

TESIS DEFENDIDA POR  
**DAVID MADRIGAL ANGULO**  
Y APROBADA POR EL SIGUIENTE COMITÉ

---

Dr. Fernando Rojas Iñiguez  
*Director del Comité*

---

Dra. Ana Isabel Martínez  
*Miembro del Comité*

---

Dr. Hugo Hidalgo Silva  
*Miembro del Comité*

---

Dr. Arturo Velázquez Ventura  
*Miembro del Comité*

---

Dr. Armando Paniagua Michel  
*Miembro del Comité*

---

Dr. Pedro Gilberto López Mariscal  
*Coordinador del programa de posgrado  
en ciencias de la computación*

---

Dr. David Hilario Covarrubias Rosales  
*Director de Estudios de Posgrado*

28 de Noviembre de 2008

**CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN SUPERIOR  
DE ENSENADA**



---

**PROGRAMA DE POSGRADO EN CIENCIAS  
EN CIENCIAS DE LA COMPUTACIÓN**

---

**Toma de Decisiones Médicas con Técnicas de Minería de Datos y  
Cómputo Móvil**

TESIS

que para cubrir parcialmente los requisitos necesarios para obtener el grado de  
MAESTRO EN CIENCIAS

Presenta

DAVID MADRIGAL ANGULO

Ensenada, Baja California, México, Noviembre del 2008

**RESUMEN** de la tesis de DAVID MADRIGAL ANGULO presentada como requisito parcial para la obtención del grado de MAESTRO EN CIENCIAS en CIENCIAS DE LA COMPUTACIÓN. Ensenada, Baja California. Noviembre de 2008.

### **Toma de decisiones médicas con técnicas de minería de datos y cómputo móvil**

Resumen aprobado por:

---

Dr. Fernando Rojas Iñiguez  
Director de Tesis

Cada vez son más las organizaciones del cuidado de la salud que requieren convertir sus datos de la memoria organizacional en conocimiento útil para apoyar los procesos del cuidado de la salud de los pacientes. Hoy en día es de gran importancia el utilizar nuevas tecnologías de apoyo, donde se pueden usar sistemas de ayuda a la toma de decisiones clínicas (CDSS) y que considere a los médicos como trabajadores de conocimiento con necesidades de movilidad y pueden tomar una decisión en cualquier lugar y en cualquier momento.

En este trabajo de tesis exploramos cómo apoyar a un proceso de toma de decisiones en un contexto hospitalario, mediante el diseño de un CDSS con modelos obtenidos mediante técnicas de minería de datos para extraer conocimiento y embeberlo en un dispositivo móvil. Estudiamos el proceso de detección y vigilancia de infecciones nosocomiales (PDVIN) en un hospital público donde obtuvimos la estructura del proceso, el conocimiento organizacional, con el cual identificamos los atributos para la minería de datos y de ahí obtener los requerimientos de soporte directamente de los dueños del proceso.

La base para la extracción del conocimiento es el formato de registro de caso de infección nosocomial (FRCIN). El FRCIN contiene todos los datos clínicos del paciente. Asimismo dentro del PDVIN se identificaron 4 eventos de generación de información/atributos durante la estadía de un paciente en el hospital. Para cada evento aplicamos el proceso de extracción de conocimiento donde se usaron de forma iterativa técnicas de generalización de conceptos, transformación de datos y reducción de atributos hasta obtener los modelos finales caracterizados con la mejor evaluación. Se obtuvieron 5 árboles de decisión con el algoritmo C4.5 y 3 clasificadores *Naive Bayes* los cuales tienen como objetivo la detección de infecciones nosocomiales (IN).

Los árboles de decisión y los clasificadores bayesianos se evaluaron utilizando validación cruzada y con un conjunto independiente que no se usó para el entrenamiento de los modelos. Los resultados de esta evaluación muestran un 90% de instancias clasificadas correctamente por los árboles de decisión y 87% por parte del clasificador *Naive Bayes*.

