UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

CENTRO UNIVERSITARIO NEZAHUALCÓYOTL

LICENCIATURA EN 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

OCTUBRE 2017

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="3">Programa educativo INGENIERÍA EN SISTEMAS INTELIGENTES</td><td colspan="6">Área de docencia: HERRAMIENTA PARA LOS SISTEMAS INTELIGENTES</td></tr>
<tr><td colspan="2">Aprobación de los HH Consejos Académico y de Gobierno</td><td colspan="2">Fecha: OCTUBRE 2017</td><td colspan="5">Programa elaborado por: Doricela Gutiérrez Cruz, Carmen Liliana Rodríguez Páez</td></tr>
<tr><td colspan="4">Nombre de la unidad de aprendizaje: Minería de Datos I</td><td colspan="5">Fecha de elaboración: Agosto 2017</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 OCTUBRE DE 2017 POR:

ÍNDICE

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

DIRECTORIO

Dr. en Ed. Alfredo Barrera Baca
RECTOR
M. en S. P. María Estela Delgado Maya
SECRETARIA DE DOCENCIA
Dr. en C.I.Amb. Carlos Eduardo Barrera Díaz
SECRETARIO DE INVESTIGACIÓN Y ESTUDIOS AVANZADOS
Dr. en C.S. Luis Raúl Ortiz Ramírez
SECRETARIO DE RECTORÍA
Dr. en A. José Edgar Miranda Ortiz
SECRETARIO DE DIFUSIÓN CULTURAL
M. en Com. Jannet Socorro Valero Vilchis
SECRETARIA DE EXTENSIÓN Y VINCULACIÓN
M. en E. Javier González Martínez
SECRETARIO DE ADMINISTRACIÓN
M. en E.U.R. Héctor Campos Alanís
SECRETARIO DE PLANEACIÓN Y DESARROLLO INSTITUCIONAL
M. en L.A. María del Pilar Ampudia García
SECRETARIA DE COOPERACIÓN INTERNACIONAL
Dra. en C.S. y Pol. Gabriela Fuentes Reyes
ABOGADA GENERAL
Lic. en Com. Gastón Pedraza Muñoz
DIRECTOR GENERAL DE COMUNICACIÓN UNIVERSITARIA
M. en R.I. Jorge Bernaldez García
SECRETARIO TÉCNICO DE LA RECTORÍA
M. en A.P. Guadalupe Santamaría González
DIRECTORA GENERAL DE CENTROS UNIVERSITARIOS UAEM Y UAP
M. en A. Ignacio Gutiérrez Padilla
CONTRALOR UNIVERSITARIO

CENTRO UNIVERSITARIO NEZAHUALCÓYOTL

DIRECTORIO

M. en I. S. C. Cuauhtémoc Hidalgo Cortés
Encargado del despacho de C.U. Nezahualcóyotl

Dr. en E. Darío Guadalupe Ibarra Zavala
Subdirector Académico

Lic. en E. Ramón Vital Hernández
Subdirector Administrativo

Dra. en C. S. María Luisa Quintero Soto
Coordinadora de Investigación y Estudios Avanzados

Lic. en A. E. Víctor Manuel Durán López
Coordinador de Planeación y Desarrollo Institucional

Dr. en R.I. Rafael Alberto Duran Gómez
Coordinador de la Licenciatura en Comercio Internacional

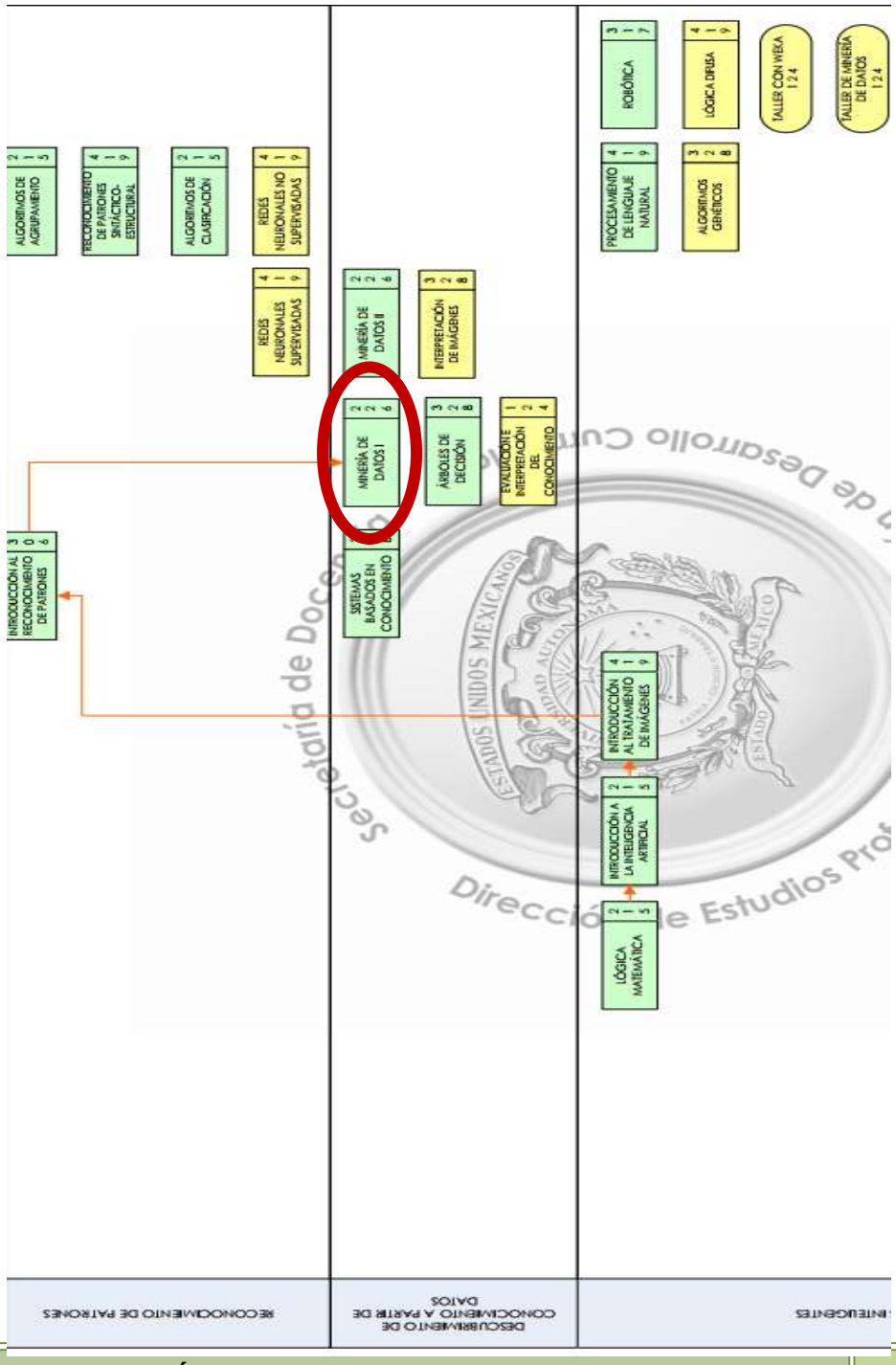
Dra. Silvia Padilla Loredo
Coordinadora de la Licenciatura en Educación para la Salud

