UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

CENTRO UNIVERSITARIO NEZAHUALCÓYOTL

INGENIERÍA EN SISTEMAS INTELIGENTES

MANUAL PARA PRÁCTICAS DEL LABORATORIO DE CÓMPUTO

ASIGNATURA:

MINERÍA DE DATOS I

ELABORARON:

DRA. DORICELA GUTIÉRREZ CRUZ
M. en C. YAROSLAF AARÓN ALBARRÁN FERNÁNDEZ
DRA. CARMEN LILIANA RODRÍGUEZ PÁEZ

AGOSTO 2019

MANUAL PARA PRÁCTICAS DEL LABORATORIO DE CÓMPUTO PARA LA ASIGNATURA MINERÍA DE DATOS I

IDENTIFICACIÓN DE LA UNIDAD DE APRENDIZAJE

<table>
<tr><td colspan="9">Espacio académico: CENTRO UNIVERSITARIO NEZAHUALCÓYOTL</td></tr>
<tr><td colspan="4">Programa educativo INGENIERÍA EN SISTEMAS INTELIGENTES</td><td colspan="5">Área de docencia: HERRAMIENTA PARA LOS SISTEMAS INTELIGENTES</td></tr>
<tr><td colspan="3">Aprobación de los HH Consejos Académico y de Gobierno</td><td colspan="2">Fecha: AGOSTO 2019</td><td colspan="4">Programa elaborado por: Doricela Gutiérrez Cruz, Carmen Liliana Rodríguez Páez</td></tr>
<tr><td colspan="5">Nombre de la unidad de aprendizaje: Minería de Datos I</td><td colspan="4">Fecha de elaboración: julio 2019</td></tr>
<tr><td>Clave</td><td>Horas de Teoría</td><td>Horas de Práctica</td><td>Total de horas</td><td>Créditos</td><td>Área curricular:</td><td>Carácter de la unidad de aprendizaje</td><td>Núcleo de formación</td><td>Modalidad</td></tr>
<tr><td>L40642</td><td>1.0</td><td>2.0</td><td>3.0</td><td>4.0</td><td>DESCUBRIMIENTO DE CONOCIMIENTO A PARTIR DE DATOS</td><td>Obligatoria</td><td>INTEGRAL</td><td>ESCOLARIZADA CON ADMINISTRACIÓN FLEXIBLE DE LA ENSEÑANZA</td></tr>
<tr><td colspan="3">Prerrequisitos (Conocimientos previos):

Introducción al Tratamiento de Imágenes
e
Introducción al Reconocimiento de Patrones</td><td colspan="3">Unidad de aprendizaje antecedente:

Introducción al Reconocimiento de Patrones</td><td colspan="3">Unidad de aprendizaje consecuente:

NINGUNA</td></tr>
<tr><td colspan="9">Programas en los que se imparte: LICENCIATURA DE INGENIERÍA EN SISTEMAS INTELIGENTES</td></tr>
</table>

EL PRESENTE MANUAL DE PRÁCTICAS HA SIDO AVALADO EN EL MES DE AGOSTO DE 2019 POR:

ÍNDICE

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

DIRECTORIO INSTITUCIONAL

Dr. en Edu. **Alfredo Barrera Baca**

RECTOR

M. en E.U. y R.
Marco Antonio Luma Pichardo
Secretario de Docencia

M. en C
Jannet Valero Vilchis
Secretaria de Rectoría

Dra. en Ed.
Sandra Chávez Marín
Secretaria de Extensión y Vinculación

M. en Dis.
Juan Miguel Reyes Viurquez
Secretario de Administración

M. en L.A.
María del Pilar Ampudia García
Secretaria de Cooperación Internacional

Dr. en C.S.
Luis Raúl Ortiz Ramírez
Abogado General

Lic. en Com.
Gastón Pedraza Muñoz
Director General de Comunicación Universitaria

M. en D.F.
Jorge Rogelio Zenteno Domínguez
Encargado del Despacho de la Contraloría Universitaria

M. en A.
José Francisco Mejía Carbajal
Secretario Particular Adjunto del Rector

Dr. en C.I.
Carlos Eduardo Barrera Díaz
Secretario de Investigación y Estudios Avanzados

Dr. en A.
José Édgar Miranda Ortiz
Secretario de Difusión Cultural

M. en E.
Javier González Martínez
Secretario de Finanzas

Dr. en C.C.
José Raymundo Marcial Romero
Secretario de Planeación y DESARROLLO Institucional

Dra. en Dis.
Mónica Marina Mondragón
Secretaría de Cultura Física y Deporte

M. en R. I.
Jorge Bernáldez García
Secretario Técnico de la Rectoría

M. en A. P.
Guadalupe Ofelia Santamaría González
Directora General de Centros Universitarios y Unidades Académicas Profesionales

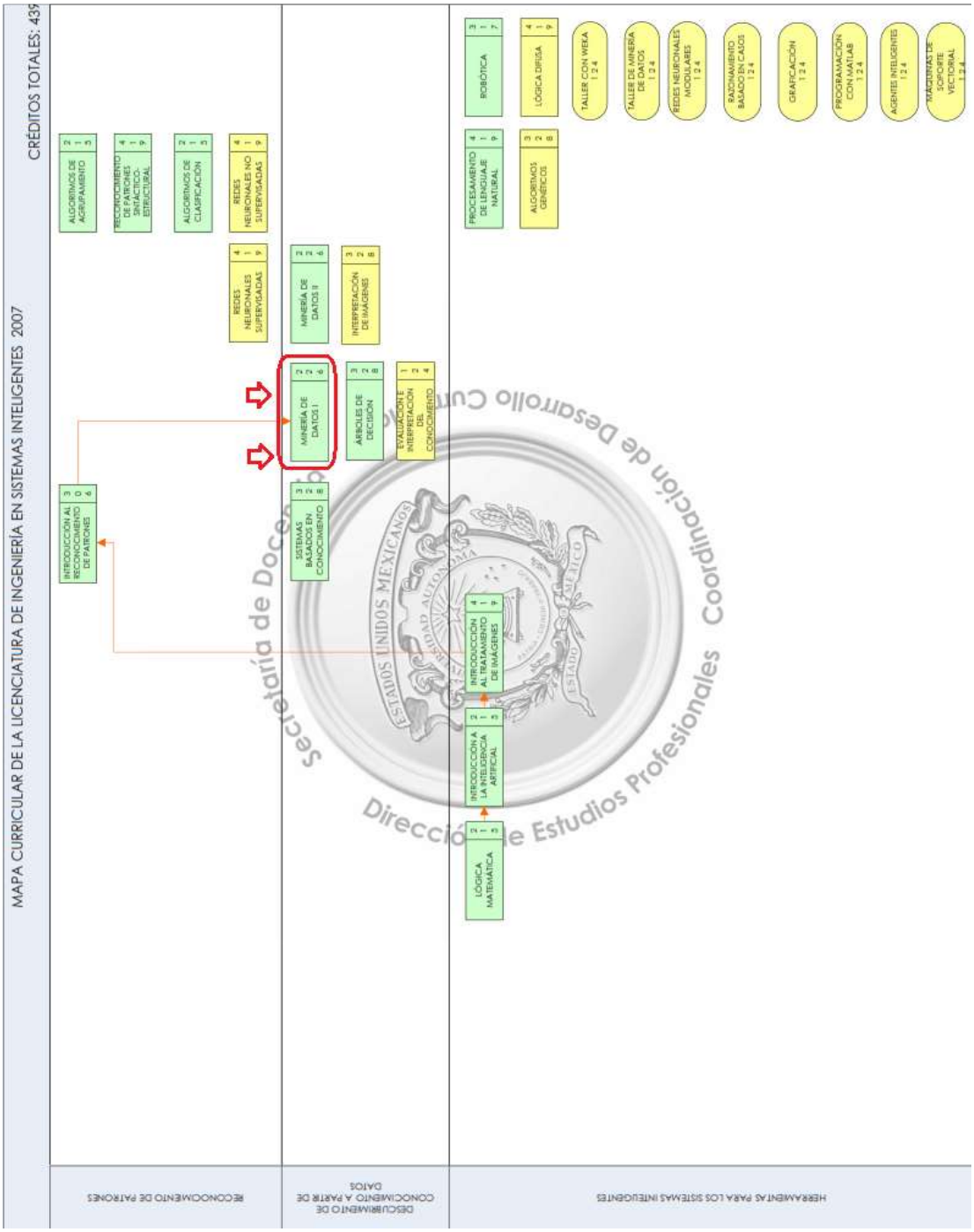
Lic. En Act.
Angelita Garduño Gómez
Secretaria particular del Rector

CENTRO UNIVERSITARIO UAEM NEZAHUALCÓYOTL

DIRECTORIO

Maestro en Derecho **Juan Carlos Medina Huicochea**	**ENCARGADO DEL DESPACHO DE LA DIRECCIÓN**
Maestro en Ciencias **José Antonio Castillo Jiménez**	Subdirector Académico
Licenciado en Economía **Ramón Vital Hernández**	Subdirector Administrativo
Doctora en Ciencias Sociales **María Luisa Quintero Soto**	Coordinadora de Investigación y Estudios Avanzados
Licenciado en Administración de Empresas **Víctor Manuel Durán López**	Coordinador de Planeación y DESARROLLO Institucional
Maestro en Ciencias **Cesar Lucio Gutiérrez Ruiz**	Coordinador de la Licenciatura en Comercio Internacional
Maestro en S.F. **Carlos Anaya Hernández**	Coordinador de la Licenciatura en Educación para la Salud
Doctor en Ingeniería de Sistemas **Ricardo Rico Molina**	Coordinador de la Licenciatura en Ingeniería en Sistemas Inteligentes
Maestro en Ciencias **Ricardo Pacheco Ruiz**	Coordinador de la Licenciatura en Ingeniería en Transporte
Maestro en Ciencias de la Computación **Erick Nicolás Cabrera Álvarez**	Coordinador de la Licenciatura en Seguridad Ciudadana Mixta
Maestro en Administración **José Ramon CS. Garcia Ibarra**	Coordinador de la Licenciatura en Seguridad Ciudadana Presencial

Ubicación de la asignatura de Minería de Datos I, dentro del programa de la Lic. En Ing. en Sistemas Inteligentes.



SECUENCIA DIDÁCTICA

PRÁCTICA 1. INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

PRÁCTICA 2. MODELOS DE PROCESO PARA PROYECTOS DE MINERÍA DE DATOS

PRÁCTICA 3. MODELOS DE PROCESO PARA PROYECTOS DE MINERÍA DE DATOS: METODOLOGÍA CRISP- DM

PRÁCTICA 4. METODOLOGÍA CRISP-DM: FASE DE COMPRENSIÓN DEL NEGOCIO O PROBLEMA

PRÁCTICA 5. METODOLOGÍA CRISP-DM: FASE DE COMPRENSIÓN DE LOS DATOS

PRÁCTICA 6. METODOLOGÍA CRISP-DM: FASE DE PREPARACIÓN DE LOS DATOS

PRÁCTICA 7. METODOLOGÍA CRISP-DM: FASE DE MODELADO

PRÁCTICA 8. METODOLOGÍA CRISP-DM: FASE DE EVALUACIÓN

PRÁCTICA 9. METODOLOGÍA CRISP-DM: FASE DE IMPLEMENTACIÓN

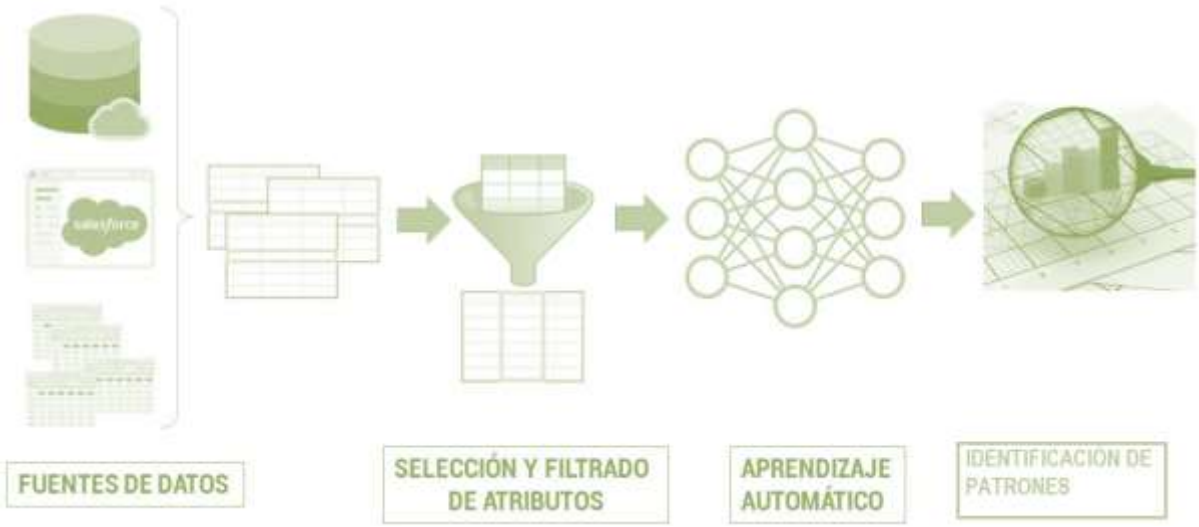
PRÁCTICA 1
INTRODUCCIÓN A LA MINERÍA DE DATOS

OBJETIVO

El alumno conocerá el concepto y la importancia de la minería de datos en distintas áreas del conocimiento.