Se desarrolló un CDSS móvil y de escritorio los cuales están embebidos dentro del flujo de trabajo del PDVIN. Asimismo permiten la valoración de riesgo de contraer infección

nosocomial mediante los modelos obtenidos. Las herramientas tienen además mecanismos para la administración de los datos del paciente durante la ejecución del proceso. El sistema de escritorio tiene soporte para la administración de la memoria organizacional y soporte para el sub-proceso de control del PDVIN.

Se realizó una evaluación para conocer la percepción de utilidad y de uso del sistema y los modelos que tienen los médicos sobre el CDSS desarrollado, mediante un conjunto de actividades de uso de los sistemas/modelo y posteriormente se realizaron entrevistas/cuestionarios. Los resultados de la evaluación se presentan de forma cualitativa. Se observa que el desarrollo de los sistemas de soporte a decisión con minería tienen a ser más aceptados cuando los dueños de los procesos se involucran en su desarrollo, en cuanto a los modelos se puede mencionar que los usuarios tienen una inclinación hacia el clasificador estadístico de *Naive Bayes*.

**Palabras Clave:** sistemas de apoyo a toma de decisiones clínicas, CDSS móviles, proceso de extracción de conocimiento en bases de datos, técnicas de minería de datos, modelos predictivos, árboles de decisión, clasificador *Naive Bayes*.

**ABSTRACT** of the thesis presented by DAVID MADRIGAL ANGULO as a partial requirement to obtain the MASTER OF SCIENCE degree in COMPUTER SCIENCE. Ensenada, Baja California, México November of 2008.

### **Making medical decisions with data mining techniques and mobile computing**

Increasingly, there are many organizations of health care that need to convert their data from memory organizational in useful knowledge to support the processes of care for patients. Nowadays is of great importance to use the new support technologies, where systems can be used to help clinical decision-making (CDSS) and to consider the doctors as knowledge workers with mobility needs and can make a decision on any place at any time.

In this thesis work we explore how to support a decision-making process in a hospital context. by designing a CDSS with models obtained from data mining to extract knowledge and embedded in a mobile device. We study the process of detection and surveillance of nosocomial infections (PDVIN) in a public hospital where we obtained the structure of the process, organizational knowledge, with which we identify the attributes for data mining and hence the requirements to obtain support directly from process's owners.

The basis for the extraction of knowledge is the format for recording cases of nosocomial infection (FRCIN). The FRCIN contains all the clinical data of patients. Also within the PDVIN we identified 4 events where information / attributes are generated during the stay of a patient at the hospital. For each event we apply the process of extracting knowledge from, by using an iterative method, generalization of concepts, data transformation and reduction of attributes to get the final models characterized with the best performance. 5 trees were obtained with the decision algorithm C4.5 Naïve Bayes classifiers a 3 which are aimed at detecting nosocomial infections (IN).

Decision trees and Bayesian classifiers were evaluated using cross-validation and whit a separate set that is not used for training the models. The results of this assessment show 90% of instances classified correctly by the decision trees and 87% by the naive Bayesian classifier.

We implemented CDSS both in mobile and desktop which are embedded within the workflow PDVIN. They also allow assessment of the risk of nosocomial infection and the models obtained. The tools also have mechanisms for the management of patient data during the process. The desktop system has support for management of organizational memory and support for the sub-process of control PDVIN.

An evaluation was performed to determine the perception of usefulness, a use of the system and the models that the doctors on the CDS developed through, a set of activities od use the systems / model and then conduct interviews / questionnaires. The evaluation results are presented in a qualitative way. It is noted that the development of decision support systems with data mining are more accepted when the owners of processes are involved in its development, in terms of the models wen can mentioned that users have an inclination toward statistical classifier Naive Bayes.

**Keywords:** Clinical decision support system (CDSS), mobile CDSS, knowledge discovery process in databases, Data mining the, predictive models, decision tree, Naïve Bayes classifier.