Dra. Ricardo Rico Molina
Coordinador de la Licenciatura en Ingeniería en Sistemas Inteligentes

D. En U. Noé Gaspar Sánchez
Coordinador de la Licenciatura en Ingeniería en Transporte

M. En CC Erick Nicolás Cabrera Álvarez
Coordinador de Licenciatura en Seguridad Ciudadana

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 Minería de Datos
PRÁCTICA 2	• KDD Proceso de Extracción de Conocimiento.
PRÁCTICA 3	• Selección de Datos e Información.
PRÁCTICA 4	• Limpieza de Datos.
PRÁCTICA 5	• Integración de Datos.


PRÁCTICA 6
• Transformación de Datos.
PRÁCTICA 7
• Reducción de Datos.
PRÁCTICA 8
• Minería de Datos.
PRÁCTICA 9
• Interpretación y Evaluación de los Patrones.
PRÁCTICA 10
• Interpretación del resultado

PRÁCTICA 1
INTRODUCCIÓN A LA MINERÍA DE DATOS

Objetivo

El alumno conocerá el concepto y la importancia de la minería de datos.

Introducción

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.

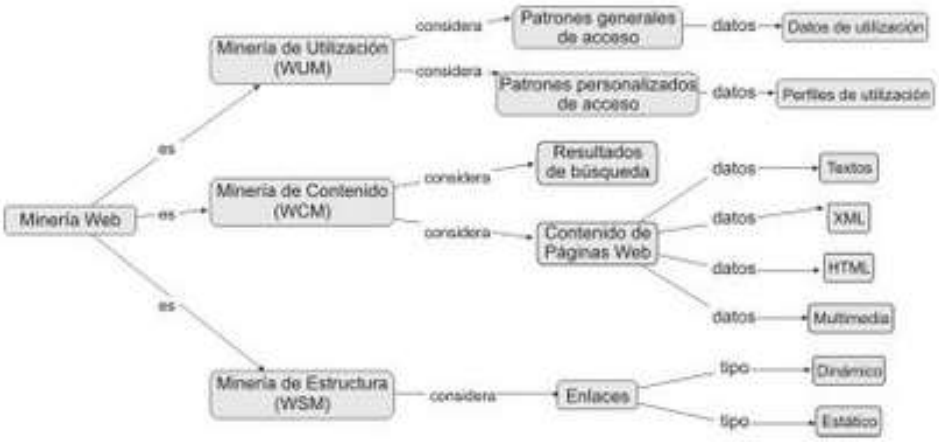
Cabe resaltar que en la metodología de MD forma parte de un proceso denominado descubrimiento de conocimiento en bases de datos «Knowledge Discovery in Databases» (KDD), que indica los pasos necesarios para reducir riesgos en la búsqueda de modelos de

conocimiento al aplicar técnicas de MD. Por ejemplo, los datos requieren un sustancial pre procesamiento para ser modelados (limpieza y preparación de datos) en el proceso KDD.

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 delos distintos autores.

1. Velarde Martínez, Apolinar, Minería de Datos. Una Introducción Conciencia Tecnológica [en línea] 2003, ( ) : Disponible en:**<http://www.redalyc.org/articulo.oa?id=94402303>** ISSN 1405-5597
2. Riquelme, José C., Ruiz, Roberto, Gilbert, Karina, Minería de Datos: Conceptos y Tendencias Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial [en línea] 2006, 10 (primavera) : Disponible en:**<http://www.redalyc.org/articulo.oa?id=92502902>** ISSN 1137-3601
3. Estrada-Danell, Rafael Isaac, Zamarripa-Franco, Roman Alberto, Zúñiga-Garay, Pilar Giselle, Martínez-Trejo, Isaías, Aportaciones desde la minería de datos al proceso de captación de matrícula en instituciones de educación superior particularesRevista Electrónica Educare [en linea] 2016, 20 (Septiembre-Diciembre) : Disponible en:**<http://www.redalyc.org/articulo.oa?id=194146862011>** ISSN



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).

PRÁCTICA 2

PROCESO KDD (*KNOWLEDGE DISCOVERY IN DATABASE)*

Objetivo

El alumno conocerá el proceso del KDD y las fases que lo componen.

Introducción

El proceso KDD se puede definir como "el proceso no trivial de identificar patrones válidos, novedosos y potencialmente útiles y en última instancia, comprensible a partir de los datos". Este proceso también es conocido por diferentes nombres que podrían ser sinónimos del mismo, entre los cuales se encuentran Data Archeology, Dependency Function Analysis, Information Recollect, Pattern Data Analysis o KnowledgeFishing.

También supone la convergencia de distintas disciplinas de investigación; podemos nombrar algunas tales como el aprendizaje automático, estadística, inteligencia artificial, sistemas de gestión de base de datos, técnicas de visualización de datos, los sistemas para el apoyo a la toma de decisión (DSS) o la recuperación de información, entre otras.

Uno de los procesos más importantes dentro del KDD es el usuario, ya que es él quien determina el dominio de la aplicación o sea, decide cómo y qué datos se utilizarán en el proceso. Por lo tanto, los pasos en el proceso global no están claramente diferenciados por ser un proceso iterativo e interactivo con el usuario experto. Las interacciones entre las Decisiones tomadas en diferentes pasos, así como los parámetros de los métodos utilizados Y la forma de representar el problema suelen ser extremadamente complejos.