INTRODUCCIÓN

La minería de datos (MD) intenta buscar el sentido a la explosión de información que actualmente puede ser almacenada en grandes volúmenes. La revolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, almacenar, distribuir, y transmitir [1]. Con este importante progreso que se encuentra en el área de tecnologías relacionadas y en su amplio uso, se continúa recogiendo y almacenando en bases de datos gran cantidad de información. El descubrir conocimiento a partir de este enorme volumen de datos es un reto en sí mismo [2]. El avance de la tecnología para la gestión de bases de datos hace posible integrar diferentes tipos de datos, tales como imagen, video, texto, y otros datos numéricos, en una base de datos sencilla, facilitando el procesamiento multimedia. La tecnología actual y su creciente demanda necesita del desarrollo de tecnologías de minería de datos más avanzadas para interpretar la información y el conocimiento de los datos que se encuentran distribuidos por todo el mundo. El acceso a grandes volúmenes de datos multimedia traerá la mayor transformación para el global de la sociedad. Por tanto, el desarrollo de la tecnología de minería de datos avanzada continuará siendo una importante área de estudio, y en consecuencia se espera gastar muchos recursos en esta área de desarrollo en los próximos años. Existen diversos dominios donde se almacenan grandes volúmenes de información en bases de datos centralizadas y distribuidas.



La MD es un intento de buscarle sentido a la información que actualmente puede ser almacenada en grandes volúmenes. La fase de minar los datos es la representación del tipo de modelo obtenido. Se concentra en la búsqueda, que tendrán unas varias formas de representación en dependencia del tipo de modelo obtenido. El análisis de los datos puede proporcionar en conjunto un verdadero conocimiento que ayude en la toma de decisiones.

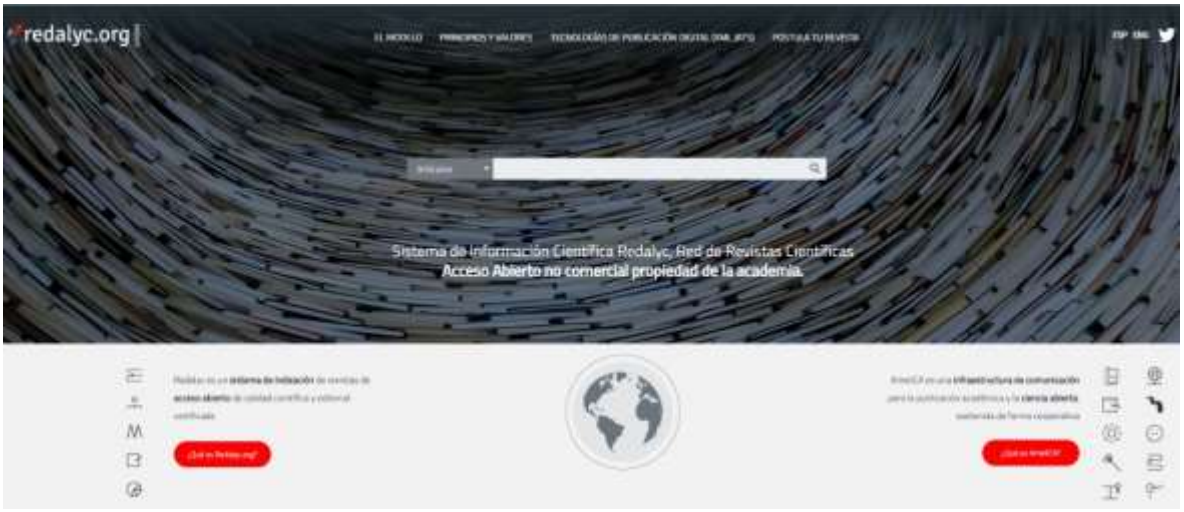
Se resalta que en la metodología de la MD forma parte de un proceso denominado descubrimiento de conocimiento en bases de datos «Knowledge Discovery in Databases» (KDD), que indica los pasos necesarios para reducir riesgos en la búsqueda de modelos de conocimiento al aplicar técnicas de MD. Por ejemplo, los datos requieren un sustancial preprocesamiento para ser modelados (limpieza y preparación de datos) en el proceso KDD. Reconocida como la tarea no trivial de extraer información implícita, previamente desconocida y potencialmente útil de bases de datos (Flrawey et. al. 1992). El proceso de descubrir conocimiento interesante de grandes cantidades de datos almacenadas en bases de datos, data warehouses u otro repositorio de información (Jiawei Han, Micheline Kamber 2001).

La Minería de Datos (MD) es definida como el procesamiento de los datos para encontrar patrones de comportamiento que sean de utilidad para la toma de decisiones, se relaciona de manera estrecha con la estadística, usando técnicas de muestreo y visualización de datos y depuración en donde la materia prima son las bases de datos. La fase de minar los datos es la representación del tipo de modelo obtenido. El análisis de los datos puede proporcionar en conjunto un verdadero conocimiento que ayude en la toma de decisiones.
Puede definirse como el uso consistente de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre-procesados, que sean de utilidad para la toma de decisiones. Se relacionan de manera estrecha con la estadística, usando técnicas de muestreo y visualización de datos. La investigación y el desarrollo para analizar grandes volúmenes de datos se hicieron cada vez más necesarios, así mismo puede realizarse a partir de archivos. No obstante, las ventajas aumentan cuando se cuenta con grandes volúmenes de datos, descubrir conocimiento de este enorme volumen de datos es un reto en sí mismo. La MD es un intento de buscarle sentido a la explosión de información que actualmente puede ser almacenada. La fase de minar los datos es la representación del tipo de modelo obtenido. Se concentra en la búsqueda, que tendrán una o varias formas de representación en dependencia del tipo de modelo obtenido. El análisis de los datos puede proporcionar en conjunto un verdadero conocimiento que ayude en la toma de decisiones.

Cabe resaltar que en la metodología de MD forma parte de un proceso denominado descubrimiento de conocimiento en bases de datos «*Knowledge Discovery in Databases*» (KDD), que indica los pasos necesarios para reducir riesgos en la búsqueda de modelos de conocimiento al aplicar técnicas de MD.

DESARROLLO

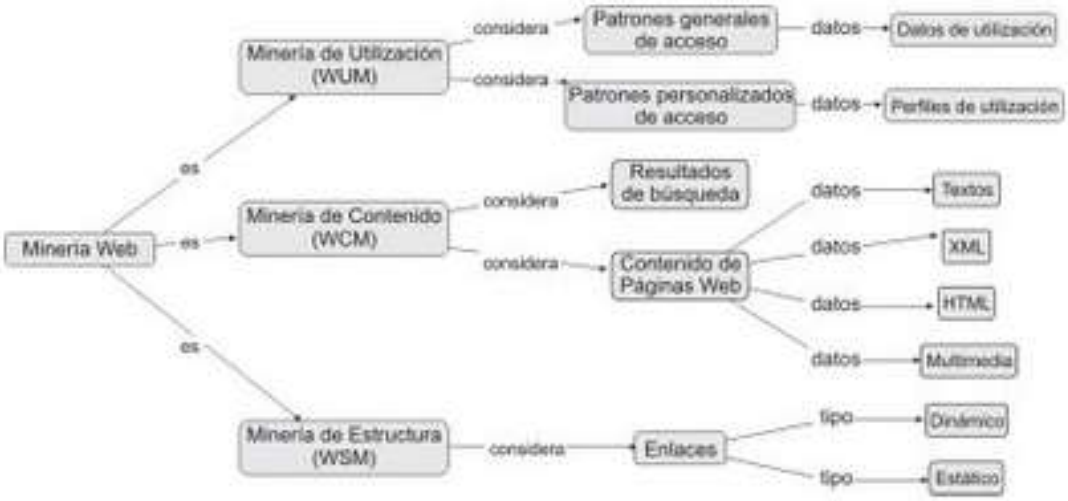
El alumno consultara los siguientes artículos de la biblioteca digital en el repositorio de REDALYC y realizara una red semántica que denote este concepto, sus aplicaciones y principales aportaciones de los distintos autores.



Descargará los siguientes artículos:

1. Velarde Martínez, Apolinar Minería de Datos. Una Introducción. Conciencia Tecnológica. 2003; (23):. [fecha de Consulta 23 de septiembre de 2019]. ISSN: 1405-5597. Disponible en: http://www.redalyc.org/articulo.oa?id=94402303
2. Ruiz, Roberto, Gilbert, Karina y Riquelme, José C. Presentación: Minería de Datos. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial. 2006;10(29): [fecha de Consulta 23 de septiembre de 2019]. ISSN: 1137-3601. Disponible en: http://www.redalyc.org/articulo.oa?id=925/92502901
3. Rodríguez Suárez, Yuniet y Díaz Amador, Anolandy Herramientas de Minería de Datos. Revista Cubana de Ciencias Informáticas. 2009;3(3-4): [fecha de Consulta 23 de septiembre de 2019]. ISSN: 1994-1536. Disponible en: http://www.redalyc.org/articulo.oa?id=3783/378343637009
4. Riquelme, José C., Ruiz, Roberto y Gilbert, Karina Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial. 2006;10(29):. [fecha de Consulta 23 de septiembre de 2019]. ISSN: 1137-3601. Disponible en: http://www.redalyc.org/articulo.oa?id=925/92502902
5. Caisés Almaguer, Yoel y Navarro Rodríguez, Ricardo Transformación de Características para la Minería de Datos. Ciencias Holguín. 2010;XVI(2): [fecha de Consulta 23 de septiembre de 2019]. ISSN: . Disponible en: http://www.redalyc.org/articulo.oa?id=1815/181517926009

6. Marcano Aular, Yelitza Josefina y Talavera Pereira, Rosalba Minería de Datos como soportea la toma de decisiones empresariales. Opción. 2007;23(52): [fecha de Consulta 23 de septiembre de 2019]. ISSN: 1012-1587. Disponible en: http://www.redalyc.org/articulo.oa?id=310/31005208



Con los artículos previamente leídos, el alumno deberá llenar la siguiente tabla y la discutirá en clase con los compañeros y el profesor.

Autor	Título del articulo	Objetivo principal	Aplicación	Revista y año
Velarde Martínez	Minería de Datos. Una Introducción			
Ruiz, Roberto	Presentación: Minería de Datos. Inteligencia Artificial.			
Rodríguez Suárez	Herramientas de Minería de Datos			

Toda la documentación elaborada deberá subirla a su portafolio de SEDUCA en la fecha indicada.

CONCLUSIONES

Anote de manera breve las principales conclusiones obtenidas al término de esta práctica

BIBLIOGRAFÍA

1. Chiotti S M, Cidisi O : Minería De Datos En Base De Datos De Servicios De Salud – Utn – Frsf, Ingar Utn- Conicet (2013).
2. Molina Félix, Luis Carlos.: Data Mining: Torturando A Los Datos Hasta Que Confiesen (2014).
3. Quesada Aznielles Yaneisis., Wong Pérez Daymi., Rosete Suárez Alejandro.: Minería De Datos Aplicada A La Gestión Hospitalaria. 14 Convención Científica De Ingeniería Y Arquitectura, CUJAE (2008)
4. Marcano Aular Yelitza Josefina Y Talavera Pereira Rosalba Minería De Datos Como Soporte A La Toma De Decisiones Empresariales Opción, Año 23, No 52 (2007): 104 – 118 ISSN 1012-1587 (2007).
5. Pérez C.: Técnicas De Muestreo Estadístico: Teoría, Práctica Y Aplicaciones Informáticas. Madrid: RA-MA pag. 700 (1999).
6. Zamarrón Sanz, Carlos., García Paz, Vanesa., Calvo Álvarez, Uxío., Pichel Guerrero, Fernanda., Rodríguez Suárez, José Ramón. Aplicación De La Minería De Dato. Universitario De Santiago De Compostela, Servicio De Neumología. Vol. 6: 156 – 166 (2006).
7. Savater, F.: El Valor De Educar. Barcelona: Ed. Ariel, 270 – 08008 (1997).
8. S. Mitra and T. Acharya. Data mining: multimedia, soft computing and bioinformatics. John Wiley & Sons, 2003.
9. José C. Riquelme, Roberto Ruiz, Karina Gilbert. Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, pp. 11-18 Asociación Española para la Inteligencia Artificial Valencia, España, vol. 10, núm. 29, primavera, 2006.
10. Pérez C.: Técnicas De Muestreo Estadístico: Teoría, Práctica Y Aplicaciones Informáticas. Madrid: RA-MA pag. 700 (1999).
11. Zamarrón Sanz, Carlos., García Paz, Vanesa., Calvo Álvarez, Uxío., Pichel Guerrero, Fernanda., Rodríguez Suárez, José Ramón. Aplicación De La Minería De Dato. Universitario De Santiago De Compostela, Servicio De Neumología. Vol. 6: 156 – 166 (2006).
12. Savater, F.: El Valor De Educar. Barcelona: Ed. Ariel, 270 – 08008 (1997).

