Data Mining Aplicado a la Predicción y Tratamiento de Enfermedades

Carlos A Vega
Investigación, Universidad Popular Autónoma del Estado de Puebla,
Puebla, Pue. C.P. 72060, México

Genoveva Rosano
Investigación, Universidad Popular Autónoma del Estado de Puebla,
Puebla, Pue. C.P. 72060, México

Juan M López
Investigación, Universidad Popular Autónoma del Estado de Puebla,
Puebla, Pue. C.P. 72060, México

José L Cendejas
Investigación, Instituto Tecnológico de Morelia,
Morelia, Mich. C.P. 58120 , México

Heberto Ferreira
Investigación, Universidad Nacional Autónoma de México, Campus
Morelia. Morelia, Mich. C.P. 58190, México

RESUMEN

Éste artículo propone una metodología de aplicación de algoritmos de predicción, utilizando técnicas de minería de datos, en la cual se incorporan mecanismos de validación a partir de los requerimientos del análisis de los datos, incluyendo la verificación de la significancia (selección y presentación de los mismos) y de mecanismos de validación de los resultados con base en métricas de calidad de la información, los cuales garantizan la efectividad en la construcción del conocimiento. Se utiliza como caso de estudio el análisis clínico de tratamiento de síndrome doloroso abdominal, apoyándose de una base de datos estandarizada en esta área de especialidad.

Palabras Claves: Aplicación de algoritmos de Data Mining, Árboles de decisiones, Data Mining aplicado a la predicción y tratamiento de enfermedades, Algoritmos de predicción, Data Mining aplicado al sector salud.

INTRODUCCIÓN

Hoy en día hay una cantidad excesiva de información que necesita ser estudiada, analizada y depurada para convertirla en conocimiento, dicha información es indispensable y necesaria para la búsqueda de soluciones reales en la toma de decisiones.

Es por esto, que los sistemas de información son tan importantes hoy en día, por que a través de ellos se almacena esta información, en donde es muy necesaria la medición de la calidad de los datos almacenados. [13]

La gran pregunta es ¿cómo puedo obtener éste conocimiento?, una respuesta es utilizando técnicas, algoritmos y mecanismos de validación de minería de datos, la cual se encarga de obtener de la información analizada los patrones de los cuales surge el conocimiento, es por esto que es indispensable medir la calidad de la información que se analizarán con la minería de datos.

Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso, posteriormente realizan un proceso de la limpieza, o preparación de los datos aquí se lleva a cabo la reducción y la transformación de las bases de datos, ya limpia se lleva a cabo la clasificación de las bases de datos, una valoración de la información y finalmente se conforma la base del conocimiento y se pueden tomar decisiones en base a la información clasificada. [14]

Pero no sólo existe el Data Mining, si no también otras herramientas como son los algoritmos de predicción, árboles de decisión, teoría de fractales y teorema de bayes. Entre los algoritmos de predicción más relevantes podemos destacar la lógica difusa ([9] fuzzy logic, la cual ha sido probada en diversos campos de la ingeniería, ecología y psicología, y se trata de un modelo de sistema inteligente conformado por una combinación de los sistemas basados en reglas de conocimiento y de la inteligencia computacional que abarca a las redes neuronales. Se puede emplear en cualquier área de control inteligente o procesamiento de datos en tiempo real), los algoritmos genéticos [8] (los cuales, generan una población de genes con posibles soluciones y los hacen evolucionar para obtener genes más aptos con mejores soluciones) entre otros.

En la actualidad son muchas las áreas del sector salud en las que ha incursionado la minería de datos desde aplicaciones muy complejas como el reconocimiento de imágenes en el cerebro, hasta los procesos para la gestión de los pacientes en hospitales, como por ejemplo en la predicción y el tratamiento de

enfermedades de especialidad analizando la sintomatología, las enfermedades y los resultados de estos tratamientos, siendo un factor importante en la atención de enfermedades tales como: cáncer, problemas del corazón, tratamiento y procesamiento de imágenes, entre otras. [12]

DESARROLLO

Existen diferentes métodos estadísticos que aplican métricas para la calidad de los datos, pero existe una carencia de dicha medición utilizando algoritmos de minería de datos.