PROCESO KDD CENTRADO EN EL USUARIO

El KDD es un proceso centrado en el usuario, que tiene la propiedad de ser altamente interactivo, y que debe ser guiado por las decisiones que toma el usuario, o también por un agente inteligente. La naturaleza centrada en el usuario del proceso KDD posee varias cuestiones actualmente en investigación. Una de ella es, como asistir al usuario en la correcta selección de herramientas y técnicas apropiadas, para lograr los objetivos del usuario. Es un desafío real, dar al sistema, la inteligencia necesaria para obtener

conocimiento e impartir el mismo, en el momento de decidir las herramientas apropiadas, para que tipo de problemas y cuando. Particularmente en KDD esto es un problema importante de abarcar, aún si el usuario es un investigador que desarrollo técnicas específicas, ya que se necesita al sistema completo para resolver un problema. Mientras que hay desarrollos de Knowledge discovery workbenches y sistemas integrados que incluyen más de un paso del proceso KDD, ellos no se involucran demasiado en las cuestiones de un sistema amigable para el usuario, para ser utilizado por un analista. Un analista, no es usualmente un experto en KDD, pero sí alguien que tiene la responsabilidad de sacar el significado de los datos usando técnicas de KDD disponibles. Para que un sistema cualquiera de KDD sea exitoso, necesita integrarse bien dentro de un ambiente existente para proveer una completa solución a un analista. Por lo tanto, un reto para los investigadores y practicantes del KDD es poner más énfasis en el proceso general de KDD y en las herramientas para soportar sus varios pasos. Se le debería prestar mayor atención a la interacción con el humano y menos a la automatización total, con el fin de soportar tanto a expertos como a usuarios novatos. El desarrollo de herramientas apropiadas de visualización, interpretación y análisis de descubrimiento de patrones son de particular importancia. Tales ambientes interactivos, a través de la reducción del tiempo para comprender la complejidad de los datos, habilitan la posibilidad de obtener soluciones prácticas a muchos de los problemas de la vida real, de una manera más rápida que lo hace el ser humano o una computadora operando independientemente.

El proceso de KDD (véase Figura 1) es interactivo e iterativo por naturaleza, e involucra una serie de pasos, en los que se incluyen decisiones tomadas por el usuario. En general, involucrando más, o menos pasos, la estructura general sigue la siguiente forma:

Entender el dominio de aplicación, cuál es el problema a resolver, y cuáles son los objetivos. 2) Seleccionar del conjunto de datos originales, un subconjunto apropiado, para el problema que deseamos resolver. Eliminando por ejemplo variables irrelevantes. 3) En la etapa de limpieza y pre procesamiento se deberían tomar decisiones con respecto a valores faltantes, atípicos, erróneos, etc. (ruido). También se podría necesitar normalizar los valores de las variables o llevar a cabo otras tareas similares. La etapa de preparación y limpieza es a veces una etapa descuidada, pero de suma importancia en este proceso, dado que grandes cantidades de datos son recolectados por medio de métodos automáticos (ej. vía web). A veces el método por el cual los datos fueron obtenidos no fue cuidadosamente controlado, y así los datos podrían contener valores fuera de rango (ej. edades negativas), combinaciones incorrectas de datos (Sexo: masculino; embarazada: si), y otros. Ingresar estos datos erróneos a los algoritmos de data mining solo lleva a entorpecer su proceso de aprendizaje,

o a conseguir resultados alejados del comportamiento real. 4) Encontrar características útiles para representar a los datos dependiendo de los objetivos. Reducción de dimensiones (ej. Kohonen) para llevar adelante el trabajo con un número reducido de variables. Eliminar columnas que varían juntas, como por ejemplo fecha de nacimiento y edad, simplemente tabular, aggregation (calcular estadísticos descriptivos), o técnicas más sofisticadas como clustering o análisis de componentes principales. 5) Elegir las herramientas de data mining adecuadas al problema a resolver, teniendo en cuenta el objetivo (predecir, explicar, clasificar, agrupar, etc). También se debe en esta etapa, establecer los parámetros de las redes utilizadas (arquitectura de la red, datos de entrenamiento, de validación y de testeo, etc.). Una vez realizada la tarea, se procede con el descubrimiento de patrones y relaciones en los datos, para presentárselos al usuario de una manera adecuada (gráficos, árboles, reglas, etc.) 6) Interpretación de los datos, llevada a cabo por el analista. 7) En esta etapa se debería consolidar el conocimiento ganado, probando los modelos creados contra los resultados obtenidos de la aplicación de estos modelos en el mundo real. Como dijimos anteriormente existen herramientas que fueron desarrolladas para caber en una de éstas metodologías, pero ninguna puso aún, real énfasis en el propósito de asistir al usuario durante el seguimiento de estos pasos en busca del conocimiento. Por tal motivo el sistema que desarrollaremos buscará adaptarse a la metodología general anteriormente mostrada, pero incorporando algunas características importantes como lo son: 9 La posibilidad de realizar un ciclo entre cualesquiera dos pasos, ya que a veces el conocimiento descubierto puede ser directamente aplicable, y otras veces puede guiar al refinamiento de los objetivos de la minería. No hay un progreso determinístico asumido desde un paso a otro. También, los pasos interpretativos y evaluativos, pueden involucrar retrocesos a cualquiera de los pasos anteriores, cualquier número de veces. La incorporación de un agente inteligente, encargado de monitorear las actividades del usuario y brindarle a esta asistencia a lo largo de todo el proceso de KDD.

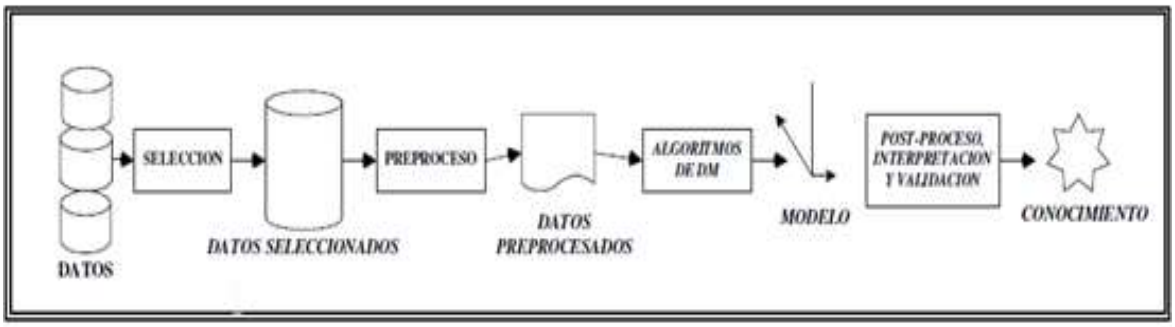


Figura 1. Proceso KDD.