PRÁCTICA 2
MODELOS DE PROCESO PARA PROYECTOS DE MINERÍA DE DATOS

OBJETIVO

El alumno conocerá los principales modelos de procesos aplicados para el diseño de proyectos de minería de datos.

INTRODUCCIÓN

El proceso de minería de datos comprende varios pasos como crear, probar y trabajar con los modelos de minería. El proyecto de minería de datos comienza con un plan bien definido de inteligencia comercial. Los analistas definen el problema a resolver y el objetivo concreto de que se desea cumplir. Cuanto mejor sea esta formulación inicial, más claras serán las directrices acerca de los datos y las funciones de minería que se utilizan para conseguir los resultados deseados.

El proyecto de minería de datos consta de las fases principales siguientes:

1. Selección y preparación de datos.
2. Creación del modelo de minería de datos (también denominada fase de *preparación*).
   Un modelo de minería de datos se crea a partir de un conjunto específico de datos de entrada.
   Durante el proceso de creación del modelo, una vez preparados los datos, debe especificar sus decisiones sobre:
   - Dónde residen los datos de entrada
   - Qué campos de los datos de entrada son apropiados
   - Qué valores se deben utilizar para la función de minería determinada que está utilizando
   - Dónde desea almacenar el modelo final
3. Prueba de un modelo y análisis de su calidad.
   Se puede probar un modelo de Clasificación o Regresión. Después, se puede analizar la calidad del modelo.
4. Utilización de un modelo que ofrece información acerca de:
   - La visualización de los resultados.
     Puede visualizar los resultados de la minería de datos para analizarlos e interpretarlos. Utilice Intelligent Miner Visualizer para ver y analizar los resultados.
   - Puntuación de los registros de datos.
     Los modelos se aplican a otros datos en la fase de aplicación de la minería de datos. Utilice Intelligent Miner para puntuar los registros de datos.
   - Análisis de un modelo y preparación para otros pasos del proceso.

Puede utilizar varias funciones para recuperar información acerca de la modelo contenida en las tablas a fin de que otros programas de aplicación realicen otros procesos.

Hoy en día, la minería de procesos se utiliza para hacer una descripción formal de los procesos a través de mecanismos que descubren, monitorean y sugieren mejoras en los procesos, por ejemplo, análisis de la atención médica en los hospitales, análisis del flujo de producción de las empresas, análisis de servicios de entrega, verificación de cumplimiento de normas y reglas de negocio, identificación de personas involucradas en los procesos y cómo estas se relacionan, entre otros. En general, este dominio se aplica en diferentes ámbitos como: organizaciones públicas y privadas, financieras, de salud, empresas tecnológicas y todo tipo de instituciones en general. En este sentido, la minería de procesos juega un importante papel en el crecimiento y éxito de las organizaciones; propiciando así la búsqueda y explotación de nuevas herramientas y métodos para el procesamiento de datos e información almacenada en los sistemas de información. Esta búsqueda ha originado el diseño de guías de trabajo, conocidos también como procedimientos, para el desarrollo de proyectos de minería de procesos, los cuales incluyen una serie de pasos o etapas a seguir para planear y guiar el desarrollo de estos proyectos. En la literatura actual se encontraron diferentes procedimientos que abarcan desde la planificación y justificación del proyecto, hasta la presentación de los resultados y las propuestas de mejoras en los procesos. Sin embargo, estos procedimientos cuentan con ciertas limitantes, como (Van der Aalst et al., 2012; Van der Heijden, 2012): a) falta de claridad en la presentación y visualización de los resultados obtenidos tras la aplicación de algún algoritmo o técnica, b) limitada participación del usuario en cada una de las etapas del proyecto, y c) falta de entendimiento de las características del entorno o contexto donde se desarrolla el proyecto. Estas limitantes traen como consecuencia proyectos no usables, disminuidos de entendimiento y hasta carentes de funcionalidad.

La minería de procesos es un conjunto de técnicas capaces de descubrir, monitorear y mejorar los procesos reales mediante la extracción de conocimiento a partir de los registros de eventos disponibles en los sistemas de información (Van der Aalst, 2011; Van der Aalst, 2016). La Figura 1 muestra un resumen sobre cómo se realizan estos análisis. En primera instancia se tienen los procesos que son soportados por los sistemas de información. Estos sistemas almacenan y coordinan los eventos que describen la historia de los procesos. Estos eventos pueden ser analizados a través de técnicas de minería de procesos que permiten descubrir el modelo real del proceso, y a partir de esta información mejorar el proceso.

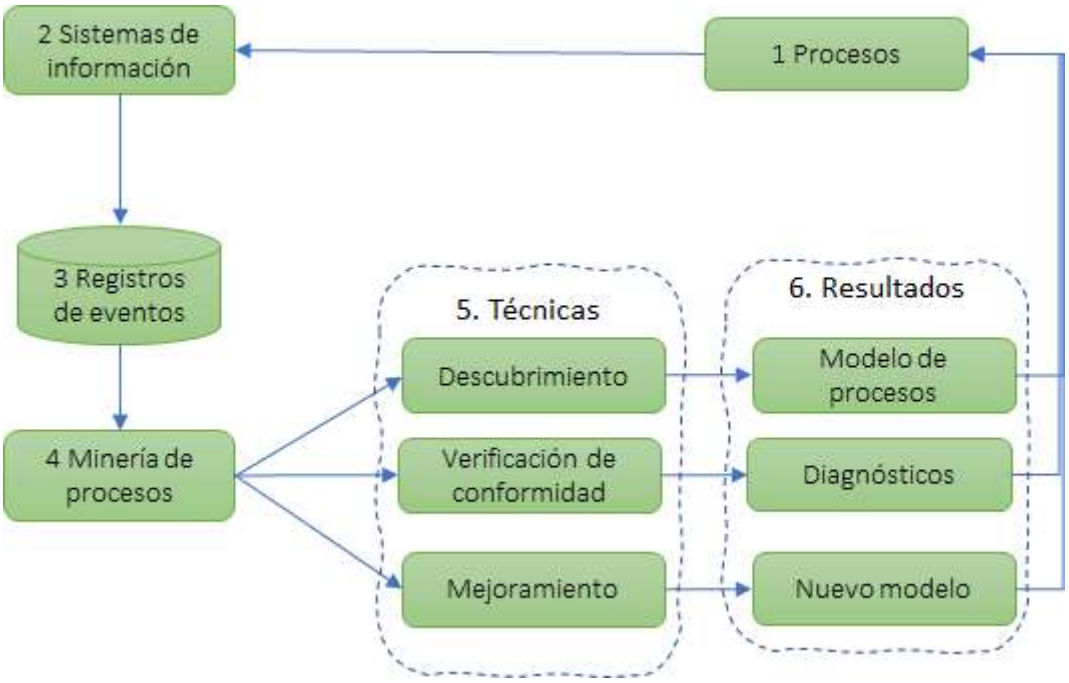


Figura 1. Minería e procesos. Adaptado de van der Aalst (2011).

La minería de procesos compensa la brecha entre el análisis de procesos basado en modelos tradicionales y las técnicas de análisis centradas en datos, tales como aprendizaje automático y minería de datos (van der Aalst, 2016). Estas áreas proporcionan técnicas para la extracción de conocimiento a partir de grandes series de datos como apoyo para la toma de decisiones (Molero et al., 2017).

Los procedimientos mencionados a pesar de su amplia variedad minimizan la participación del usuario en cada una de sus etapas, lo cual trae como consecuencia el desarrollo de proyectos de minería de procesos con limitaciones de usabilidad (Zhao et al., 2006) y accesibilidad (Horberry et al., 2015), puesto que no se consideran las necesidades y requerimientos de los usuarios durante el desarrollo del proyecto. Los puntos de vista de los usuarios, como sus preferencias, necesidades y decisiones, sirven para desarrollar aplicaciones de minería de procesos eficientes. Asimismo, la participación del usuario permite la creación de proyectos personalizados (Ho et al., 2002), es decir, proyectos hechos a la medida, basados en los requerimientos y necesidades de los usuarios.

DESARROLLO

El alumno conformara el contenido de la siguiente tabla basándose en la problemática a abordar para el empleo del proyecto de minería de datos, como un primer acercamiento de los principales componentes.

TAREAS GENERALES	Descripción
Describir el contexto y valorar la situación donde se desarrollará el proyecto	
Especificar los objetivos del proyecto	
Especificar los requerimientos de los usuarios y demás interesados	
Definir los objetivos de minería de procesos	
Elaborar un plan general del proyecto	

ACTIVIDADES	Descripción
Identificar usuarios y otros interesados válidos en el proyecto.	
Identificar las características relevantes de los usuarios.	
Identificar las metas y tareas de los usuarios.	
Identificar los entornos técnicos de la solución.	
Definir los objetivos del proyecto.	
Identificar los principales criterios de desempeño.	
Identificar las necesidades de los usuarios e interesados.	
Determinar los requisitos de los usuarios e interesados. **Resolver los conflictos entre los requisitos.**	
Garantizar la calidad de los requisitos.	
Elaborar el reporte de requisitos.	
Definir los objetivos de la minería de procesos.	
Definir un plan de datos, de comunicación, infraestructura, recursos humanos, entre otros.	

Salidas esperadas
Reporte sobre el contexto del proyecto. **Definición de requerimientos.** **Definición de los objetivos del proyecto y de la minería de procesos.** **Plan general del proyecto**

BIBLIOGRAFÍA

1. Chiotti S M, Cidisi O : Minería De Datos En Base De Datos De Servicios De Salud – Utn – Frsf, Ingar Utn- Conicet (2013).

2. Molina Félix, Luis Carlos.: Data Mining: Torturando A Los Datos Hasta Que Confiesen (2014).
3. Quesada Aznielles Yaneisis., Wong Pérez Daymi., Rosete Suárez Alejandro.: Minería De Datos Aplicada A La Gestión Hospitalaria. 14 Convención Científica De Ingeniería Y Arquitectura, CUJAE (2008)

4. Marcano Aular Yelitza Josefina Y Talavera Pereira Rosalba. Minería De Datos Como Soporte A La Toma De Decisiones Empresariales Opción, Año 23, No 52 (2007): 104 – 118 ISSN 1012-1587 (2007).

5. Pérez C.: Técnicas De Muestreo Estadístico: Teoría, Práctica Y Aplicaciones Informáticas. Madrid: RA-MA pag. 700 (1999).
6. Zamarrón Sanz, Carlos., García Paz, Vanesa., Calvo Álvarez, Uxío., Pichel Guerrero, Fernanda., Rodríguez Suárez, José Ramón. Aplicación De La Minería De Dato. Universitario De Santiago De Compostela, Servicio De Neumología. Vol. 6: 156 – 166 (2006).

7. ***Savater, F.: El Valor De Educar. Barcelona: Ed. Ariel, 270 – 08008 (1997).***
8. Van der Aalst, W. y Weijters, A. (2004). Process mining: a research agenda. Computers in industry, 53(3), pp. 231-244.
9. Van der Aalst, W. (2011). Process Mining: Discovery, Conformance and Enhancement of Business Processes (Springer-Verlag). Berlin.
10. Van der Aalst, W., Adriansyah, A., de Medeiros, A., Arcieri, F., Baier, T., Blickle, T., et al. (2012). Process mining manifesto. International Conference on Business Process Management Workshops.
11. Van der Aalst, W. (2016). Process Mining: Data Science in Action (Springer-Verlag). Berlin. Van der Heijden, T. (2012) Process mining project methodology: Developing a general approach to apply process mining in practice, Master of Science in Operations Management and Logistics, School of Industrial Engineering, Eindhoven University of Technology (TU/e).
12. Van Eck, M., Lu, X., Leemans, S. y van der Aalst, W. (2015). PM2 : A Process Mining Project Methodology. International Conference on Advanced Information Systems Engineering. Zhao, Y., Chen, Y., y Yao, Y. (2006). User-centered interactive data mining. Cognitive Informatics, 5th IEEE International Conference, 1, pp. 457-466.

PRÁCTICA 3
MODELOS DE PROCESO PARA PROYECTOS DE MINERÍA DE DATOS: METODOLOGÍA CRISP-DM

OBJETIVO

El alumno conocerá el proceso de la metodología CRISP-DM general y las fases que lo componen.

INTRODUCCIÓN

La metodología de CRISP-DM está descrita en términos de un modelo de proceso jerárquico, consistente en un conjunto de tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada, e instancia de procesos. En el nivel superior , el proceso de minería de datos es organizado en un número de fases; cada fase consiste en varias tareas genéricas de segundo nivel. Este segundo nivel lo llaman genérico porque está destinado a ser bastante general para cubrir todas las situaciones posibles de minería de datos. Las tareas genéricas están destinadas a ser tan completas y estables como sea posible.