## **Dedicatorias**

*A mis padres Silvia y Juan  
Por darme la vida y enseñarme a vivirla*

*A mis hermanos Juan Gonzalo, José Luis y Alejandro  
Por su cariño y apoyo*

## **Agradecimientos**

*Al Dr. Fernando y a la Dra. Ana por sus sabios consejos pero sobre todo por su paciencia y amistad.*

*Al Dr. Paniagua por su tiempo, dedicación y sus buenos consejos, al Dr. Hugo y al Dr. Arturo por sus amplios comentarios.*

*A los profesores del departamento de ciencias de la computación  
Que contribuyeron en mi formación*

*A mis tíos Pablo, Gaby, Carmen y Mauricio por siempre estar al pendiente y hacerme sentir cerca de casa*

*A mis amigos del posgrado Abdel, Alejandra, Alfonso, Angel, Antonio, Cesar, Daniela, Gama, Gustavo, Jehovani, Jessica, Jose luis, Juan, Leo, Liz, Luis, Miroslava, Nancy, Pablo, Raúl, Rolando, Salvador y Víctor por brindarme su sincera amistad, cada uno de ustedes puso algo muy interesante en esta aventura.*

*A mis amigos de Colima Astrid, Betsy, Eduardo, Lili, Lily, Lucia, Luis, Mich, Mónica, Norma, Rita y Roberto por siempre darme ánimos*

*Al Centro de Investigación Científica y de Educación Superior de Ensenada y al Consejo Nacional de Ciencia y Tecnología por su apoyo económico. ¡No te acabes beca Conacyt!*





## CONTENIDO (continuación)

	<b>Página</b>
II.4. Sistemas de apoyo a la toma de decisiones.....	20
II.5. Sistemas de apoyo a la toma de decisiones en un contexto clínico .....	25
II.6. Cómputo móvil aplicado al contexto médico .....	28
II.7. Trabajo previo.....	29
II.8. Resumen.....	32
<b>Capítulo III. Descubrimiento de conocimiento en bases de datos.....</b>	<b>33</b>
III.1. Introducción .....	33
III.2. Descubrimiento de conocimiento en bases de datos.....	33
III.2.1. Proceso de descubrimiento de conocimiento en bases de datos .....	35
III.2.2. Procesamiento previo y transformación de la información .....	38
III.3. Minería de datos.....	40
III.3.1. Tareas de minería datos.....	42
III.3.2. Métodos de minería de datos .....	43
III.3.3. Evaluación del conocimiento adquirido.....	62
III.4. Resumen.....	65
<b>Capítulo IV. Proceso de detección y vigilancia de infecciones nosocomiales</b> <b>.....</b>	<b>66</b>
IV.1. Introducción .....	66
IV.2. Estudio del proceso de detección y vigilancia de infecciones nosocomiales .....	67
IV.2.1. Descripción textual del proceso .....	70
IV.3. Modelado del caso de estudio.....	73
IV.3.1. Grafica rica global del proceso de detección y vigilancia de IN .....	74
IV.3.2. Grafica rica del sub-proceso de detección de IN .....	75
IV.3.3. Grafica rica del sub-proceso de información de IN .....	75
IV.3.4. Grafica rica del sub-proceso de control de IN .....	76
IV.3.5. Grafica de generación de información de acuerdo a eventos.....	77

## CONTENIDO (continuación)

	<b>Página</b>
IV.4. Requerimientos de diseño del CDSS .....	77
IV.4.1.    Requerimientos del CDSS .....	78
IV.4.2.    Requerimientos resultantes del caso de estudio.....	78
IV.5. Propuesta de soporte al PDVIN .....	80
IV.6. Resumen.....	81
<b>Capítulo V. Proceso de Extracción de Conocimiento .....</b>	<b>83</b>
V.1. Introducción .....	83
V.2. Proceso de extracción de conocimiento.....	83
V.3. Herramienta para la extracción de conocimiento .....	87
V.4. Generación de información del PDVIN.....	87
V.5. Procesamiento previo de la información y creación de la vista minable.....	91
V.6. Transformación de datos y generalización de conceptos.....	91
V.7. Reducción de atributos.....	98
V.8. Análisis de exploración.....	103
V.9. Modelos predictivos definitivos.....	104
V.9.1.    Modelos predictivos por el clasificador <i>Naive Bayes</i> .....	104
V.9.2.    Modelos predictivos por árboles de decisión.....	112
V.10. Resumen.....	121
<b>Capítulo VI. Diseño e implementación de CDSS .....</b>	<b>123</b>
VI.1. Introducción .....	123
VI.2. Diseño arquitectónico .....	124
VI.3. Funcionalidad (casos de uso).....	126
VI.4. Comportamiento (diagrama de secuencia) .....	130