[1] y [2] Definen tres pasos para evaluar la calidad de los datos: 1) Obtener las reglas de asociación. 2) Seleccionar las reglas de asociación compatibles. 3) Agregar un factor de confianza como criterios de calidad de los datos de las transacciones. Cabe destacar que éste método presenta dos problemas: el primero, para extraer todas las reglas de asociación necesita un análisis profundo y segundo no existen parámetros o fórmulas exactas para medir la calidad de los datos. En la metodología propuesta en éste material trata de resolver estos problemas debido a que se plantea un levantamiento de requerimientos con la validación de los especialistas en la materia de la aplicación.

[3] Propone un modelo de tres pasos para calcular la calidad de los datos de entrada de las operaciones en su modelo DQM (Data Quality Measurement using Data Mining). En el primer paso extraen las reglas de asociación y estas son adaptadas por dependencias funcionales (T), en el segundo paso separan las reglas de asociación compatible e incompatible y finalmente en el paso tres calculan la calidad de los datos con la Fórmula 1:

$$Q(T) = \frac{n - nc + \sum_{i=1}^{nc} cf_i - \sum_{j=1}^{n-nc} cf_j'}{n} \qquad \text{(Fórmula 1)}$$

Donde:

nc es el número de reglas compatibles,
cf_i es el factor de confianza de la
$_i$ regla de asociación compatible y,
cf_j es el factor de confianza de la regla de asociación incompatible.

Data Mining utiliza diferentes algoritmos de predicción, a continuación se presenta una tabla comparativa de los más relevantes y utilizados. Ver Tabla 1.

Tabla 1: Comparación de los algoritmos de Data Mining más relevantes.

Algoritmo de predicción	Utiliza (Top Down Induction Trees)	Valor Continuo	Permite ejemplos con valores descono-cidos	Utiliza métodos de división	Método de Poda	Aplicación
J48	✓	✓	✓	✓	Post-Data	Aprendizaje exhaustivo
J48Graft	✓	✓	✓	✓	Post-Data	Aprendizaje exhaustivo.
BFTree	✓	✓		✓	Costo Superior	Árboles de decisión
NBTree	✓	✓		✓	Costo Superior	Árboles de decisión
C45	✓	✓		✓	Post-Data	Árboles de decisión
Redes Bayesianas	✓			✓	Poliárboles	Árboles de decisión

METODOLOGÍA DE TRABAJO

A continuación se presenta la metodología con la cual se realizó la presente investigación. Figura 1.

i. Propuesta de investigación:
Se realizaron varias entrevistas con Médicos Especialistas, en las cuales se plantearon diversas problemáticas en el sector salud y las posibles soluciones aplicando las tecnologías de la información con el apoyo de técnicas de minería de datos.

ii. Investigación:
En el proceso de investigación se analizaron distintas enfermedades de especialidad e investigaciones realizadas en el área de la salud, las cuales fueron abordadas con técnicas de

minería de datos.

En pláticas realizadas con Médicos Especialistas se determino que el tratamiento de la información, tiene que ser con datos verídicos, sin omisiones en la información del paciente y del medio en el que se requiere hacer el estudio de dichos datos.

iii. Estandarización de la Información:
Se revisó que la información cumpla con los requerimientos para su análisis, debe ser consistente, sin omisiones y verídica.

iv. Preparación de la Información:
Se crearon las estructuras y formatos necesarios en los archivos que serán analizados por las herramientas de Data Mining.

v. Aplicación del Algoritmo:
Con la información lista para ser analizada se aplicaron las técnicas de minería de datos, haciendo uso de algoritmos que generaron árboles de decisiones.

vi. Obtención de Resultados:
Se obtuvieron frecuencias de los datos y la descomposición de la información en árboles de decisiones, estos datos fueron valorados por médicos especialistas.

Al finalizar la metodología se tuvo una retroalimentación con los médicos especialistas.

Existen diferentes herramientas de minería de datos, por lo que es necesario analizarlos y compararlos para elegir la mejor herramienta a utilizar. Para realizar ésta comparación se utilizarán [10] los once factores de McCall (el cual evalúa tres áreas de trabajo del software: 1) La operación del producto, el cual define la rapidez de su comprensión y operación por parte del usuario. 2) La revisión del producto, el cual se refiere al tiempo y trabajo necesario para corregir errores y la adaptación de los sistemas. 3) La transición del producto, refiriéndose esto a la capacidad de adaptarse a los rápidos cambios de hardware y [11] el modelo de calidad FURPS, el cual define a la funcionalidad, usabilidad, fiabilidad, rendimiento y capacidad de soporte como los factores para definir la calidad del software.