Completo significa que cubre tanto al proceso entero de minería de datos y todas las aplicaciones de minería de datos posibles.

Estable significa que el modelo debería ser válido para acontecimientos normales y aún para desarrollos imprevistos como técnicas de modelado nuevo. El tercer nivel , el nivel de tarea especializado, es el lugar para describir como las acciones en las tareas genéricas deberían ser realizadas en ciertas situaciones específicas. Por ejemplo, en el segundo nivel podría haber una tarea genérica llamada limpieza de datos. El tercer nivel describe como esta tarea se diferencia en situaciones diferentes, como la limpieza de valores numéricos contra la limpieza de valores categóricos, o si el tipo de problema es agrupamiento o el modelado predictivo. La descripción de fases y tareas como pasos discretos realizados en un orden específico representa una secuencia idealizada de eventos. En la práctica, muchas de las tareas pueden ser realizadas en una orden diferente, y esto a menudo será necesario volver a hacer tareas anteriores repetidamente y repetir ciertas acciones.

El cuarto nivel , la instancia de proceso, es un registro de las acciones, decisiones, y de los resultados de una minería de datos real contratada. Una instancia de proceso esta organizado según las tareas definidas en los niveles más altos, pero representa lo que en realidad pasó en un contrato particular más bien que lo que pasa en general

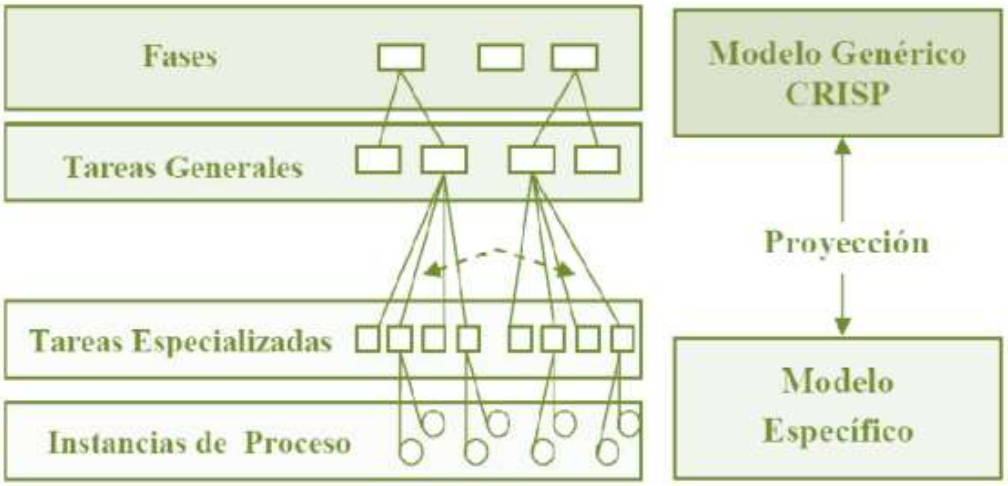


Figura 1: Cuatro niveles de la metodología CRISP-DM

La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas.

El modelo de proceso corriente para la minería de datos proporciona una descripción del ciclo de vida del proyecto de minería de datos. Este contiene las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción, no es posible identificar todas las relaciones. Las relaciones podrían existir entre cualquier tarea de minería de datos según los objetivos, el contexto, y –lo más importante- el interés del usuario sobre los datos. El ciclo de vida del proyecto de minería de datos consiste en seis fases, mostrado en la Figura 2. La secuencia de las fases no es rígida. El movimiento hacia adelante y hacia atrás entre fases diferentes es siempre requerido.

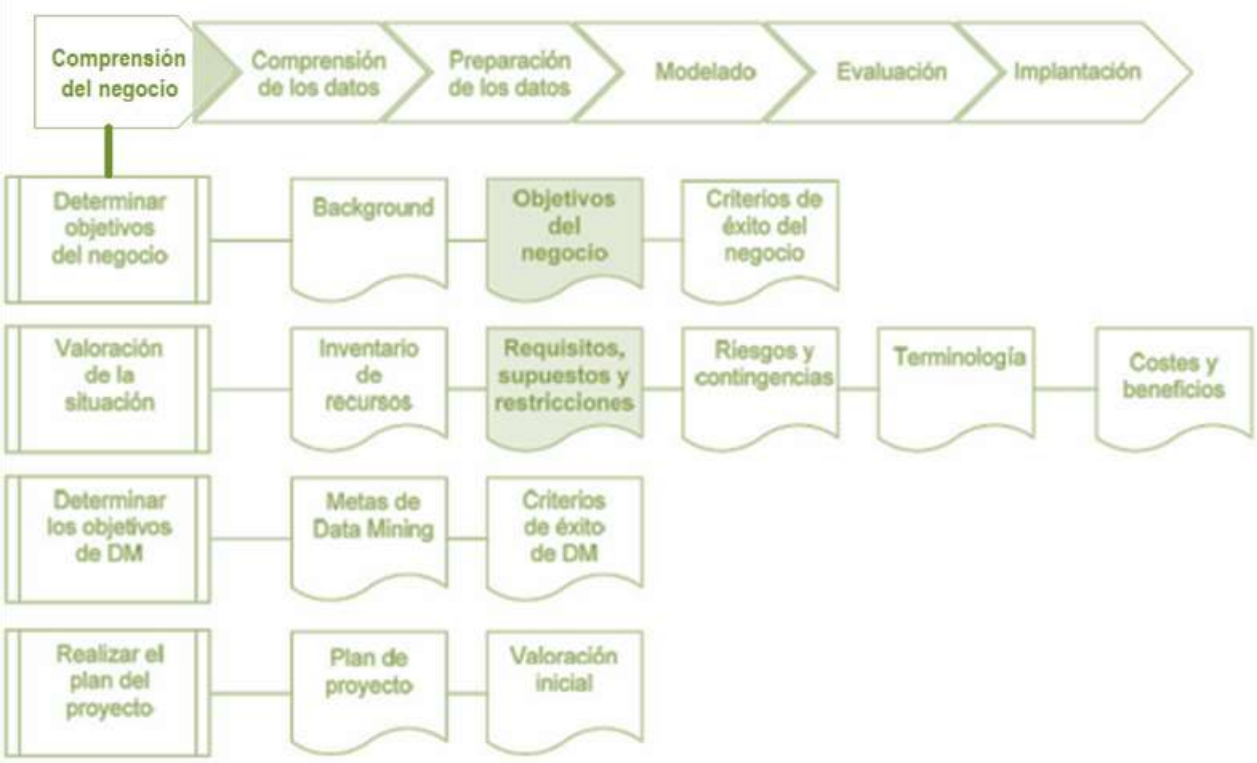


Figura 2. *Fase de comprensión del negocio ([CRISP-DM, 2000]).*

El resultado de cada fase determina que la fase, o la tarea particular de una fase, tienen que ser realizados después. Las flechas indican las más importantes y frecuentes dependencias entre fases.

En primer lugar, se lleva a cabo una investigación documental acerca del conocimiento denominado "Entendimiento del Negocio", el uso de técnicas para la reducción de dicho conocimiento y las principales metodologías de Explotación de Información. La revisión bibliográfica incluye artículos de revistas especializadas y de trabajos presentados en congresos, diversos libros, investigaciones llevadas a cabo por organismos de prestigio internacional, sitios web de entidades e instituciones vinculadas con la temática objeto de estudio, etc.

En segundo lugar, se realiza la construcción de un marco teórico para avanzar en el desarrollo del modelo de proceso de conceptualización del entendimiento del negocio; con base en el estudio de lo recopilado durante el ejercicio previo y en función de la información procesada, se consolida el informe final del estado de arte del proyecto. En tercer lugar, se plantea la problemática de investigación del presente trabajo, a partir de la complejidad que involucra la comprensión clara del dominio del negocio en proyectos de Explotación de Información. En cuarto lugar, y considerando la información procesada previamente, se diseña el nuevo modelo de proceso y se definen las fases, las tareas y los productos respectivos. Adicionalmente, se identifican las técnicas adecuadas aplicables al nuevo modelo para alcanzar la comprensión del negocio, modelando las mismas en función de los objetivos trazados.

Comprensión del negocio	•Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio,luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un planpreliminar diseñado para alcanzar los objetivos.
Comprensión de los datos	•La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permitenfamiliarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en losdatos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.
Preparación de los datos	•La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final [los datosque serán provistos en las herramientas de modelado] de los datos en brutos iniciales. Las tareas de preparación de datosprobablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección detablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.
Modelado	•En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valoresóptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienenrequerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudonecesario.
Evaluación	•En esta etapa en el proyecto, usted ha construido un modelo (o modelos) que parece tener la alta calidad de unaperspectiva de análisis de datos. Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutadospara crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, unadecisión en el uso de los resultados de minería de datos debería ser obtenida.
Desarrollo	•La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar elconocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el clientepueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de unaorganización, por ejemplo, en tiempo real la personalización de página Web o la repetida obtención de bases de datos demercadeo.

DESARROLLO

El alumno conformara el contenido de la siguiente tabla basándose en la problemática a abordar para el empleo del proyecto de minería de datos, considerando de manera general las fases de la metodología CRISP MD

TAREAS GENERALES	DESCRIPCIÓN
Comprensión del negocio	
Comprensión de los datos	
Preparación de los datos	
Modelado	
Evaluación	
Desarrollo	

CONCLUSIONES

Anote de manera breve las principales conclusiones obtenidas al término de esta práctica

BIBLIOGRAFÍA

1. Chiotti S M, Cidisi O : Minería De Datos En Base De Datos De Servicios De Salud – Utn – Frsf, Ingar Utn- Conicet (2013).

2. Molina Félix, Luis Carlos.: Data Mining: Torturando A Los Datos Hasta Que Confiesen (2014).
3. Quesada Aznielles Yaneisis., Wong Pérez Daymi., Rosete Suárez Alejandro.: Minería De Datos Aplicada A La Gestión Hospitalaria. 14 convención Científica De Ingeniería Y Arquitectura, CUJAE (2008)

4. Marcano Aular Yelitza Josefina Y Talavera Pereira Rosalba. Minería De Datos Como Soporte A La Toma De Decisiones Empresariales Opción, Año 23, No 52 (2007): 104 – 118 ISSN 1012-1587 (2007).

5. Pérez C.: Técnicas De Muestreo Estadístico: Teoría, Práctica Y Aplicaciones Informáticas. Madrid: RA-MA pag. 700 (1999).

6. Zamarrón Sanz, Carlos., García Paz, Vanesa., Calvo Álvarez, Uxío., Pichel Guerrero, Fernanda., Rodríguez Suárez, José Ramón. Aplicación De La Minería De Dato. Universitario De Santiago De Compostela, Servicio De Neumología. Vol. 6: 156 – 166 (2006).

7. Savater, F.: El Valor De Educar. Barcelona: Ed. Ariel, 270 – 08008 (1997).
8. Van der Aalst, W. y Weijters, A. (2004). Process mining: a research agenda. Computers in industry, 53(3), pp. 231-244.

9. León León, O., Asato España, J. (2009). La Importancia del Modelado de Procesos de Negocio como Herramienta para la Mejora e Innovación. Revista Panorama Administrativo, http://admon.itc.mx/ojs/index.php/panorama/article/view/155/156.

10. Amón Uribe, I., Jiménez Ramírez, C. (2009). Hacia una Metodología para la Selección de Técnicas de Depuración de Datos. Revista Avances en Sistemas e Informática. 6(1): 185-190. ISSN 1657-7663

PRÁCTICA 4
METODOLOGÍA CRISP-DM: FASE DE COMPRENSIÓN DEL NEGOCIO O PROBLEMA

OBJETIVO

- El alumno conocerá las subfases que componen la primera fase de la metodología CRISP MD y desarrollara cada una de ellas en la integración de su propuesta de aplicación.

INTRODUCCIÓN

La etimología de la Palabra *Negocio*, proviene de las palabras en latín *nec* y *otium*, es decir lo que no es ocio, e implica acción o movimiento, y constituye el elemento central a gestionar, si se persigue el éxito de cualquier organización[1]. Para los romanos, la palabra *otium* era lo que se realizaba en el tiempo libre, sin ningún tipo de recompensa; por consiguiente, negocio para ellos era lo que se hacía por dinero[2].

¿Qué es el Negocio? El negocio es un sistema complejo, está compuesto por una organización en la cual generalmente se diferencian elementos tangibles e intangibles, tales como son los equipos de trabajo formales y las funciones que estos desempeñan. Algunas de las funciones que se distinguen, sólo se llevan a cabo en un solo departamento o sección de la organización respectiva, mientras que otras tantas son comunes a diversos departamentos de esta. Las empresas son estructuras formales, que se construyen para llevar adelante un negocio determinado, pudiendo ser creadas para perdurar en el tiempo o bien, por el contrario, para explotar un negocio por un período de tiempo determinado[3].