## CONTENIDO (continuación)

	<b>Página</b>
VI.5. Implementación .....	134
VI.5.1. Implementación del diseño arquitectónico propuesto.....	134
VI.5.2. Implementación del diseño de la funcionalidad propuesto.....	136
VI.5.3. Escenario de uso del prototipo del CDSS desarrollado .....	140
VI.6. Resumen.....	141
<b>Capítulo VII. Evaluación de los modelos predictivos y del CDSS .....</b>	<b>142</b>
VII.1. Introducción .....	142
VII.2. Evaluación de los modelos predictivos.....	142
VII.2.1. Evaluación de los modelos mediante validación cruzada.....	143
VII.2.2. Evaluación de los modelos predictivos con el conjunto de prueba.....	146
VII.3. Pruebas con usuarios al CDSS.....	150
VII.3.1. Definición del problema de evaluación. ....	150
VII.3.2. Diseño de la evaluación .....	150
VII.3.3. Actividades realizadas.....	151
VII.3.4. Análisis cualitativo de los resultados encontrados en la evaluación.....	152
VII.3.5. Discusión de los resultados .....	156
VII.4. Resumen.....	157
<b>Capítulo VIII. Conclusiones.....</b>	<b>158</b>
VIII.1. Conclusiones .....	158
VIII.2. Aportaciones .....	161
VIII.3. Trabajo a futuro .....	162
<b>Referencias.....</b>	<b>163</b>
<b>Apéndice A. Modelado del PDVIN con Rads y DFD.....</b>	<b>168</b>
<b>Apéndice B. Valores del atributo antibiótico y germen.....</b>	<b>176</b>

**CONTENIDO (continuación)****Página**

<b>Apéndice C. Modelos predictivos del obtenidos del análisis de exploración.....</b>	<b>178</b>
<b>Apéndice D. actividades de la evaluación.....</b>	<b>190</b>

## LISTA DE FIGURAS

	<b>Página</b>
Figura 1. Metodología de investigación. ....	9
Figura 2. Clasificación continua de los problemas de decisión, propuesta por Gory y Scott-Morton (1971), extendida de la clasificación discreta de Simon (1960) .....	16
Figura 3. Perspectivas de toma de decisiones .....	17
Figura 4. Ciclo de inteligencia organizacional. ....	19
Figura 5. Elementos de los DSS. ....	22
Figura 6. Proceso de extracción de conocimiento el cual se destaca el de minería de datos (Fayyad et al., 1996.) .....	37
Figura 7. Árbol de decisión.....	46
Figura 8. Árbol de clasificación según las clases p, r, s. ....	49
Figura 9. Red Bayesiana .....	56
Figura 10. Pseudocódigo del algoritmo K2 .....	59
Figura 11. El PDVIN, consta de tres sub-procesos en el cual se muestran los elementos de información utilizados y actividades que tiene cada uno. ....	69
Figura 12. Grafica rica global del proceso de detección y vigilancia de infecciones nosocomiales.....	74
Figura 13. Grafica rica para el sub proceso de detección epidemiológica de infecciones nosocomiales.....	75
Figura 14. Grafica rica para el sub-proceso de información de infecciones nosocomiales. .	76
Figura 15. Grafica rica para el sub-proceso de control de infecciones nosocomiales. ....	76
Figura 16. Grafica de generación de información de acuerdo a evento durante todo el proceso de detección y vigilancia de infecciones nosocomiales. ....	77
Figura 17. Soporte propuesto al PDVIN.....	80
Figura 18. Proceso global de extracción del conocimiento. ....	86
Figura 19. Eventos de generación de información. ....	88
Figura 20. Relevancia de atributos. ....	99
Figura 21. Relevancia de Atributos II.....	100
Figura 22. Relevancia de Atributos III. ....	100
Figura 23. Relevancia de Atributos IV. ....	101
Figura 24. Instancias clasificadas correctamente.....	102
Figura 25. Comparación de las medidas de desempeño. ....	103
Figura 26. Modelo de extracción de conocimiento dinámico con el clasificador NaiveBayes. ....	105
Figura 27. Clasificador bayesiano en Ingreso del Paciente. ....	106
Figura 28. clasificado Naive Bayes en revisión del paciente. ....	109
Figura 29. Clasificador Naive Bayes en Ingreso de Análisis de laboratorio. ....	111
Figura 30. Modelo de extracción de conocimiento dinámico con árboles de decisión .....	113
Figura 31. Modelo de árbol de decisión para el ingreso del paciente.....	114
Figura 32. Modelo de árbol de decisión para la revisión del paciente.....	117
Figura 33. Modelo de árbol de decisión para la revisión del paciente con cirugía.....	118