Para poder determinar la solución de software de Data Mining de mejor aplicabilidad se realizó una tabla comparativa entre los productos Orange, Weka, RapidMiner, Jhepwork y Knime utilizando los factores de McCall y FURPS. Tabla 2.

Tabla 2: Tabla comparativa de los diferentes productos de software de Data Mining

Software de Datamining	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**Total**
WEKA	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	50
ORANGE	✓		✓	✓	✓	✓		✓		✓	✓	✓	✓	✓	50
RAPID MINER	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	47
JHEPWORK	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓			✓	38
KNIME	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	42

En la tabla 3 se presentan los elementos evaluados en la tabla comparativa.

Tabla 3: Elementos o factores de evaluación.

Elementos a Evaluar			
1	Flexibilidad	8	Fácil de Configurar
2	Portabilidad	9	Amigable
3	Interoperabilidad	10	Íconos con Ayuda
4	Manual Técnico	11	Seguridad
5	Manual de Usuario	12	Actualizaciones
6	Ayuda en Línea	13	Soporte Técnico
7	Fácil de Instalar	14	Independencia de Hw

De la misma manera se realizó un estudio comparativo de las funcionalidades de las cinco herramientas de software. Ver tabla 4.

Tabla 4: Tabla comparativa de las funcionalidades de software de Data Mining.

Herramientas de data mining	**Procesamiento de datos**	**Modelos Predictivos**	**Métodos de descripción de datos**	**Graficación**	**Validación del modelo**
Weka	✓	✓	✓	✓	✓
Orange	✓	✓	✓	✓	✓
RapidMiner	✓	✓		✓	
JHepWork	✓	✓		✓	
Knime	✓	✓		✓	

Para poder llevar a cabo el tratamiento de la información en aplicaciones médicas, específicamente en análisis y tratamientos de enfermedades, se requiere contar con bancos de datos estandarizados a los requerimientos de información de los especialistas, con atributos definidos, y rangos de valores permitidos, los cuales posibiliten la aplicación de técnicas y algoritmos de minería de datos que permitan generar una mayor exactitud en predicción del conocimiento [13].

Hoy en día, se requieren herramientas que permitan importar y exportar datos en diferentes formatos, como lo son: bases de datos especificas, documentos en hojas de cálculo como Excel, archivos estándar en formatos de Software de Datamining como .ARFF, o incluso que permitan el registro de la información directamente a una base de datos estandarizada.

Por lo anterior, para el proyecto se desarrollaron dos herramientas de importación y exportación de datos, que permiten la interacción de la información con programas de minería de datos como WEKA [16] y Orange Canvas [17], lo cual permite aprovechar las bondades de ambas herramientas para la generación de los resultados de manera gráfica a través de árboles de decisión.

El trabajo propuesto consiste en dos etapas 1) la implementación de un algoritmo para importar y exportar los datos de fuentes locales y externas y 2) la generación de arboles de decisión.

A diferencia del modelo DQM el modelo propuesto evalúa la calidad de los datos al determinar la validez de los resultados con base a un análisis de confiabilidad definiendo el coeficiente de viabilidad “nombre” en base al comportamiento de los datos.

El objetivo es encontrar la detección apropiada de las diferentes sintomatologías que se pueden presentar en el Síndrome doloroso abdominal. Esta enfermedad puede derivarse en diferentes diagnósticos de cierta gravedad de salud, siendo estos normalmente, Apendicitis en sus diferentes grados de complicación, Colecistitis, Piocolecisto, Embarazo Ectopico, Aborto incompleto, Peritonitis, Lesión Vesical Tratogenia, Psesis Abdominal, Irritación Peritonal, Absceso Hepáico Amibiano, Oclusión Intestinal, Quiste Bilateral, Infección y Fastitis de pared abdominal, Abdomen Agudo, entre otras.