Por otra parte, se considera que un negocio es mucho más que un simple producto o un servicio determinado, añadiendo que el negocio se define en base a su misión, sus objetivos, las estrategias a implementar para alcanzar los mismos, sus competidores, sus clientes y los productos o servicios que ofrece[4].

---

[1] Morales, M., Cancino, C. (2009). La RSE como Herramienta Estratégica del Negocio. Revista Economía & Administración. Facultad de Economía y Negocios. Universidad de Chile. Nro. 158. Pág. 48-57. ISSN 076-4793.

[2] Cedeño Ruíz, T. F., Ceballos Centeno, M. C., Guevara Catagua, L. P., Vadiviezo Macías, J. L. (2010). Adecuación de un Ambiente Administrativo-Pedagógico e Implementación de un Plan de Capacitación para Administrar Pequeñas Microempresas Dirigido a los Integrantes del Comité de Desarrollo Comunitario "José Lívido Intriago" de la Comunidad "Las Mercedes N° 1" del Cantón Santa Ana. Tesis de Grado. Universidad Técnica de Manabí, Facultad de Ciencias Administrativas y Económicas. Manabí, Ecuador.

[3] Cancino, C., Morales, M. (2010). La Racionalidad de Comprometerse con el Negocio. Revista Trend Management. Edición Especial: Management Made in Chile.

[4] Ochoa, M. (2006). Uso de Técnicas de Educción para el Entendimiento de Negocio. Tesis de Maestría. Universidad Politécnica de Madrid

Proceso de Negocio

Para alcanzar sus objetivos, una empresa organiza sus actividades por medio de un conjunto de procesos denominados "Procesos de Negocios". Cada uno de ellos involucra un conjunto de datos que son producidos y procesados mediante una serie de tareas específicas, en la que intervienen determinados agentes (como pueden ser los empleados o departamentos de la propia organización), de acuerdo con un flujo de trabajo determinado. Asimismo, todos estos procesos se hallan regulados por un conjunto de reglas de negocio, que determinan la estructura de la información y las políticas respectivas de la propia empresa[5] [6].

En la actualidad, las empresas y la comunidad científica están centrado su atención a los procesos de negocio, ya que reconocen que ellos constituyen un recurso fundamental para el desempeño y la obtención de ventajas competitivas en el mercado. De acuerdo con[7] (Jennings, 2000), los procesos de negocio están centrados en el mercado. Por consiguiente, se deben revisar y ajustar los mismos de manera constante para incorporar posibles mejoras que permitan: mejorar la calidad de los productos o servicios, aumentar la satisfacción del cliente, reducir costos, optimizar la eficiencia en las operaciones del negocio, encontrar nuevos negocios u oportunidades para reemplazar servicios o productos existentes o bien introducir nuevos.
A pesar de la importancia que tienen los procesos de negocios, la gran mayoría de las organizaciones no representa esquemáticamente cómo son sus procesos. Es decir, no es una práctica habitual que los mismos sean representados mediante modelos, y en tal sentido, que sirvan para la confección de una base que facilite la toma de decisiones. Asimismo, la gran mayoría de los modelos no consideran los sistemas informáticos que apoyan las actividades de sus procesos de negocio, a pesar de ser un aspecto fundamental de cómo se realiza el trabajo en el ámbito empresarial. Por el contrario, aquellos modelos que sí lo hacen, sólo consideran este aspecto implícitamente o bien de una manera insuficiente[8].

Entender el Negocio

Existe un tipo especial de conocimiento llamado "Entendimiento del Negocio" al cual hacen referencia las metodologías de desarrollo de sistemas de información. A pesar que en algunas de ellas se expone dicho conocimiento, se hace evidente la falta de información acerca de las herramientas que integran el soporte específico del proceso de educción y los parámetros involucrados en el proceso respectivo para alcanzar la comprensión del negocio. Asimismo, se desconoce la información relevante a manejar para lograr una comprensión adecuada del negocio ya que estas metodologías no describen dicha información. Por el contrario, todas las metodologías de desarrollo de proyectos de software dan por

[5] García Molina, J., Ortín, M. J., Moros, B., Nicolás, J. (2007). De los Procesos del Negocio a los Casos de Uso. Técnica Administrativa, ISSN: 1666-1680 (en línea), 6(4).
[6] Ortín, M. J., García Molina, J., Moros, B., Nicolás, J. (2001). El Modelo del Negocio como base del Modelo de Requisitos. Grupo de Investigación de Ingeniería del Software. Departamento de Informática y Sistemas. Facultad de Informática. Universidad de Murcia. España.
[7] Jennings, N. R., Norman, T. J., Faratin, P., O'Brien, P., Odgers, B. (2000). Autonomous Agents for Business Process Management. Applied Artificial Intelligence. Vol. 14(2). Pág. 145-189.
[8] Jiménez Quintana, C., Farías Valenzuela, L., Pinto, F., Neriz Jara, L. (2003). Análisis de Modelos de Procesos de Negocios en Relación a la Dimensión Informática. Departamento de Ingeniería Informática y Ciencias de la Computación. Universidad de Concepción. Nº 9. ISSN 0717-4195.

garantizado el conocimiento que se tiene del dominio del negocio, sin facilitar al ingeniero los instrumentos necesarios para abordar de manera precisa el entendimiento del entorno de trabajo. Siguiendo con la misma línea de trabajo, una *protofase* que articula un conjunto de técnicas y herramientas asociadas para lograr la reducción del entendimiento del negocio, independientemente del sistema de información a desarrollar, ya sea que se trate de un sistema basado en conocimiento, un sistema de gestión tradicional o bien, un sistema de Explotación de Información.

Asimismo, se menciona que actualmente, la mayoría de los negocios utilizan algún tipo particular de sistemas de información; sin embargo, muchas compañías no están conformes con la calidad de estos, aduciendo que son difíciles de manejar, son poco confiables, ofrecen un soporte de negocio incompleto u obsoleto, y no se complementan con otros sistemas adicionales. En la mayoría de los casos, esto es consecuencia de un entendimiento incorrecto del negocio al momento de llevar a cabo el desarrollo de estos sistemas. Esto es común en aquellas empresas que usan numerosos sistemas informáticos pequeños, que, aunque soportan ciertos procesos del negocio, funcionan de manera autónoma, de manera tal que es usual que la información generada por dichos sistemas sea inconsistente. Por tanto, en base a lo expresado en el párrafo anterior, se concluye que un entendimiento adecuado del negocio contribuye al desarrollo de estos sistemas de información.

Esta primera fase inicial, se basa en el entendimiento de los objetivos del proyecto y la comprensión de los requerimientos de este desde el punto de vista del negocio, a fin de definir el problema a resolver y diseñar una planificación preliminar para el cumplimiento efectivo de los objetivos en cuestión.

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema (Figura 1), es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea permitirá obtener resultados fiables. Para obtener el mejor provecho de Data Mining, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Data Mining y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. Una descripción de cada una de las principales tareas que componen esta fase es la siguiente:

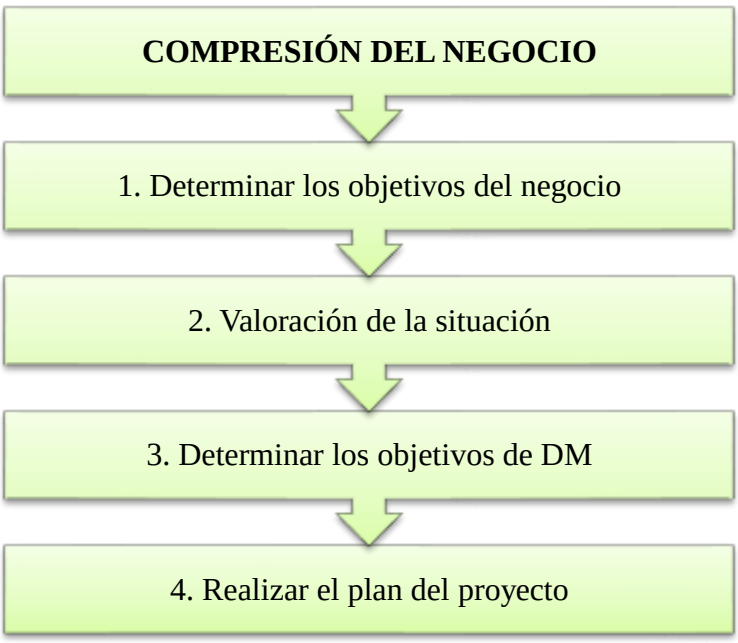


Figura 1. Fase 1 de la metodología CRISP MD

Determinar los objetivos del negocio. Esta es la primera tarea para desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Data Mining y definir los criterios de éxito. Los problemas pueden ser diversos como, por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio, califica el resultado del proceso de DM, o de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.

Evaluación de la situación. En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de DM, considerando aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de DM?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de Data Mining.

Determinación de los objetivos de DM. Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de DM, como, por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de DM será, por ejemplo, determinar el perfil de los clientes respecto de su capacidad de endeudamiento.
Producción de un plan del proyecto. Finalmente, esta última tarea de la primera fase de CRISP-DM, tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada paso.

Lo anterior puede resumirse de la siguiente manera, considerando las actividades asociadas alas componentes de la fase correspondiente:

FASE	TAREAS COMPONENTES	ACTIVIDADES ASOCIADAS
Comprensión del negocio	Determinar los objetivos del negocio.	✓ Background ✓ Objetivos del negocio ✓ Criterios de éxito del negocio
	Evaluar la situación.	✓ Inventarios de recursos ✓ Requisitos, supuestos y requerimientos. ✓ Riesgos y contingencias. ✓ Terminología. ✓ Costos y beneficios.
	Determinar objetivos del proyecto de explotación de la información.	✓ Las metas del proyecto de explotación de información. ✓ Criterios de éxito de proyecto de explotación de información.
	Realizar el plan del proyecto.	✓ Plan del proyecto. ✓ Valoración inicial de herramientas.

DESARROLLO

El alumno conformara el contenido de la siguiente tabla basándose en la problemática a abordar para el empleo del proyecto de minería de datos, considerando la primera fase de la metodología CRISP MD y sus actividades asociadas las componentes que le corresponde.

FASE 1. COMPRENSIÓN DEL NEGOCIO	
COMPONENTES	ACTIVIDADES ASOCIADAS
Determinar los objetivos del negocio.	
Evaluar la situación.	
Determinar objetivos del	

proyecto de explotación de la información.	
Realizar el plan del proyecto.	

CONCLUSIONES

Anote de manera breve las principales conclusiones obtenidas al término de esta práctica

BIBLIOGRAFÍA

Básico:

1. Advances in knowledge discovery. Fayyad usama m. Pearson (1996)
2. Introducción a la minería de datos. Hernández orallo José. Pearson (2004)
3. Inteligencia artificial un enfoque moderno. Stuart Russell. Alfa omega (2004).
4. Inteligencia artificial e ingeniería del conocimiento. Gonzalo pajares Martin Sanz. Alfa omega (2006)
5. Inteligencia artificial. Pedro Ponce cruz. Alfa omega (2010).
6. Inteligencia artificial técnicas, métodos y aplicaciones. José t. Palma Méndez, roque Martin morales. Alfa omega (2008).
7. Essentials of artificial inteligence. Matt Ginsberg. Morgan kaufman (1993)
8. Artificial intelligence. George f. Luger. Pearson (2009).

9. The mechanical mind in history. Philip husbands, Owen Holland. The mit press (2008).
10. Computational intelligence a logical approach. David Poole, Alan Mack worth. Oxford (1998).
11. Data mining. Ian h. Witten, eibe frank. Morgan kaufman (2011).
12. Introduction to data mining. Pang-ning tan, michael steinbach. Pearson (2006).
13. Data mining concepts and techiniques. Jawei han, Michael kamber jian pei. Morgan Kaufman (2011)

Complementario:

1. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.
2. Mitchell, T., *Machine Learning*, McGraw-Hill, 1997.
3. Quinlan, J.R., *C4.5 Programs for Machine Learning*, Morgan-Kaufmann, 1993
4. José Hernández Orallo, M.José Ramírez Quintana, "Introducción a la Minería de datos", Editorial Pearson 2004.
5. Cesar Perez López & Santin González Daniel, "Minería de datos : técnicas y herramientas"Editorial Thomson Paraninfo

PRÁCTICA 5
METODOLOGÍA CRISP-DM: FASE DE COMPRENSIÓN DE LOS DATOS

OBJETIVO

- El alumno conocerá las subfases que componen la segunda fase de la metodología CRISP-MD y desarrollara cada una de ellas en la integración de su propuesta de aplicación.

INTRODUCCIÓN

La segunda fase, es la fase de comprensión de los datos (Figura 1), que consiste en la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y abundantes accesos a la base de datos al objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas.

Figura 1. Fase 2 de la metodología CRISP MD

Las principales tareas para desarrollar en esta fase del proceso son:

Recolección de datos iniciales. La primera tarea en esta segunda fase del proceso de CRISP-DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.

Descripción de los datos. Después de adquiridos los datos iniciales, estos deben ser descritos.
Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

Exploración de datos. A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.