Desarrollo

Conteste las siguientes preguntas y envíelas a su portafolio de SEDUCA

1) ¿Es la extracción de conocimientos desde bases de datos?
2) ¿Búsqueda de patrones interesantes y de regularidades importantes en grandes bases de datos?
3) ¿Proceso que consta de una serie de fases, mientras que la minería de datos es solo una de estas fases?
4) ¿Cuáles son las fases del KDD o Proceso de Extracción del Conocimiento?

5) ¿Cuáles son las tareas de la minería de datos?
6) ¿Cuáles son las técnicas de la minería de datos?

7) Explique cada una de las etapas del proceso KDD

- Selección de Datos e Información
- Limpieza de Datos
- Integración de Datos
- Transformación de Datos
- Reducción de datos
- Data Mining.
- Interpretación y Evaluación de los Patrones
- Interpretación del resultado

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).***

PRÁCTICA 3

Selección de Datos e Información

Objetivo

El alumno conocerá los distintos repositorios donde pueda obtener información.

Introducción

Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original. Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentran en bases de datos y otras fuentes muy diversas, tanto internas como externas. Muchas de estas fuentes son las que se utilizan para el trabajo transaccional; parte de información interna de la organización, los almacenes de datos pueden recoger información externa:

• Demografías (censo), páginas amarillas, pictografías (perfiles por zonas), uso de Internet, información de otras organizaciones.
• Datos compartidos en una industria o área de negocio, organizaciones y colegios profesionales, catálogos, etc.
• Datos resumidos de áreas geográficas, distribución de la competencia, evolución de la economía, información de calendarios y climatológicas, programaciones televisivas deportivas, catástrofes,..
• Bases de datos externas compradas a otras compañías.

Debido a la heterogeneidad de los datos disponibles en la actualidad, las posibles fuentes de datos a considerar:

- Datos de Marketing

Este tipo de datos se refiere a aquellos datos almacenados en bases de datos, data marts o data warehouses

- Datos de Servidor

Estos datos corresponden a los generados por la interacción de las personas o usuarios que acceden a un servidor, por ejemplo, navegar un sitio web en particular o utilizan un servicio donde queda registro de su actividad (por ejemplo: Facebook, twitter, etc.).

- Metadatos

La última fuente de datos es acerca del contenido del sitio mismo. Usualmente esta información es generada dinámica y automáticamente luego de una actualización del contenido o estructura del sitio. Los metadatos de un sitio proporcionan información topológica acerca del mismo. (Por ejemplo: páginas vecinas, enlaces entre páginas).

Figura 1. Fuentes de Datos

El análisis posterior será mucho más sencillo si la fuente es unificada y accesible (interna).

Desarrollo

Conteste las siguientes preguntas y envíelas a su portafolio de SEDUCA

1) ¿Son los datos adecuados para describir el o los fenómenos que el análisis está intentando explicar?
2) ¿Hay un campo común entre los datos que serán utilizados y otros datos de otros repositorios de datos?

3) ¿Pueden ser consolidados todos los datos en un repositorio de datos (base datos, data mart o data warehouse)?
4) ¿Qué información interna y/o externa está disponible para el análisis?
5) ¿Hay alguna información redundante en los datos?
6) ¿Existen datos demográficos disponibles?

7) Una vez que se ha definido el objetivo, se deben seleccionar los datos y la información. Buscar un repositorio de datos para así continuar con el siguiente paso Limpieza 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. Büchner, A.; Mulvenna, M.; Norwood, M. (2000). *Data-Driven Marketing*. EM-ElectronicMarkets, Vol. 8, No. 3, 07/98
2. Carlos A. I. Fernández – Fundamentos de los Agentes Inteligentes – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid.
3. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0. Step-by-step data mining guide.
4. Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206.
5. Molina Félix, Luis Carlos. Data Mining: Torturando los datos hasta que confiesen. Coordinador del programa de Data mining (UOC). 2002
6. SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System

PRÁCTICA 4

Limpieza de Datos

Objetivo

El alumno conocerá como limpiar datos de un repositorio.

Introducción

Limpieza de datos: Se aumenta la calidad de los datos al nivel requerido mediante técnicas de análisis selectivo. Este proceso consiste en la eliminación de datos erróneos o inconsistentes

Detección de datos anómalos

- Selección de datos (muestreo, ya sea verticalmente, eliminando atributos, u horizontalmente, eliminando tuplas).
- Redefinición de atributos (agrupación o separación)
- Los datos sucios en algunos casos deben ser eliminados ya que pueden contribuir a un análisis inexacto y resultados incorrectos

Se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera de rango, obteniéndose al final una estructura de datos adecuada para su posterior transformación.

Desarrollo

A continuación en la Figura 1 y 2 se muestra un ejemplo de limpieza donde los datos proporcionados por la unidad de cuidados paliativos, fueron un universo total de 3365 casos, con el objetivo de descubrir patrones de calidad esto es: Sin valores nulos o anómalos.