DESARROLLO DEL MODELO

En la figura 2 se presenta el modelo de preparación de datos que es generado producto de la investigación.

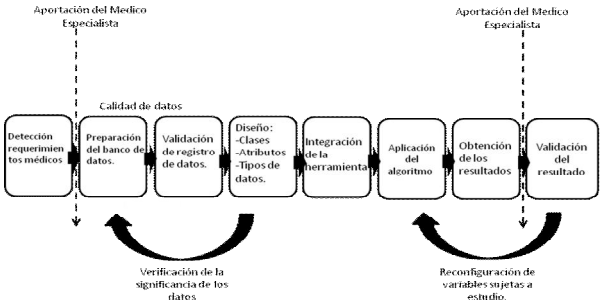


Figura 2: Modelo de aplicación de algoritmos de predicción utilizando técnicas de Data Mining

El modelo propone una detección adecuada de los requerimientos de información, para el caso de estudio los requerimientos que los especialistas determinen, la aplicación de la verificación de la significancia de los datos, la preparación de la información, la aplicación del algoritmo de Data Mining, la obtención de los resultados y la aplicación de un mecanismo de validación de resultados con base a métricas de calidad de los datos, la cual garantiza la efectividad en la aplicación de las herramientas de minería de datos.

En este procedimiento se administran dos o más versiones de algoritmos de predicción de Data Mining utilizando árboles de decisiones. Las reglas de aplicación son similares en contenido, validaciones y otras características. Este método genera resultados confiables si existe una correlación alta entre los resultados de las diferentes aplicaciones. Los patrones de los resultados pueden variar un poco entre las aplicaciones pero las tendencias en las predicciones deben de ser similares lo cual garantizaría su efectividad.

RESULTADOS OBTENIDOS

Una vez realizada la investigación se obtuvieron los siguientes datos, ver figura 3.

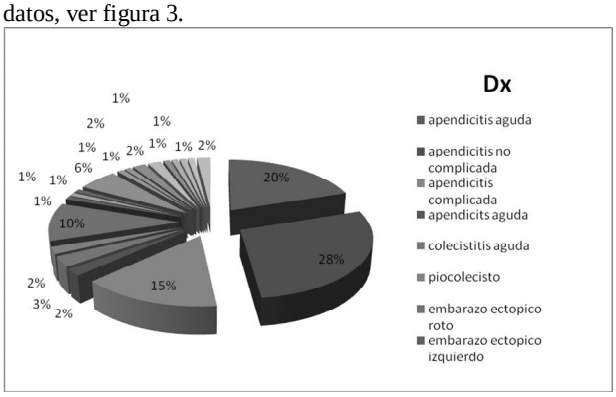


Figura 3: Datos obtenidos por WEKA en el Caso de Estudio Síndrome Doloroso Abdominal.

La confiabilidad de reaplicación de las pruebas muestra hasta donde los puntajes obtenidos en un instrumento pueden ser generalizados a través del tiempo. En la medida que la confiabilidad es mayor, menos susceptibles son los puntajes de ser modificados por las condiciones aleatorias asociadas con la situación de medición o con los cambios de los propios sujetos. El coeficiente de confiabilidad obtenido es una medida de la estabilidad de la prueba.

Como se observa en la figura 4, analizando los primeros 4 niveles de descomposición de los árboles de decisión se encuentra que en el primer y segundo nivel la variación es mínima (con una sola diferencia en la predicción para cada nivel), teniendo una probabilidad del 86% lo cual demuestra una alta confiabilidad para los algoritmos de predicción utilizados, siguiendo con el análisis de la tabla 6 en el tercer nivel la variación tiende a aumentar por las diferentes combinaciones generadas durante la ramificación del árbol, sin embargo, la probabilidad es del 71% lo cual demuestra ser elevada debido a que se encuentra un nivel superior de descomposición y para el cuarto nivel analizado sube nuevamente la confiabilidad a un 86%.

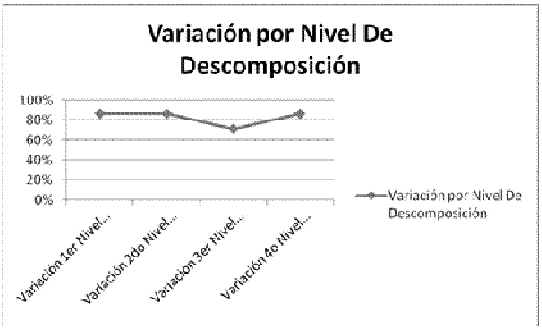


Figura 4. Variación de los niveles de descomposición en los árboles de decisión.