Lo anterior puede resumirse de la siguiente manera, considerando las actividades asociadas alas componentes de la fase correspondiente:

FASE	TAREAS COMPONENTES	ACTIVIDADES ASOCIADAS
Comprensión de datos	Recolectar los datos iniciales	✓ Reporte de recolección de datos iniciales
	Descubrir datos	✓ Reporte de descripción de los datos
	Explorar los datos	✓ Reporte de exploración de datos
	Verificar la calidad de datos	✓ Reporte de la calidad de datos

Puntualmente cada una de esas subfases puede tener una ejecución, como a continuación se muestra:

RECOLECCIÓN DE DATOS INICIALES	
Tarea	**Recolectar datos iniciales**
	Adquiera en el proyecto los datos (o el acceso a los datos) listados en los recursos del proyecto. Esta colección inicial incluye carga de datos, si es necesario para la comprensión de los datos. Por ejemplo, si usted usa un instrumento específico para la comprensión de los datos, esto perfectamente se entiende para abrir sus datos en esta herramienta. Este esfuerzo posiblemente conduce a los pasos iniciales de preparación de datos. Note: si usted adquiere datos de múltiples fuentes, la integración es una cuestión adicional, aquí o más tarde en las fases de preparación de datos más.
Salida	**Informe de colección de datos inicial**
	Liste el conjunto de dato(s) adquirido(s), juntos con sus posiciones, los métodos usados para adquirirlos, y algunos de los problemas encontrados.

	Registre los problemas encontrados y algunas de las resoluciones alcanzadas. Esto ayudará con la réplica (observación) futura de este proyecto o con la ejecución de proyectos similares futuros.

DESCRIBIR LOS DATOS	
Tarea	Describir los datos
	Examine las propiedades "gruesas" o "superficiales" de los datos e informe adquiridos en los resultados.
Salida	Informe de descripción de datos
	Describa los datos que han sido adquiridos, incluyendo el formato de los datos, la cantidad de datos (por ejemplo, el número de registros y campos en cada tabla), los identificadores de los campos, y cualquier otro rasgo superficial que ha sido descubierto. Evalúe si los datos adquiridos satisfacen las exigencias relevantes.

EXPLORAR LOS DATOS	
Tarea	Explorar los datos
	Esta tarea dirige interrogantes de minería de datos usando preguntas, visualización, y técnicas de reporte. Estos incluyen la distribución de atributos claves (por ejemplo, el atributo objetivo de una tarea de predicción) relacionados entre pares o pequeños números de atributos, los resultados de simples agregaciones, las propiedades de las subpoblaciones significativas, y análisis estadísticos simples. Estos análisis directamente pueden dirigir los objetivos de minería de datos; ellos también pueden contribuir o refinar la descripción de datos e informes de calidad, y alimentar en la transformación y otros pasos de preparación de datos necesarios para análisis futuros.
Salida	Informe de exploración de datos
	Describa los resultados de esta tarea, incluyendo primeras conclusiones o hipótesis iniciales y su impacto sobre el resto del proyecto. Si es apropiado, incluya gráficos y plots para indicar las características de datos que sugieren más examen de subconjuntos de datos interesantes

VERIFICAR LA CALIDAD DE LOS DATOS	
Tarea	Verificar la calidad de los datos
	Examine la calidad de los datos, dirigiendo preguntas como: ¿Los datos están completos? (¿Esto cubre todos los casos requeridos)? ¿Son correctos, o estos contienen errores y, si hay errores, que tan comunes son estos? ¿Hay valores omitidos en los datos? Si es así, ¿cómo se representan estos, donde ocurre esto, y que tan comunes son estos?

Salida	Informe de calidad de datos
	Liste los resultados de la verificación de calidad de datos; si existen problemas de calidad, liste las posibles soluciones. Las soluciones a los problemas de calidad de datos generalmente dependen tanto del conocimiento de los datos y como del negocio.

DESARROLLO

El alumno conformara el contenido de la siguiente tabla basándose en la problemática a abordar para el empleo del proyecto de minería de datos, considerando la SEGUNDA fase de la metodología CRISP MD y sus actividades asociadas a las componentes que le corresponde.

FASE 2. COMPRENSIÓN DE DATOS	
COMPONENTES	ACTIVIDADES ASOCIADAS
Recolectar los datos iniciales	
Descubrir datos	
Explorar los datos	
Verificar la calidad de datos	

CONCLUSIONES

Anote de manera breve las principales conclusiones obtenidas al término de esta práctica

BIBLIOGRAFÍA

Básico:

1. Advances in knowledge discovery. Fayyad usama m. Pearson (1996)
2. Introducción a la minería de datos. Hernández orallo José. Pearson (2004)
3. Inteligencia artificial un enfoque moderno. Stuart Russell. Alfa omega (2004).
4. Inteligencia artificial e ingeniería del conocimiento. Gonzalo pajares Martin Sanz. Alfa omega (2006)
5. Inteligencia artificial. Pedro Ponce cruz. Alfa omega (2010).
6. Inteligencia artificial técnicas, métodos y aplicaciones. José t. Palma Méndez, roque Martin morales. Alfa omega (2008).
7. Essentials of artificial inteligence. Matt Ginsberg. Morgan kaufman (1993)
8. Artificial intelligence. George f. Luger. Pearson (2009).
9. The mechanical mind in history. Philip husbands, Owen Holland. The mit press (2008).
10. Computational intelligence a logical approach. David Poole, Alan Mack worth. Oxford (1998).
11. Data mining. Ian h. Witten, eibe frank. Morgan kaufman (2011).
12. Introduction to data mining. Pang-ning tan, michael steinbach. Pearson (2006).
13. Data mining concepts and techiniques. Jawei han, Michael kamber jian pei. Morgan Kaufman (2011)

Complementario:

1. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.

PRÁCTICA 6
METODOLOGÍA CRISP-DM: FASE DE PREPARACIÓN DE LOS DATOS

OBJETIVO

- El alumno conocerá las subfases que componen la tercera fase de la metodología CRISP-MD y desarrollara cada una de ellas en la integración de su propuesta de aplicación.

INTRODUCCIÓN

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.
Esta fase se encuentra relacionada con la fase de modelado, puesto que, en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas. Es así que las fases de preparación y modelado interactúan de forma permanente. La figura 1, ilustra las áreas de que se compone ésta, e identifica sus salidas.

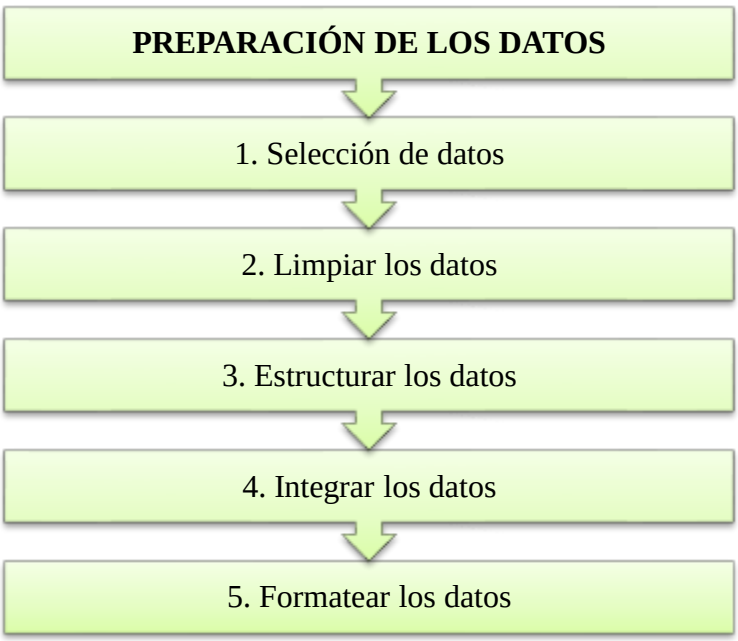


Figura 1. Fase 3 de la metodología CRISP MD

Una descripción de las tareas involucradas en esta fase es la siguiente:

Selección de datos. En esta etapa, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas.

Limpieza de los datos. Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos al objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.

Estructuración de los datos. Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.

Integración de los datos. La integración de los datos involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

Formateo de los datos. Esta tarea consiste principalmente, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de DM en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

Lo anterior puede resumirse de la siguiente manera, considerando las actividades asociadas a las componentes de la fase correspondiente:

FASE	TAREAS COMPONENTES	ACTIVIDADES ASOCIADAS
Preparación de datos	Selección de datos	✓ Inclusión/ exclusión de datos
	Limpiar los datos	✓ Reporte de calidad de datos limpios
	Estructurar los datos	✓ Derivación de atributos ✓ Generación de registros
	Integrar los datos	✓ Unificación de datos
	Formatear los datos	✓ Reporte de calidad de datos

Puntualmente cada una de esas subfases puede tener una ejecución, como a continuación se muestra:

Selección de datos	
Tarea	Selección de datos
	Decidir qué datos serán usados para el análisis. Los criterios incluyen la importancia a los objetivos de la minería de datos, la calidad, y las restricciones técnicas como límites sobre el volumen de datos o los tipos de datos. Note que la selección de datos cubre la selección de atributos (columnas) así como la selección de registros (filas) en una tabla.
Salida	Razonamiento para la inclusión/exclusión
	Listar los datos para ser incluidos/excluidos y los motivos para estas decisiones

Limpieza de datos	
Tarea	Limpieza de datos
	Elevar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede implicar la selección de los subconjuntos de datos limpios, la inserción de datos por defectos adecuados, o técnicas más ambiciosas tales como la estimación de datos faltantes mediante modelado
Salida	Informe de la limpieza de los datos
	Describa que decisiones y acciones fueron tomadas para dirigir los problemas de calidad de datos informados durante la tarea de Verificación de Calidad de Datos de los Datos de la fase de Comprensión de Datos. Las transformaciones de los datos para una apropiada limpieza y el posible impacto en el análisis de resultados deberían ser considerados

Construir datos	
Tarea	Construir datos
	Esta tarea incluye la construcción de operaciones de preparación de datos tales como la producción de atributos derivados o el ingreso de nuevos registros, o la transformación de valores para atributos existentes.
Salida	Atributos derivados
	Los atributos derivados son los atributos nuevos que son construidos de uno o más atributos existentes en el mismo registro. Ejemplo: área = longitud * anchura.

Integrar datos	
Tarea	Integrar datos
	Estos son los métodos por el cual la información es combinada de múltiples tablas o registros para crear nuevos registros o valores.
Salida	Combinación de datos
	La combinación de tablas se refiere a la unión simultánea de dos o más tablas que tienen información diferente sobre el mismo objeto. Ejemplo: una cadena de venta al público tiene una tabla con la información sobre las características generales de cada tienda (Por ejemplo, el espacio, el tipo de comercio), otra tabla con datos resumidos de las ventas (por ejemplo, el beneficio, el cambio porcentual en ventas desde el año anterior), y el otro con información sobre los datos demográficos del área circundante. Cada una de estas tablas contiene un registro para cada tienda. Estas tablas pueden ser combinadas simultáneamente en una nueva tabla con un registro para cada tienda, combinando campos de las tablas fuentes. Los datos combinados también cubren agregaciones. La agregación se refiere a operaciones en la que nuevos valores son calculados de información resumida de múltiples registros y/o tablas. Por ejemplo, convirtiendo una tabla de compra de clientes donde hay un registro para cada compra en una tabla nueva donde hay un registro para cada cliente, con campos tales como el número de compras, el promedio de la cantidad de compra, el porcentaje de ordenes cobrados a tarjeta de crédito, el porcentaje de artículos bajo promoción, etc.

Formatear datos	
Tarea	Formatear datos
	Formateando transformaciones se refiere a modificaciones principalmente sintácticas hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado.
Salida	Datos reformateados
	Algunas herramientas tienen requerimientos sobre el orden de los atributos, tales como el primer campo que es un único identificador para cada registro o el último campo es el campo resultado que el modelo debe predecir. Podría ser importante cambiar el orden de los registros en el conjunto de datos. Quizás la herramienta de modelado requiere que los registros sean clasificados según el valor del atributo de resultado. Comúnmente, los registros del conjunto de datos son ordenados al principio de algún modo, pero el algoritmo que modela necesita que ellos estén en un orden moderadamente arbitrario. Por ejemplo, cuando se usa redes neuronales, esto es generalmente mejor para los registros para ser presentados en un orden aleatorio, aunque algunas herramientas manejen esto automáticamente sin la intervención explicita del usuario. Además, hay cambios puramente sintácticos hechos para satisfacer las exigencias de la herramienta de modelado específica. Ejemplos: el quitar de comas de adentro de campos de texto en ficheros de datos delimitados por coma, corta todos los valores a un

	máximo de 32 caracteres.

DESARROLLO

El alumno conformara el contenido de la siguiente tabla basándose en la problemática a abordar para el empleo del proyecto de minería de datos, considerando la tercera fase de la metodología CRISP MD y sus actividades asociadas a las componentes que le corresponde.