	A	B	C	D	E	F	G
1		agosto	92	ca.cvcr	2009	ca.co	
2	f	julio	52	ca.co	2009	ca.ce	83
3	f	junio	55	ca.ce	2009	ca.ab	5
4	f	mayo	61	ca.ceu	2009	ca.ceu	91
5	f	octubre	79	ca.c.v	2009	ca.a	2
6	f	septiembre	56	ca.meo	2009	ca.c.v	228
7	f	mayo	75	ca.pu	2009	ca.cvcr	293
8	f	julio	69	ca.ce	2009	ca.es	81
9	m	noviembre	69	sa.te.b	2009	ca.e	65
10	f	enero	61	ca.c.v	2009	ca.h	164
11	f	agosto	74	ca.pu	2009	ca.av	2
12	f	diciembre	57	ca.meo	2009	ca.meo	229
13	m	enero	74	ca.p	2009	ca.ov	146
14	m	mayo	74	sa.te.b	2009	ca.pi	59
15	m	diciembre	52	ca.meo	2009	ca.p	283
16	m	febrero	66	ca.meo	2009	ca.pu	280
17	f	agosto	62	ca.m	2009	ca.r	151
18	m	agosto	54	ca.ce	2009	ca.ri	160
19	f	junio	50	ca.meo	2009	ca.te	26
20	m	junio	40	ca.c.v	2009	ca.v	48
21	m	junio	62	ca.meo	2009	ca.sl	8
22	f	diciembre	52	ca.meo	2009	ca.m	626

Figura 1

23	f	febrero	77	ca.ce	2009	ca.o
24	f	agosto	49	ca.meo	2009	ca.os
25	m	marzo	60	ca.meo	2009	ca.pa
26	m	enero	79	ca.pu	2009	le
27	m	septiembre	63	ca.meo	2009	ca.mm
28	m	agosto	52	ca.meo	2009	sa.te.b
29	f	enero	58	ca.es	2009	suma
30	m	diciembre	60	ca.ov	2009	
31	f	febrero	67	ca.meo	2009	derecho
32	f	julio	61	ca.meo	2009	0
33	m	diciembre	60	ca.meo	2009	1
34	f	julio	57	ca.meo	2009	2
35	f	enero	57	ca.meo	2009	3
36	m	diciembre	41	ca.meo	2009	4
37	f	agosto	63	ca.meo	2009	5
38	m	agosto	48	ca.pu	2009	6
39	m	marzo	49	ca.meo	2009	7
40	m	diciembre	81	ca.m	2009	8
41	f	marzo	70	sa.te.b	2009	9
42	m	octubre	58	ca.pi	2009	
43	f	julio	70	ca.pu	2009	
44	f	mayo	60	ca.ce	2009	

Figura 2

La BD contiene datos anómalos y fuera de las variables que se pretenden abordar, la fase de la limpieza es eliminar cada uno de ellos mediante Excel, para que la inconsistencia se reduzca notoriamente. Los datos contienen variables representativas que abordan la problemática: *Sexo, año, edad y diagnóstico.* Que son elementales para la siguiente fase, Integración de Datos.

1) Con la BD que el alumno obtuvo del repositorio de datos, deberá realizar la limpieza y eliminar datos inconsistentes.

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

Biliografía:

1) W. Kim, B. Choi, E. Hong, S. Kim, and D. Lee. A taxonomy of dirty data. Data Mining and Knowledge Discovery, 7:8199, 2003.
2) M. Castejón, J. B. Ordieres, F. J. Martínez, and E. P. Vergara. Outlier detection and data cleaning in multivariate non-normal samples: The PAELLA algorithm. Data Mining and Knowledge Discovery, (9):171 187, 2004.
3) Carlos A. I. Fernández – Fundamentos de los Agentes Inteligentes – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid.
4) Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0. Step-by-step data mining guide.
5) Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206.
6) Molina Félix, Luis Carlos. Data Mining: Torturando los datos hasta que confiesen. Coordinador del programa de Data mining (UOC). 2002
7) SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System

PRÁCTICA 5

Integración de Datos

Objetivo

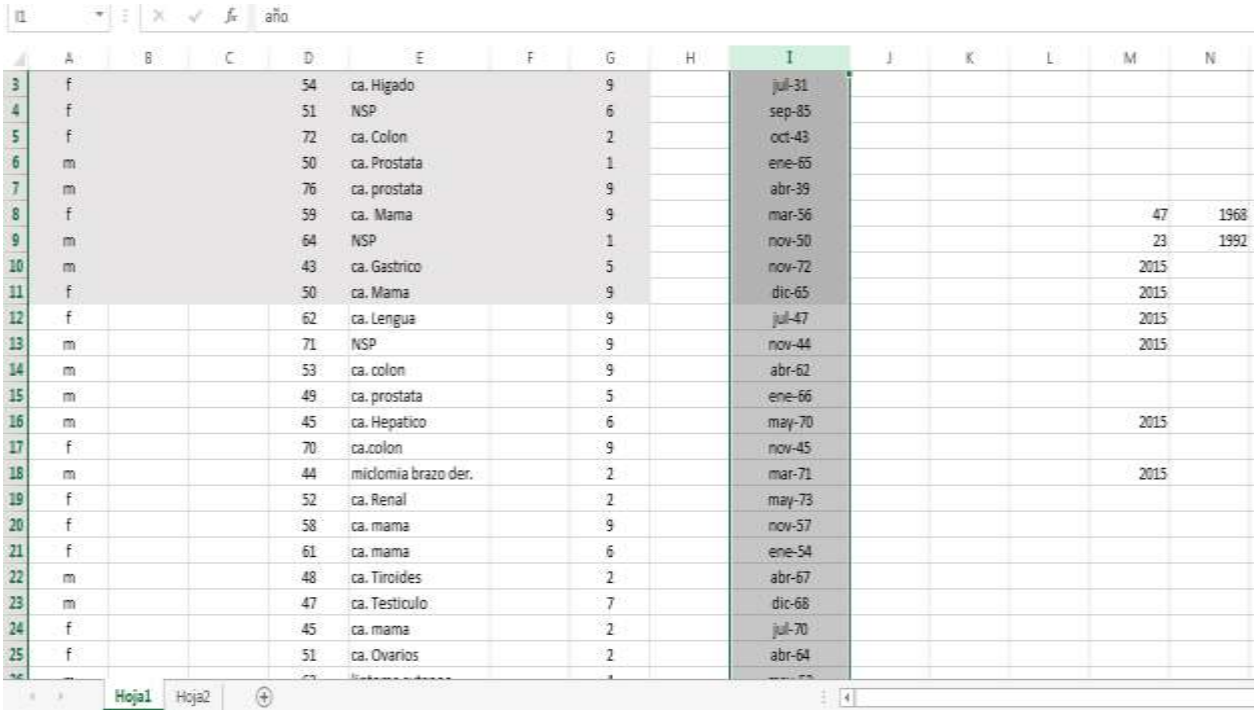
El alumno conocerá la fase de integración de Datos.

