginas [12-15].

[26] Sistemas y herramientas de minería de datos. Ejemplos; [Citado 2013 Octubre 18] Disponible en: http://www.oocities.org/es/mineria.datos/sistemas_herramientas_mineria_datos.pdf

[27] Data Mining [StatSoft Making the World More Productive]; [Citado 2013 Octubre 18] Disponible en: http://www.statsoft.com/Solutions/Cross-Industry/Data-Mining

[28] Ramón García-Martínez. Metodología CRISP-DM; 2000 [Citado 2013 Octubre 20] Disponible en: http://www.iidia.com.ar/rgm/CD-TIpEI/TEI-2-CRISP-DM-GdP-material.pdf

[29] Metodología para proyectos de Minería de Datos; 2008 [Citado 2013 Octubre 20] Disponible en: http://jpgarcia.cl/2008/07/25/metodologia-para-proyectos-de-mineria-de-datos/

[30] Diccionario de la lengua española. Impresión; 2005 [Citado 2013 Octubre 21] Disponible en: http://www.wordreference.com/definicion/impresi%C3%B3n

[31] Machine Learning Group at the University of Waikato. Weka; [Citado 2013 Octubre 21] Disponible en: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

2. Bibliografía

- Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining Practical Machine Learning Tools and Techniques. 3ra Ed; 2011.
- Mario Castillo Hernández. Toma de Decisiones en las empresas. 1ra Ed; 2008.
- Peter Rob, Carlos Coronel. Sistemas de Bases de datos; 5ta Ed.
- Michael J. A. Berry, Gordon S. Linoff. Mastering data mining the art and science of customer relationship management; 2da Ed; 2004.
- J. Bigus, McGrawHill .Data Mining with neural networks; 1996.

VII - ANEXOS

El siguiente listado contiene los anexos correspondientes al presente trabajo de grado.

Anexo1. Glosario

Anexo2. Post-Mortem

1. Metodología propuesta vs. Metodología realmente utiliza-da.
2. Actividades propuestas vs. Actividades realizadas.
3. Efectividad en la estimación de tiempos del proyecto
4. Costo estimado vs. Costo real del proyecto

Actas de Reunión

- Acta de Reunión - Agosto 08
- Acta de Reunión - Agosto 21
- Acta de Reunión - Septiembre 05
- Acta de Reunión - Septiembre 18
- Acta de Reunión - Octubre 04
- Acta de Reunión - Octubre 09
- Acta de Reunión - Octubre 19
- Acta de Reunión - Octubre 21
- Acta de Reunión - Octubre 23
- Acta de Reunión - Octubre 28
- Acta de Reunión Cliente - Octubre 31
- Acta Categorización impresión - Noviembre 05
- Acta de Reunión - Noviembre 06
- Acta de Reunión - Noviembre 08
- Acta de Reunión - Noviembre 13
- Acta de Reunión - Noviembre 15

Reportes MegaTrack

- Registros-MegaTrack
- Registros-MegaTrack-BuenasMalas
- Análisis impresión Buenas-Malas

Arquitectura de la solución

Descripción Diagramas Tabla de Hechos

Manuales

- Manual de Usuario
- Manual de instalación

Archivos herramienta Weka

- Registros-Dirigido
- Registros-NoDirigido

Cronograma – Plan de trabajo Proyecto

Carta cliente – Printer On Line Integral Document SAS

Presentación Trabajo de Grado