## LISTA DE FIGURAS (continuación)

	<b>Página</b>
Figura 34. Modelo de árbol de decisión para la fase de detección de IN. ....	119
Figura 35. Modelo de árbol de decisión para la fase de detección de IN de solo cirugía. ....	120
Figura 36. Ciclo de extracción de conocimiento. ....	121
Figura 37. Diseño arquitectónico del sistema. ....	124
Figura 38. Diagrama de casos de uso para el sistema móvil y de base ....	127
Figura 39. Diagrama de secuencia del caso de uso “Valoración” ....	131
Figura 40. Diagrama de secuencia del caso de uso “Valorar censo” ....	133
Figura 41. Diseño de la implementación del sistema ....	135
Figura 42. Desarrollo de la interfaz de usuario para (a) computadora de escritorio y (b) dispositivo móvil. ....	137
Figura 43. Flujo de información a través del censo, censo valorado (a), información básica del paciente (b), fecha de detección de IN y sitio de infección que tiene el paciente (c), motivo de ingreso y diagnóstico del paciente (d). ....	138
Figura 44. Flujo de información, factores de riesgo del paciente (a), antecedentes quirúrgicos, información microbiológica (b) y el desenlace del paciente (d). ....	138
Figura 45. Opciones de valoración que tiene el médico tomador de decisión, se encuentra en la fase de ingreso de análisis de laboratorio, con el modelo de Naive Bayes(a), valoración con Naive Bayes (b).....	139
Figura 46. Opción de valoración que tiene el médico tomador de decisión con árboles de decisión C4.5 (a), regla utilizada para llegar a ese resultado (b). ....	140
Figura 47. Instancias clasificadas correctamente con Naive Bayes.....	144
Figura 48. Comparación de las medidas de precisión y alcance de acuerdo a los eventos de generación de información con Naive Bayes. ....	144
Figura 49. Instancias clasificadas correctamente con árboles de decisión. ....	145
Figura 50. Comparación de las medidas de precisión y alcance de acuerdo a los eventos de generación de información con árboles de decisión. ....	146
Figura 51. Porcentaje de instancias clasificadas s clasificadas correctamente. ....	147
Figura 52. Comparativa de las medidas de precisión y alcance del clasificador Naive Bayes. ....	148
Figura 53. Porcentaje de instancias clasificadas s clasificadas correctamente ....	149
Figura 54. Comparativa de las medidas de precisión y alcance de en arboles de decisión	149
Figura 55. Rad detectando vigilando la infección nosocomial ....	169
Figura 56. Rad identificando casos sospechosos por recorrido ....	170
Figura 57. Rad sobre controlando la infección nosocomial.....	170
Figura 58. Vigilando infecciones nosocomiales. ....	171
Figura 59. Rad Llenado de la forma FRCIN ....	172
Figura 60. Vista global del diagrama del flujo de datos ....	173
Figura 61. DFD sobre en la fase de detección de infección nosocomial ....	173
Figura 62. DFD de llenado de la forma FRCIN ....	174
Figura 63. DFD sobre la fase de vigilancia ....	174
Figura 64. DFD sobre el proceso de control de infección nosocomial.....	175