Introducción

Se basa en combinar múltiples tablas o registros para crear nuevos registros o valores. El combinar tablas hace referencia a unir dos o más tablas que presentan diferente información sobre los mismos objetos. La combinación de datos también incluye la agregación. La agregación consiste en operaciones donde se obtienen nuevos valores mediante la unión de información de varios registros o tablas. Esta tarea comprende así mismo operaciones relativas a construcción de datos tales como la producción de atributos derivados, nuevas muestras completas, o transformaciones de los valores de atributos ya existentes. Los atributos derivados se pueden construir con uno o más atributos presentes en el mismo patrón. En el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización, consolidando los datos de una forma necesaria para la fase siguiente. Combina datos de múltiples procedencias incluyendo múltiples bases de datos, que podrían tener diferentes contenidos y formatos.

Desarrollo

En la Figura que se muestra a continuación, cada dato se agrupo con su respectiva variable y se obtuvo un nuevo valor, que sería la cantidad de pacientes por *Diagnóstico.*



	A	D	E	G	I	M	N
3	f	54	ca. Higado	9	jul-31		
4	f	51	NSP	6	sep-85		
5	f	72	ca. Colon	2	oct-43		
6	m	50	ca. Prostata	1	ene-65		
7	m	76	ca. prostata	9	abr-39		
8	f	59	ca. Mama	9	mar-56	47	1968
9	m	64	NSP	1	nov-50	23	1992
10	m	43	ca. Gastrico	5	nov-72	2015	
11	f	50	ca. Mama	9	dic-65	2015	
12	f	62	ca. Lengua	9	jul-47	2015	
13	m	71	NSP	9	nov-44	2015	
14	m	53	ca. colon	9	abr-62		
15	m	49	ca. prostata	5	ene-66		
16	m	45	ca. Hepatico	6	may-70	2015	
17	f	70	ca.colon	9	nov-45		
18	m	44	miclomia brazo der.	2	mar-71	2015	
19	f	52	ca. Renal	2	may-73		
20	f	58	ca. mama	9	nov-57		
21	f	61	ca. mama	6	ene-54		
22	m	48	ca. Tiroides	2	abr-67		
23	m	47	ca. Testiculo	7	dic-68		
24	f	45	ca. mama	2	jul-70		
25	f	51	ca. Ovarios	2	abr-64		

1) El alumno tiene que realizar la fase de integración y crear si es posible sus nuevas variables.

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. Carlos A. I. Fernández – Fundamentos de los Agentes Inteligentes – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid.
2. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0. Step-by-step data mining guide.
3. Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206.
4. Molina Félix, Luis Carlos. Data Mining: Torturando los datos hasta que confiesen. Coordinador del programa de Data mining (UOC). 2002
5. SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System

PRÁCTICA 6

Transformación de Datos

Objetivo

El alumno conocerá la fase de transformación de datos

Introducción

Consisten principalmente en modificaciones sintácticas llevadas a cabo sobre datos sin que supongan un cambio para la técnica de minería aplicada. Las transformaciones discretas de los datos tienen la ventaja de que mejoran la comprensión de las reglas descubiertas al transformar los datos de bajo nivel en datos de alto nivel y también reduce significativamente el tiempo de ejecución del algoritmo de búsqueda. Su principal desventaja es que se puede reducir la exactitud del conocimiento descubierto, debido a que puede causar la perdida de alguna información. Existen diferentes métodos de transformación de variables continuas a discretas que se pueden agrupar según distintas aproximaciones: métodos locales (realizan la transformación discreta en una región del espacio de las instancias, por ejemplo, utilizando un subconjunto de las instancias), métodos globales (utilizan el espacio de las instancias), métodos supervisados (utilizan la información de la clave (valor del atributo objetivo).

Consisten principalmente en modificaciones sintácticas llevadas a cabo sobre los datos, sin que supongan un cambio en el significado de los mismos. Estas transformaciones pueden ser necesarias para la técnica de MD aplicada. Las estrategias anteriormente descritas no son mutuamente excluyentes. Existen técnicas de pre procesado que podrían seguir dos o más de las vías indicadas y habría que clasificarlas como una combinación de ambas (por ejemplo, la compactación de datos, que reduce e integra).

Desarrollo

La búsqueda y descubrimiento de patrones en esta etapa que conlleva la de minería de datos y en donde son aplicadas tareas de descubrimiento. Este proceso de transformación de datos solo se realizó para los relacionados con el diagnostico, ya que estos eran oncológicos o no oncológicos.

1	sexo	edad	dx	dh	ano
2	m	81-98	cacvcr	2	2009
3	f	41-60	caco	2	2009
4	f	41-60	cace	2	2009
5	f	61-80	caceu	2	2009
6	f	61-80	cacv	9	2009
7	f	41-60	cameo	2	2009
8	f	61-80	capu	2	2009
9	f	61-80	cace	2	2009
10	m	61-80	sateb	1	2009
11	f	61-80	cacv	0	2009
12	f	61-80	capu	5	2009
13	f	41-60	cameo	1	2009
14	m	61-80	cap	1	2009
15	m	61-80	sateb	9	2009
16	m	41-60	cameo	2	2009
17	m	61-80	cameo	3	2009
18	f	61-80	cam	2	2009
19	m	41-60	cace	9	2009
20	f	41-60	cameo	2	2009
21	m	21-40	cacv	6	2009
22	m	61-80	cameo	2	2009
23	f	41-60	cameo	6	2009

24	f	61-80	cace	2	2009
25	f	41-60	cameo	2	2009
26	m	41-60	cameo	1	2009
27	m	61-80	capu	9	2009
28	m	61-80	cameo	1	2009
29	m	41-60	cameo	1	2009
30	f	41-60	caes	0	2009
31	f	41-60	caov	9	2009
32	f	61-80	cameo	2	2009
33	f	61-80	cameo	2	2009
34	m	41-60	cameo	2	2009
35	f	41-60	cameo	3	2009
36	f	41-60	cameo	2	2009
37	m	41-60	cameo	0	2009
38	f	61-80	cameo	9	2009
39	m	41-60	capu	1	2009
40	m	41-60	cameo	4	2009
41	m	81-98	cam	9	2009
42	f	61-80	sateb	2	2009
43	m	41-60	capi	9	2009
44	f	61-80	capu	9	2009
45	f	41-60	cace	2	2009
46	f	41-60	cameo	2	2009

1) El alumno deberá encontrar las modificaciones y transformaciones posibles en su repositorio 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. T.Y. Lin. Attribute transformation for data mining I: Theoretical explorations. International Journal of Intelligent Systems, 17:213222, 2002.
2. Carlos A. I. Fernández – Fundamentos de los Agentes Inteligentes – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid.
3. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0. Step-by-step data mining guide.
4. Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206.
5. Molina Félix, Luis Carlos. Data Mining: Torturando los datos hasta que confiesen. Coordinador del programa de Data mining (UOC). 2002
6. SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System

PRÁCTICA 7

Reducción de Datos

Objetivo

El alumno conocerá la reducción de datos.