FASE 3. PREPARACIÓN DE LOS DATOS	
COMPONENTES	**SALIDAS**
Selección de datos	
Limpiar los datos	
Estructurar los datos	
Integrar los datos	
Formatear los datos	

CONCLUSIONES

Anote de manera breve las principales conclusiones obtenidas al término de esta práctica

BIBLIOGRAFÍA

Básico:

1. Advances in knowledge discovery. Fayyad usama m. Pearson (1996)
2. Introducción a la minería de datos. Hernández orallo José. Pearson (2004)
3. Inteligencia artificial un enfoque moderno. Stuart Russell. Alfa omega (2004).
4. Inteligencia artificial e ingeniería del conocimiento. Gonzalo pajares Martin Sanz. Alfa omega (2006)
5. Inteligencia artificial. Pedro Ponce cruz. Alfa omega (2010).
6. Inteligencia artificial técnicas, métodos y aplicaciones. José t. Palma Méndez, roque Martin morales. Alfa omega (2008).
7. Essentials of artificial inteligence. Matt Ginsberg. Morgan kaufman (1993)
8. Artificial intelligence. George f. Luger. Pearson (2009).
9. The mechanical mind in history. Philip husbands, Owen Holland. The mit press (2008).
10. Computational intelligence a logical approach. David Poole, Alan Mack worth. Oxford (1998).
11. Data mining. Ian h. Witten, eibe frank. Morgan kaufman (2011).
12. Introduction to data mining. Pang-ning tan, michael steinbach. Pearson (2006).
13. Data mining concepts and techiniques. Jawei han, Michael kamber jian pei. Morgan Kaufman (2011)

Complementario:

1. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.
2. Mitchell, T., *Machine Learning*, McGraw-Hill, 1997.
3. Quinlan, J.R., *C4.5 Programs for Machine Learning*, Morgan-Kaufmann, 1993
4. José Hernández Orallo, M.José Ramírez Quintana, "Introducción a la Minería de datos", Editorial Pearson 2004.
5. Cesar Perez López & Santin González Daniel, "Minería de datos : técnicas y herramientas"Editorial Thomson Paraninfo

PRÁCTICA 7
METODOLOGÍA CRISP-DM: FASE DE MODELADO

OBJETIVO

- El alumno conocerá las subfases que componen la cuarta fase de la metodología CRISP-MD y desarrollara cada una de ellas en la integración de su propuesta de aplicación.

INTRODUCCIÓN

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas para utilizar en esta fase se eligen en función de los siguientes criterios:

o Ser apropiada al problema.
o Disponer de datos adecuados.
o Cumplir los requisitos del problema.
o Tiempo adecuado para obtener un modelo.
o Conocimiento de la técnica.
La figura 1 ilustra las tareas y resultados que se obtienen en esta fase. Una descripción de las principales tareas de esta fase.

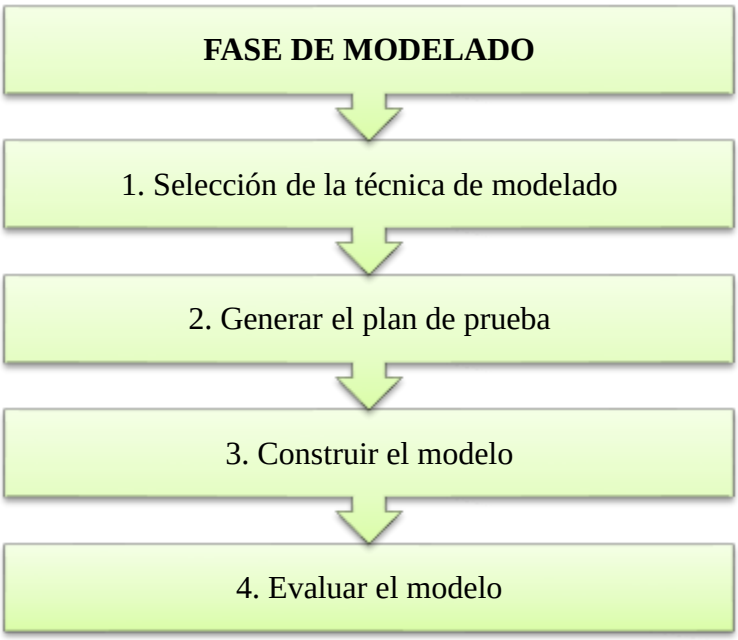


Figura 1. Fase 4 de la metodología CRISP MD

Previamente al modelado de los datos, se debe determinar un método de evaluación de los modelos que permita establecer el grado de bondad de ellos. Después de concluir estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo.

Selección de la técnica de modelado. Esta tarea consiste en la selección de la técnica de DM más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de DM existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbour o razonamiento basado en casos (CBR); si el problema es de predicción, análisis de regresión, redes neuronales; o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.

Generación del plan de prueba. Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez de este. Por ejemplo, en una tarea supervisada de DM como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

Construcción del Modelo. Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

***Evaluación del modelo*.** En esta tarea, los ingenieros de DM interpretan los modelos de acuerdo con el conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Data Mining aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc.…).

Lo anterior puede resumirse de la siguiente manera, considerando las actividades asociadas a las componentes de la fase correspondiente:

FASE	TAREAS COMPONENTES	ACTIVIDADES ASOCIADAS
Modelado	Seleccionar una técnica de modelado	✓ La técnica modelada ✓ Supuestos del modelo
	Generar el plan de pruebas	✓ Plan de pruebas
	Construir el modelo	✓ Configuración de parámetros ✓ Modelo ✓ Descripción del modelo
	Evaluar el modelo	✓ Evaluar el modelo ✓ Revisión de la configuración de

		parámetros

Puntualmente cada una de esas subfases puede tener una ejecución, como a continuación se muestra:

Selección de la técnica de modelado	
Tarea	Escoger la técnica de modelado
	Como primer paso en modelado, seleccionar la técnica de modelado real que está por ser usado. Aunque usted haya podido seleccionar una herramienta durante la fase de Comprensión del negocio, esta tarea se refiere a la técnica de modelado específico, por ejemplo, un árbol decisión construido con C4.5, o la generación de red neuronales Back-Propagación. Si múltiples técnicas son aplicadas, se realizan esta tarea separadamente para cada técnica.
Salida	Técnicas de modelado
	Documente la técnica de modelado real que está por ser usado. Presunciones del modelado Muchas técnicas de modelado hacen presunciones específicas sobre los datos -por ejemplo, que todos los atributos tengan distribuciones uniformes, no encontrar valores no permitidos, el atributo de clase debe ser simbólico, etc. Registrar cualquiera de tales presunciones hechas.

Generación de la prueba de diseño	
Tarea	Generar la prueba de diseño
	Antes de que nosotros en realidad construyamos un modelo, tenemos que generar un procedimiento o el mecanismo para probar la calidad y validez del modelo. Por ejemplo, en tareas de minería de datos supervisados como la clasificación, esto es común usar tasas de errores como medida de calidad para modelos de minería de datos. Por lo tanto, típicamente separamos el conjunto de datos en una serie y en un conjunto de prueba, construimos el modelo sobre el conjunto de series, y estimamos su calidad sobre el conjunto de prueba separado.
Salida	Prueba de diseño
	Describir el plan intencionado para el entrenamiento, la prueba, y la evaluación de los modelos. Un componente primario del plan determina como dividir un conjunto de datos disponible en datos de entrenamiento, datos de prueba, y conjunto de datos de validación.

Construcción del modelo	
Tarea	Construir el modelo
	Ejecutar la herramienta de modelado sobre el conjunto de datos preparados para crear uno o más modelos
Salida	Parámetro de ajustes

	Con cualquier herramienta de modelado, hay a menudo un gran número de parámetros que pueden ser ajustados. Listar los parámetros y sus valores escogidos, también con el razonamiento para elegir los parámetros de ajustes.
	Modelos
	Estos son los modelos reales producidos por la herramienta de modelado, no un informe.
	Descripciones del modelo
	Describir los modelos obtenidos. Informar sobre la interpretación de los modelos y documentar cualquier dificultad encontrada con sus significados.

Evaluación del modelo	
Tarea	Evaluar el modelo
	El ingeniero de minería de datos interpreta los modelos según su conocimiento de dominio, los criterios de éxitos de minería de datos, y el diseño de prueba deseado. El ingeniero de minería de datos juzga el éxito de la aplicación del modelado y descubre técnicas más técnicamente; él se pone en contacto con analistas de negocio y expertos en el dominio luego para hablar de los resultados de la minería de datos en el contexto de negocio. Por favor note que esta tarea sólo se considera modelos, mientras que la fase de evaluación también toma en cuenta todos los otros resultados que fueron producidos en el curso del proyecto. El ingeniero de minería de datos intenta clasificar los modelos. Él evalúa los modelos según los criterios de evaluación. Tanto como es posible, él también tiene en cuenta objetivos del negocio y criterios de éxito de negocio. En los grandes proyectos de minería de datos, el ingeniero de minería de datos aplica una sola técnica más de una vez, o genera resultados de minería de datos con varias técnicas diferentes. En esta tarea, él también compara todos los resultados según los criterios de evaluación.
Salida	Evaluación de modelos
	Resumir los resultados de esta tarea, listar las calidades de los modelos generados (por ejemplo, en términos de exactitud), y clasificar su calidad en relación con cada otro.
	Parámetros de ajustes revisados
	Según la evaluación del modelo, revise los parámetros de ajuste y témplelos para la siguiente corrida en la tarea de Construcción del Modelo. Repetir la construcción y evaluación del modelo hasta que crea que usted ha encontrado el/los mejor/es modelo/s. Documentar todo como las revisiones y las evaluaciones.

DESARROLLO

El alumno conformara el contenido de la siguiente tabla basándose en la problemática a abordar para el empleo del proyecto de minería de datos, considerando la cuarta fase de la metodología CRISP MD y sus actividades asociadas a las componentes que le corresponde.

FASE 4. FASE DE MODELADO	
COMPONENTES	**SALIDAS**
Seleccionar una técnica de modelado	
Generar el plan de pruebas	
Construir el modelo	
Evaluar el modelo	

CONCLUSIONES

Anote de manera breve las principales conclusiones obtenidas al término de esta práctica

BIBLIOGRAFÍA

Básico:

1. Advances in knowledge discovery. Fayyad usama m. Pearson (1996)
2. Introducción a la minería de datos. Hernández orallo José. Pearson (2004)
3. Inteligencia artificial un enfoque moderno. Stuart Russell. Alfa omega (2004).
4. Inteligencia artificial e ingeniería del conocimiento. Gonzalo pajares Martin Sanz. Alfa omega (2006)
5. Inteligencia artificial. Pedro Ponce cruz. Alfa omega (2010).
6. Inteligencia artificial técnicas, métodos y aplicaciones. José t. Palma Méndez, roque Martin morales. Alfa omega (2008).
7. Essentials of artificial inteligence. Matt Ginsberg. Morgan kaufman (1993)
8. Artificial intelligence. George f. Luger. Pearson (2009).
9. The mechanical mind in history. Philip husbands, Owen Holland. The mit press (2008).
10. Computational intelligence a logical approach. David Poole, Alan Mack worth. Oxford (1998).
11. Data mining. Ian h. Witten, eibe frank. Morgan kaufman (2011).
12. Introduction to data mining. Pang-ning tan, michael steinbach. Pearson (2006).
13. Data mining concepts and techiniques. Jawei han, Michael kamber jian pei. Morgan Kaufman (2011)

Complementario:

1. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.
2. Mitchell, T., *Machine Learning*, McGraw-Hill, 1997.
3. Quinlan, J.R., *C4.5 Programs for Machine Learning*, Morgan-Kaufmann, 1993
4. José Hernández Orallo, M.José Ramírez Quintana, "Introducción a la Minería de datos", Editorial Pearson 2004.
5. Cesar Perez López & Santin González Daniel, "Minería de datos : técnicas y herramientas"Editorial Thomson Paraninfo

PRÁCTICA 8
METODOLOGÍA CRISP-DM: FASE DE EVALUACIÓN

OBJETIVO

- El alumno conocerá las subfases que componen la quinta fase de la metodología CRISP-MD y desarrollara cada una de ellas en la integración de su propuesta de aplicación.

INTRODUCCIÓN

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse, además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Hay que considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Las *matrices de confusión* Edelstein, 1999 son muy empleadas en problemas de clasificación y consisten en una tabla que indica cuantas clasificaciones se han hecho para cada tipo, la diagonal de la tabla representa las clasificaciones correctas. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo.

Las tareas involucradas en esta fase del proceso son las siguientes:

Evaluación de los resultados. En los pasos de evaluación anteriores, se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación con los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo, en un problema real si el tiempo y restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.

Determinación de futuras fases. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de DM.

La figura 1 ilustra las tareas y resultados que se obtienen en esta fase.