Aplicando la fórmula para obtener el coeficiente de correlación por el método de los puntajes directos, el cual se expresa en la fórmula siguiente:

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left[N\sum X^2 - \left(\sum X\right)^2\right]\left[N\sum Y^2 - \left(\sum Y\right)^2\right]}}$$ (Fórmula 2)

En donde:

r , es el coeficiente de correlación entre las dos administraciones de la prueba.

N = número de sujetos

ΣXY = resultado de sumar el producto de cada valor de X por su correspondiente valor en Y.

ΣX = suma total de los valores de X (primera aplicación).

ΣY = suma total de los valores de Y (segunda aplicación).

Σx^2 = resultado de sumar los valores de X elevados al cuadrado.

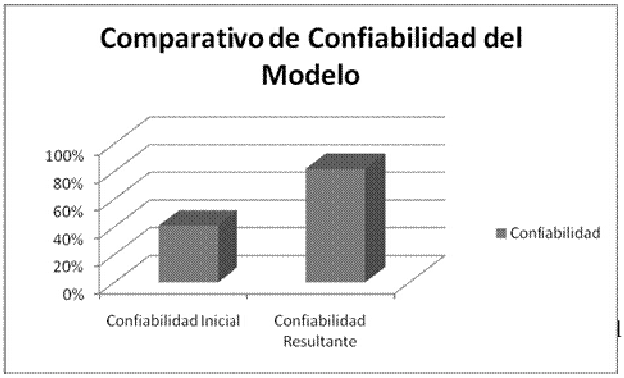
Σy^2 = resultado de sumar los valores de Y elevados al cuadrado.

$(\Sigma x)^2$ = suma total de los valores de X, elevada al cuadrado.

$(\Sigma y)^2$= suma total de los valores de Y, elevada al cuadrado.

Conteniendo la variable X los valores de cada resultado encontrado de la primer medición y Y el valor de cada resultado de las siguientes mediciones y sustituyendo los valores correspondientes en la fórmula, se encontró un resultado del 92%, lo cual indica que existe una alta correlación entre los resultados obtenidos de la aplicación de los 7 algoritmos de predicción de Data Mining utilizando arboles de decisión, lo cual hace notar que la predicción del análisis de los datos es confiable, en cuanto a la estabilidad de las puntuaciones a través del tiempo con los datos existentes.

Como parte del proceso de comprobación del modelo se realizó un comparativo entre el resultado generado sin la aplicación del modelo obteniendo un 41% de confiabilidad con los datos iniciales, en contraste con el 82% de confiabilidad combinatoria en los 4 niveles utilizando el modelo propuesto, lo cual garantiza una eficacia superior en su utilización y aplicación. Ver figura 5.



Para validar los resultados de la predicción del modelo sobre un escenario de prueba real, se tomaron en cuenta 10 registros aleatorios y se analizaron sus resultados. Este instrumento, a su vez, se repitió por 10 grupos de pruebas encontrando una eficacia del 90% en los resultados.

Es de destacar, que en el caso de estudio el dolor abdominal crónico, sólo había 110 registros en la base de datos del Hospital Regional de Tuxtla Gutiérrez, Chiapas, donde como en varios hospitales encuestados, se detecta una falta de información debido a la falta de estandarización de estas bases de datos.

Se recomienda para trabajos futuros obtener una base de datos más completa, con un mayor número de casos, con información estandarizada que permita replicar el modelo y los escenarios de prueba.

Por último, también se recomienda tomar los resultados del modelo para evaluar y replicar este, con otro grupo de enfermedades.