Introducción

Reducir el tamaño de los datos, encontrando las características más significativas dependiendo del objetivo del proceso. Se pueden utilizar métodos de transformación para reducir el número efectivo de variables a ser consideradas, o para encontrar otras representaciones de los datos.

- Reducción de dimensiones (la extracción irrelevante y débil de atributo), compresión de datos (reemplazando valores de datos con datos alternativos codificados), reducción de tamaño (reemplazando valores de datos con representación alternativa más pequeña), una generalización de datos (reemplazando valores de datos de niveles conceptuales bajos con niveles conceptuales más altos).

Reducción de datos: Consiste en decidir qué datos deben ser utilizados para el análisis. El criterio que se sigue incluye la relevancia con respecto a los objetivos que se persiguen en la MD, y limitaciones técnicas tales como pueden ser volúmenes máximos de datos o bien tipos de datos concretos. Nos centraremos en este caso en esta perspectiva del pre procesamiento: reducir el volumen de datos seleccionando los más relevantes para su posterior uso por algoritmos de MD.

Desarrollo

La descripción de atributos más relevantes se conforma en la tabla 1, mientras que la tabla 2 representa los atributos de los distintos tipos de cáncer que se abordan en la problemática, así la reducción de los datos y sus dimensiones hacen más óptima la comprensión de los objetivos.

Atributo	Definición	Descripción
S	Sexo	Género de los pacientes
A	Año	Año de en qué se reporta en la unidad
Dx	Cáncer	Es el tipo de cáncer que padeció el paciente

Tabla 1. Descripción de los atributos más relevantes

Atributo	Definición
ca co	cáncer de corazón
ca ce	cáncer cerebral
ca ab	cáncer abdominal
ca ceu	cáncer cervicouterino
ca a	cáncer de apéndice
ca c v	cáncer de columna vertebral
ca cvcr	cáncer de columna vertebral y cáncer de recto
ca es	cáncer de esófago
ca e	cáncer de estómago
ca h	cáncer de hígado
ca av	cáncer de la ampolla de váter
ca meo	cáncer de médula ósea

ca ov	cáncer de ovario
ca pi	cáncer de piel
ca p	cáncer de próstata
ca pu	cáncer de pulmón
ca r	cáncer de recto
ca ri	cáncer de riñón
ca te	cáncer de testículos
ca v	cáncer de vejiga
ca sl	cáncer del sistema linfático
ca m	cáncer mama
ca o	cáncer oral
ca os	cáncer óseo
ca pa	cáncer páncreas
le	Leucemia
ca mm	Mieloma múltiple
sa te b	sarcoma de tejido blando

Tabla 2. Descripción de los atributos asociados a los tipos de cáncer

1) El alumno deberá reducir los datos posibles en su repositorio.

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. Carlos A. I. Fernández – Fundamentos de los Agentes Inteligentes – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid.
2. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0. Step-by-step data mining guide.
3. Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206.
4. Molina Félix, Luis Carlos. Data Mining: Torturando los datos hasta que confiesen. Coordinador del programa de Data mining (UOC). 2002
5. SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System

PRÁCTICA 8

Data Mining

Objetivo

El alumno realizará el análisis y construcción del modelo de minería de datos.

Introducción

Consiste en la búsqueda de los patrones de interés que pueden expresarse como un modelo o simplemente que expresen dependencia de los datos.
Se tiene que especificar un criterio de preferencia para seleccionar un modelo de un conjunto de posibles modelos. También se tiene que especificar la estrategia de búsqueda a utilizar (normalmente está determinado en el algoritmo de minería).

La MD en 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, 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

Desarrollo

Para realizar esta actividad se utilizó el software WEKA. Esto se debe a que los datos incluidos en el conjunto de datos son en su mayoría categóricos, se seleccionó por utilizar árboles de decisión. De los algoritmos disponibles, se utilizó el J48 correspondiente al algoritmo C4.5 En su ejecución se utilizaron las especificaciones que por default tiene WEKA, así como el método de validación cruzada estratificada. En el

análisis de la figura se observa que por genero los pacientes que fueron atendidos en su mayoría por algún tipo de cáncer para el periodo reportado fueron las mujeres en un 91.8%.

```
Number of Leaves  :     138

Size of the tree :      235


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2932               87.1322 %
Incorrectly Classified Instances       433               12.8678 %
Kappa statistic                          0.7269
Mean absolute error                      0.118
Root mean squared error                  0.269
Relative absolute error                 37.0597 %
Root relative squared error             67.4235 %
Total Number of Instances             3365

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.802     0.084      0.86      0.802     0.83       0.896    m
                 0.918     0.198      0.878     0.918     0.897      0.899    f
                 0         0          0         0         0          0.723    m
Weighted Avg.    0.871     0.153      0.87      0.871     0.87       0.898
```

1) El alumno deberá realizar la Minería de Datos a su repositorio, concluyendo con las fases anteriores.

Conclusiones

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

Bibliografía:

1. Riquelme, José C, Ruiz Roberto, Gilbert, Karina : Minería de Datos: Conceptos y Tendencias Departamento de Lenguajes y Sistemas Informáticos Universidad de Sevilla (2006)
2. 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)
3. 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)
4. Carlos A. I. Fernández – Fundamentos de los Agentes Inteligentes – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid.
5. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0. Step-by-step data mining guide.
6. Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206.
7. Molina Félix, Luis Carlos. Data Mining: Torturando los datos hasta que confiesen. Coordinador del programa de Data mining (UOC). 2002
8. SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System

PRÁCTICA 9

Evaluación e Interpretación de los Patrones

Objetivo

El alumno deberá identificar los nuevos patrones y producir varias hipótesis.