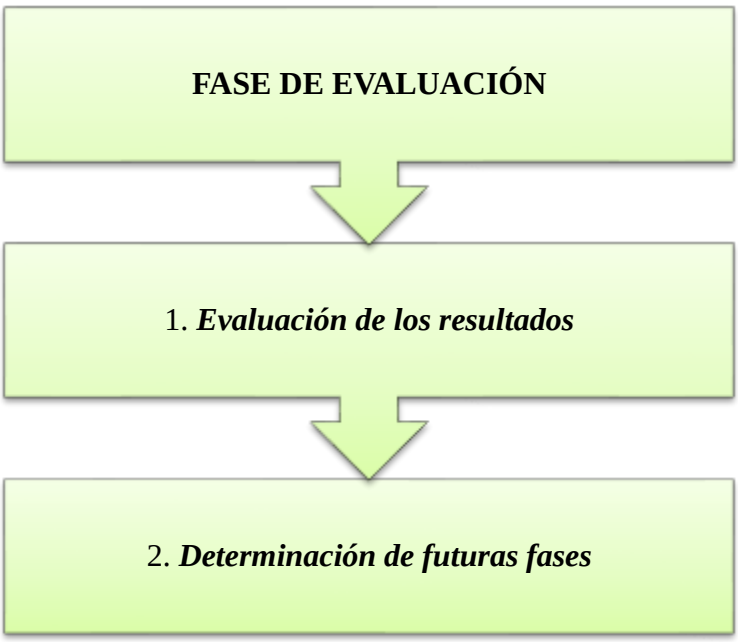


Figura 1. Fase 5 de la metodología CRISP MD

Lo anterior puede resumirse de la siguiente manera, considerando las actividades asociadas a las componentes de la fase correspondiente:

FASE	TAREAS COMPONENTES	ACTIVIDADES ASOCIADAS
Evaluación	Evaluar resultado	✓ Valoración de resultados mineros con respecto con respecto al éxito del negocio ✓ Modelos aprobados
	Revisar	✓ Revisión del proceso
	Determinar próximos pasos	✓ Listar posibles acciones

Puntualmente cada una de esas subfases puede tener una ejecución, como a continuación se muestra:

EVALUACIÓN DE LOS RESULTADOS	
Tarea	**Evaluar los resultados**
	Los pasos de la evaluación anterior trata con factores como la exactitud y la generalidad del modelo. Este paso evalúa el grado al que el modelo responde (encuentra) los objetivos de negocio y procura determinar si hay alguna

	decisión de negocio por el que este modelo es deficiente. Otra opción de evaluación es probar el/los modelo/s sobre aplicaciones de prueba en la aplicación real, si el tiempo y las restricciones de presupuesto lo permiten. Además, la evaluación también verifica otros resultados generados por la minería de datos. Los resultados de la minería de datos implican modelos que necesariamente son relacionados con los objetivos originales de negocio y todas los otros descubrimientos que no son relacionados necesariamente con los objetivos originales de negocio, pero también podría revelar desafíos adicionales, información, o insinuaciones para futuras direcciones.
Salida	**Evaluación de los resultados de la minería de datos en lo que concierne a criterios de éxito de negocio**
	Resumir los resultados de evaluación en términos de criterios de éxito de negocio, incluyendo una declaración final en cuanto si el proyecto ya encuentra los objetivos iniciales de negocio.
	Modelos aprobados
	Después de la evaluación de modelos en lo que concierne a criterios de éxito de negocio, los modelos generados que encuentran los criterios seleccionados son los modelos aprobados.

PROCESO DE REVISIÓN	
Tarea	**Revisar el proceso**
	En este punto, los modelos resultantes pasan a ser satisfactorios y a satisfacer las necesidades de negocio. Ahora es apropiado hacer una revisión más cuidadosa de los compromisos de la minería de datos para determinar si hay cualquier factor importante o tarea que de algún modo ha sido pasada por alto. Esta revisión también cubre cuestiones de calidad -por ejemplo: ¿Construimos correctamente el modelo? ¿Usamos sólo los atributos que nos permitieron usar y que están disponibles para análisis futuros?
Salida	**Revisión de proceso**
	Resumir la revisión de proceso y destacar las actividades que han sido omitidas y/o aquellas que deberían ser repetidas.

Determinación de los próximos pasos	
Tarea	**Determinar los próximos pasos**
	Según los resultados de la evaluación y la revisión de proceso, el equipo de proyecto decide cómo proceder. El equipo decide si hay que terminar este proyecto y tomar medidas sobre el desarrollo si es apropiado, tanto iniciar más iteraciones, o comenzar nuevos proyectos de minería de datos. Esta tarea incluye los análisis de recursos restantes y del presupuesto, que puede influir en las decisiones.
Salida	**Lista de posibles acciones**
	Listar las acciones futuras potenciales, con los motivos a favor y en contra de

	cada opción.
	Decisión
	Describir la decisión en cuanto a cómo proceder, junto con el razonamiento.

DESARROLLO

El alumno conformara el contenido de la siguiente tabla basándose en la problemática a abordar para el empleo del proyecto de minería de datos, considerando la quinta fase de la metodología CRISP MD y sus actividades asociadas a las componentes que le corresponde.

FASE 5. FASE DE EVALUACIÓN	
COMPONENTES	**SALIDAS**
Evaluar resultado	
Revisar	
Determinar próximos pasos	

CONCLUSIONES

Anote de manera breve las principales conclusiones obtenidas al término de esta práctica

BIBLIOGRAFÍA

Básico:

1. Advances in knowledge discovery. Fayyad usama m. Pearson (1996)
2. Introducción a la minería de datos. Hernández orallo José. Pearson (2004)
3. Inteligencia artificial un enfoque moderno. Stuart Russell. Alfa omega (2004).
4. Inteligencia artificial e ingeniería del conocimiento. Gonzalo pajares Martin Sanz. Alfa omega (2006)
5. Inteligencia artificial. Pedro Ponce cruz. Alfa omega (2010).
6. Inteligencia artificial técnicas, métodos y aplicaciones. José t. Palma Méndez, roque Martin morales. Alfa omega (2008).
7. Essentials of artificial inteligence. Matt Ginsberg. Morgan kaufman (1993)
8. Artificial intelligence. George f. Luger. Pearson (2009).
9. The mechanical mind in history. Philip husbands, Owen Holland. The mit press (2008).
10. Computational intelligence a logical approach. David Poole, Alan Mack worth. Oxford (1998).
11. Data mining. Ian h. Witten, eibe frank. Morgan kaufman (2011).
12. Introduction to data mining. Pang-ning tan, michael steinbach. Pearson (2006).
13. Data mining concepts and techiniques. Jawei han, Michael kamber jian pei. Morgan Kaufman (2011)

Complementario:

1. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.
2. Mitchell, T., *Machine Learning*, McGraw-Hill, 1997.
3. Quinlan, J.R., *C4.5 Programs for Machine Learning*, Morgan-Kaufmann, 1993
4. José Hernández Orallo, M.José Ramírez Quintana, “Introducción a la Minería de datos“, Editorial Pearson 2004.
5. Cesar Perez López & Santin González Daniel, “Minería de datos : técnicas y herramientas”Editorial Thomson Paraninfo

PRÁCTICA 9
METODOLOGÍA CRISP-DM: FASE DE IMPLEMENTACIÓN

OBJETIVO

- El alumno conocerá las subfases que componen la sexta fase de la metodología CRISP-MD y desarrollara cada una de ellas en la integración de su propuesta de aplicación.

INTRODUCCIÓN

En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, como por ejemplo, en análisis de riesgo crediticio, detección de fraudes, etc. Generalmente un proyecto de Data Mining no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento.

Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Las tareas que se ejecutan en esta fase son las siguientes:

Plan de implementación. Para implementar el resultado de DM en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación.

Monitorización y Mantenimiento. Si los modelos resultantes del proceso de Data Mining son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

Informe Final. Es la conclusión del proyecto de DM realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto.

Revisión del proyecto*:* En este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar.

La figura 1 ilustra las tareas y resultados que se obtienen en esta fase.

Figura 1. Fase 6 de la metodología CRISP MD

Puntualmente cada una de esas subfases puede tener una ejecución, como a continuación se muestra:

DESARROLLO DEL PLAN	
Tarea	**Desarrollar el plan**
	De acuerdo al desarrollo de los resultados de minería de datos en el negocio, esta tarea toma los resultados de la evaluación y determina una estrategia para el desarrollo. Si un procedimiento general ha sido identificado para crear el/los modelo/s relevante/s, este procedimiento es documentado aquí para el desarrollo posterior.
Salida	**Desarrollo del plan**
	Resumir la estrategia de desarrollo, incluyendo los pasos necesarios y como realizarlos.

PLAN DE SUPERVISIÓN Y MANTENIMIENTO	
Tarea	**Planear la supervisión y el mantenimiento**
	La supervisión y el mantenimiento son cuestiones importantes si los resultados de minería de datos son parte del negocio cotidiano y de su ambiente. La preparación cuidadosa de una estrategia de mantenimiento ayuda evitar largos periodos innecesarios de uso incorrecto de resultados de minería de datos. Para supervisar el desarrollo de los resultados de la minería de datos, el proyecto necesita un plan detallado de proceso de supervisión. Este plan tiene en cuenta el tipo específico de desarrollo.
Salida	**Supervisión y plan de mantenimiento**
	Resumir la estrategia de supervisión y mantenimiento incluyendo los pasos necesarios y como realizarlos

INFORME DEFINITIVO DE PRODUCTO	
Tarea	**Producir el informe final**
	En el final del proyecto, el líder del proyecto y su equipo sobrescribe un informe final. Según el plan de desarrollo, este informe puede ser sólo un resumen del proyecto y sus experiencias (si estas aún no han sido documentadas como una actividad en curso) o esto puede ser una presentación final y comprensiva de los resultados de minería de datos.
Salida	**Informe definitivo**
	Esto es el informe escrito final del compromiso de la minería de datos. Esto incluye todo el desarrollo anterior, el resumen y la organización de los resultados.
	Presentación final
	También a menudo habrá una reunión en la conclusión del proyecto en el que los resultados son presentados verbalmente al cliente.

REVISIÓN DEL PROYECTO	
Tarea	**Revisar el proyecto**
	Evaluar lo que fue correcto y lo que se equivocó, lo que fue bien hecho y lo que necesita para ser mejorado.
Salida	**Documentación de la experiencia**
	Resumir la experiencia importantes ganadas durante el proyecto. Por ejemplo, trampas, accesos engañosos, las insinuaciones para seleccionar las mejores técnicas de miner ía de datos ensituaciones similares podrían ser la parte de esta documentaci ón. En proyectos ideales, ladocumentación de la experiencia también cubre cualquier informe que ha sido escrito por miembros individuales del proyecto durante las fases del proyecto y sus tareas.

DESARROLLO

El alumno conformara el contenido de la siguiente tabla basándose en la problemática a abordar para el empleo del proyecto de minería de datos, considerando la sexta fase de la metodología CRISP MD y sus actividades asociadas a las componentes que le corresponde.

FASE 6. FASE DE IMPLEMENTACIÓN	
COMPONENTES	**SALIDAS**
PLAN DE IMPLEMENTACIÓN.	
PLAN DE SUPERVISIÓN Y MANTENIMIENTO	
INFORME DEFINITIVO DE PRODUCTO	
REVISIÓN DEL PROYECTO	

CONCLUSIONES

Anote de manera breve las principales conclusiones obtenidas al término de esta práctica

BIBLIOGRAFÍA

Básico:

1. Advances in knowledge discovery. Fayyad usama m. Pearson (1996)
2. Introducción a la minería de datos. Hernández orallo José. Pearson (2004)
3. Inteligencia artificial un enfoque moderno. Stuart Russell. Alfa omega (2004).
4. Inteligencia artificial e ingeniería del conocimiento. Gonzalo pajares Martin Sanz. Alfa omega (2006)
5. Inteligencia artificial. Pedro Ponce cruz. Alfa omega (2010).
6. Inteligencia artificial técnicas, métodos y aplicaciones. José t. Palma Méndez, roque Martin morales. Alfa omega (2008).
7. Essentials of artificial inteligence. Matt Ginsberg. Morgan kaufman (1993)
8. Artificial intelligence. George f. Luger. Pearson (2009).
9. The mechanical mind in history. Philip husbands, Owen Holland. The mit press (2008).
10. Computational intelligence a logical approach. David Poole, Alan Mack worth. Oxford (1998).
11. Data mining. Ian h. Witten, eibe frank. Morgan kaufman (2011).
12. Introduction to data mining. Pang-ning tan, michael steinbach. Pearson (2006).
13. Data mining concepts and techiniques. Jawei han, Michael kamber jian pei. Morgan Kaufman (2011)

Complementario:

1. Goldberg, D.E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 1989.
2. Mitchell, T., *Machine Learning*, McGraw-Hill, 1997.
3. Quinlan, J.R., *C4.5 Programs for Machine Learning*, Morgan-Kaufmann, 1993
4. José Hernández Orallo, M.José Ramírez Quintana, “Introducción a la Minería de datos“, Editorial Pearson 2004.
5. Cesar Perez López & Santin González Daniel, “Minería de datos : técnicas y herramientas”Editorial Thomson Paraninfo


ESTADOS UNIDOS MEXICANOS
UNIVERSIDAD AUTÓNOMA