DISCUSIÓN Y CONCLUSIONES

Es importante destacar que los datos son la materia prima de la minería de datos y si éstos son incorrectos todo el análisis y por lo tanto el conocimiento generado también será incorrecto, lo que conlleva a una solución falsa que puede ocasionar una mala toma de decisiones dentro de una empresa o en nuestro caso en el tratamiento de una enfermedad y puede llegar a causar la muerte de las personas. Es decir, debe de haber una integridad de la información entre la que se encuentra en la vida real y la que se encuentra registrada en las bases de datos y sistemas de información empresariales. En el caso de estudio y con los datos iniciales el porcentaje de confiabilidad era bajo de sólo el 41%, debido principalmente a la falta de una estandarización de información con mediciones diferentes, campos incompletos, rangos de valores no definidos, diferentes escalas de valoración y diagnósticos con resultados diferentes. Al aplicar el modelo el porcentaje de confiabilidad se incremento a un 82% de confiabilidad.

El síndrome doloroso abdominal es una enfermedad muy común entre los pacientes, que tiende a ser mal interpretada y muchas veces se trata con medicamentos generales que no ayudan a resolver el padecimiento especifico y se mal interpreta con enfermedades como Gastritis, Infecciones Estomacales, Inflamaciones en el abdomen por mencionar algunas, con el fin de poder brindar un mejor diagnóstico en el tratamiento de esta enfermedad es necesario identificar y asociar sintomatologías especificas de una apendicitis con la ayuda de las herramientas de minería de datos para “Reducir el riesgo en el tratamiento y diagnostico de esta enfermedad.”

El Modelo de preparación de datos para su aplicación en técnicas de minería de datos es una base importante para el desarrollo de nuevas aplicaciones de minería de datos que puedan generar diagnósticos completos y prevenciones en enfermedades de especialidad [18].

REFERENCIAS

[1] Strong, D.M,Lee Y.W,wang, R.Y., “Data Quality in Contex”, communication of ACM, 40(5),1997.

[2] Pipino,L.,lee,W., Wang,y., “Data Quality Assessment”, 1998.

[3] Saeed Farzi, Ahmad Baraani Dastjerdi., “Data Quality Measurement using Data Mining”, International Journal of Computer Theory and Engineering, Vol. 2, No. 1 February, 2010

[4] Estopa, Rosa., Valero, Antoni., “Adquisición de Conocimiento Especializado y Unidades de Significación Especializada en Medicina”., sin fecha.

[5] Zavala Vaca, Hugo Alejandro., Ferreira Medina, Herberto., “Análisis comparativo de herramientas de monitoreo y control de redes, utilizando software libre para Institutos de Investigación”., CIGA., CIECO., sin fecha.

[6] Ruiz Bolívar, Carlos., “Confiabilidad”., Programa Interinstitucional Doctorado en Educación., sin fecha.

[7] Segura Bolívar, John Alexander., Obregón Neira, Nelson., “ Un modelo de lógica difusa y conjuntos difusos para el pronóstico de los niveles medios diarios del río Magdalena, en la estación limnigráfica de Puerto Salgar, Colombia”., revista de ingeniería #22 facultad de ingeniería universidad de los andes noviembre 2005.

[8] Mathew, Tom V., “Genetic Algorithm”., Indian Institute of Technology Bombay., sin fecha.

[9] I. Aydin, M. Karakose, E. Akin., “The Prediction Algorithm Based on Fuzzy
Logic Using Time Series Data Mining Method”., World Academy of Science, Engineering and Technology 51 2009.

[10] Gillies, A. 1997. “Software Quality: Theory and Management”., Thomson, London.

[11] Grady, Robert B., 1992., “Practical software metrics for project management and process improvement”.

[12] Villalobos J. Ángel, Expediente-e. Telemedicina, “Minería de Datos” y apoyo por Teléfono”, México, 2009.

[13] Laudon Kennet, Sistema de Información Gerencial. Administración de la empresa digital, Pearson Educación, México, Octava Edición, 2004

[14] Ian H. Witten & Eibe Frank, Data Mining, Machine Learning tools and techniques, Editorial Morgan Kaufmann, San Francisco, CA, Segunda Edición, 2005.

[16]Sitio de la Universidad de Waikato apartado de software especializado en minería de dato. http://www.cs.waikato.ac.nz/~ml/weka/

[17] Sitio oficial del Software Orange Canvas. http://www.ailab.si/orange/

[18] D. Canlas Rubén. Data Mining in Healthcare: Current Applications and Issues. Carnegie Mellon University, Australia 2009. PAG 3-9.