Introducción

Se identifican verdaderamente patrones interesantes que representan conocimiento usando diferentes técnicas incluyendo análisis estadísticos y lenguajes de consultas.
La interpretación de los mejores modelos (visualización, simplicidad, posibilidad de integr ación, ventajas colaterales) ayudará a la selección del modelo(s) final
La fase de MD puede producir varias hipótesis de modelos.Será necesario establecer qué m odelos son los más válidos (técnicas habituales son el uso de conjuntos de test indepen dientes) .

Desarrollo

En el análisis de la Figura se observa que los tipos de cáncer de mayor incidencia para el periodo reportado fueron: cáncer de próstata con 71.4%, cáncer de mama 68.7%, cáncer de páncreas 68.6%, cáncer de esófago 65.7%, en tanto que el de menor incidencia fue el cáncer de ojo con 9.1%. En Clasificación de tipo de cáncer reportados durante el período 2009-2015

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1623               48.2318 %
Incorrectly Classified Instances      1742               51.7682 %
Kappa statistic                          0.4331
Mean absolute error                      0.0292
Root mean squared error                  0.1326
Relative absolute error                 66.5038 %
Root relative squared error             89.5113 %
Total Number of Instances             3365

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.224     0.078      0.215     0.224      0.22       0.652    cacvcr
                 0.267     0.001      0.5       0.267      0.348      0.893    caco
                 0.145     0.014      0.207     0.145      0.17       0.725    cace
                 0.198     0.014      0.281     0.198      0.232      0.798    caceu
                 0.487     0.038      0.485     0.487      0.486      0.802    cacv
                 0.467     0.065      0.342     0.467      0.395      0.799    cameo
                 0.473     0.052      0.449     0.473      0.46       0.811    capu
                 0.05      0.004      0.125     0.05       0.071      0.772    sateb
                 0.714     0.059      0.527     0.714      0.607      0.877    cap
                 0.687     0.114      0.579     0.687      0.628      0.847    cam
                 0.658     0.005      0.735     0.658      0.694      0.895    caes
                 0.541     0.02       0.549     0.541      0.545      0.857    caov
                 0.458     0.007      0.529     0.458      0.491      0.843    capi
                 0.375     0.011      0.407     0.375      0.39       0.843    cae
                 0         0          0         0          0          0.871    le
                 0         0          0         0          0          0.483    cameo
                 0         0          0         0          0          0.482    capu
                 0.522     0.016      0.619     0.522      0.567      0.846    cari
                 0.344     0.016      0.5       0.344      0.408      0.798    car
                 0         0          0         0          0          0.673    casl
                 0.167     0.002      0.417     0.167      0.238      0.764    cao
                 0         0          0         0          0          0.497    caab
                 0.488     0.006      0.513     0.488      0.5        0.792    camm
                 0         0          0         0          0          0.497    caa
                 0         0          0         0          0          0.496    caav
                 0.417     0.002      0.741     0.417      0.533      0.828    cav
                 0         0          0         0          0          0.497     cam
                 0.569     0.005      0.685     0.569      0.622      0.811    caos
                 0         0          0         0          0          0.499    caos
                 0.323     0.017      0.438     0.323      0.372      0.769    cah
                 0         0          0         0          0          0.499     caes
                 0         0          0         0          0          0.497    caceu
                 0         0.001      0         0          0          0.493    caes
                 0         0          0         0          0          0.497     cam
                 0.569     0.005      0.685     0.569      0.622      0.811    caos
                 0         0          0         0          0          0.499    caos
                 0.323     0.017      0.438     0.323      0.372      0.769    cah
                 0         0          0         0          0          0.499     caes
                 0         0          0         0          0          0.497    caceu
                 0         0.001      0         0          0          0.493    caes
                 0         0          0         0          0          0.986    cam
                 0.686     0.01       0.722     0.686      0.703      0.916    capa
                 0         0          0         0          0          0.5       cari
                 0.241     0.003      0.389     0.241      0.298      0.827     cah
                 0         0          0         0          0          0.497     caos
                 0         0          0         0          0          0.494    cae
                 0.091     0.001      0.2       0.091      0.125      0.809     cao
                 0.423     0.005      0.407     0.423      0.415      0.915    cate
```

1) El alumno deberá interpretar los resultados que obtuvo a partir de la 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. Carlos A. I. Fernández – Fundamentos de los Agentes Inteligentes – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid.
2. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0. Step-by-step data mining guide.
3. Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206.
4. Molina Félix, Luis Carlos. Data Mining: Torturando los datos hasta que confiesen. Coordinador del programa de Data mining (UOC). 2002
5. SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System

PRÁCTICA 10

Interpretación del Resultado

Objetivo:

El alumno va a desarrollar una herramienta capaz de asistir a un usuario a través de todas las etapas del proceso de descubrimiento de conocimiento en bases de datos.

Introducción

Consiste en entender los resultados del análisis y sus implicaciones y puede llevar a regresar a algunos de los pasos anteriores. En esta etapa se debería consolidar el conocimiento ganado, probando los modelos creados contra los resultados obtenidos de la aplicación de estos modelos en el mundo real.
Elaboración de informes para su distribución
•Usar el nuevo conocimiento de forma independiente
•Incorporarlo a sistemas ya existentes (verificar con el conocimiento ya usado para evitar i nconsistencias y posibles conflictos).

La monitorización del sistema en acción dará lugar a nuevos casos que realimentarán el cic lo del KDD Las condiciones iniciales pueden variar, invalidando el modelo adquirido.

Desarrollo

El alumno debe presentar los resultados en un formato entendible. Por esta razón las técnicas de visualización son importantes para que los resultados sean útiles, dado que los modelos matemáticos o descripciones en formato de texto pueden ser difíciles de interpretar para los usuarios finales.

Conclusiones

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

Bibliografía:

7. Carlos A. I. Fernández – Fundamentos de los Agentes Inteligentes – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid.
8. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). CRISP-DM 1.0. Step-by-step data mining guide.
9. Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206.
10. Molina Félix, Luis Carlos. Data Mining: Torturando los datos hasta que confiesen. Coordinador del programa de Data mining (UOC). 2002
11. SK Gupta; Vasudha Bhatnagar; SK Wasan. A proposal for Data Mining Management System


ESTADOS UNIDOS MEXICANOS
UNIVERSIDAD AUTÓNOMA