## LISTA DE FIGURAS (continuación)

	<b>Página</b>
Figura 65. Red Bayesiana .....	183
Figura 66. Clasificador naive bayes para la detección de Neumonía .....	187
Figura 67. Ejemplo de una regla en la PDA .....	191
Figura 68. Pantalla inicial del sistema FRCIN .....	192
Figura 69. Para dar de alta un paciente con sus respectivos datos.....	192
Figura 70. Pantalla inicial del CDSS Móvil .....	194
Figura 71. Censo de pacientes .....	194
Figura 72. Información básica del paciente .....	195
Figura 73. Sitio de infección del paciente.....	196
Figura 74. Factores de riesgo del paciente.....	198
Figura 75. valoración con Naive Bayes .....	199
Figura 76. Valoración con redes bayesianas.....	200
Figura 77. Valoración con árboles de decisión.....	200
Figura 78. Valoración con árboles de decisión.....	201
Figura 79. Regla utilizada para la decisión.....	201
Figura 80. Censo de pacientes .....	202
Figura 81. Antecedentes Quirúrgicos .....	203
Figura 82. Valoración con árboles de decisión.....	204
Figura 83. Valoración con árboles de decisión.....	204
Figura 84. Actualizar los cambios en el sistema FRCIN.....	205
Figura 85. Resultado de la búsqueda de “Paco“ .....	206
Figura 86. Antecedentes quirúrgicos .....	206

## LISTA DE TABLAS

	<b>Páginas</b>
Tabla I. Matriz de confusión para las tuplas positivas y negativas.....	64
Tabla II. Simbología de la grafica rica. ....	73
Tabla III. Información en la forma FRCIN.....	89
Tabla IV. Conjuntos de datos generados durante el proceso del pre-procesamiento de la información.....	91
Tabla V. Generalización del atributo desenlace .....	92
Tabla VI Generalización de conceptos del atributo sitio de infección. ....	93
Tabla VII. Generalización de conceptos del atributo edad.....	94
Tabla VIII. Generalización de conceptos del atributo DE.....	95
Tabla IX. Generalización de conceptos del atributo FR.....	95
Tabla X. Generalización de conceptos del atributo contaminación.....	96
Tabla XI. Generalización de conceptos del atributo Edad_2.....	96
Tabla XII. Generalización de conceptos del atributo DE_2.....	97
Tabla XIII. Generalización de conceptos del atributo DSEH_2.....	97
Tabla XIV. Comparación de modelos en ingreso del paciente.....	105
Tabla XV. Distribución de probabilidad condicional del atributo desenlace.....	106
Tabla XVI. Distribución de probabilidad condicional del atributo inmunocompromiso...106	106
Tabla XVII. Distribución de probabilidad condicional del atributo Servicio.....	106
Tabla XVIII. Distribución de probabilidad condicional del atributo Edad_2.....	106
Tabla XIX. Distribución de probabilidad condicional del atributo DE_2.....	107
Tabla XX. Comparación de modelos en revisión del paciente.....	109
Tabla XXI. Distribución de probabilidad condicional del atributo desenlace.....	109
Tabla XXII. Distribución de probabilidad condicional del atributo enfermedades degenerativas.....	109
Tabla XXIII. Distribución de probabilidad condicional del atributo inmunocompromiso.109	109
Tabla XXIV. Distribución de probabilidad condicional del atributo servicio.....	110
Tabla XXV. Distribución de probabilidad condicional del atributo sitio de infección. ....	110
Tabla XXVI. Distribución de probabilidad condicional del atributo DE_2.....	110
Tabla XXVII. Distribución de probabilidad condicional del atributo edad.....	110
Tabla XXVIII .Comparación de modelos en ingreso de análisis de laboratorio.....	111
Tabla XXIX. Distribución de probabilidad condicional del atributo microorganismo. ....	112
Tabla XXX. Distribución de probabilidad condicional del atributo DE.....	112
Tabla XXXI. Distribución de probabilidad condicional del atributo sitio de infección.....	112
Tabla XXXII. Distribución de probabilidad condicional del atributo Edad_2.....	112
Tabla XXXIII. Comparación de modelo en revisión del paciente.....	116
Tabla XXXIV. Comparación de los modelos revisión del paciente con solo cirugía .....	117
Tabla XXXV. Comparación de los modelos en detección de IN .....	119
Tabla XXXVI. Comparación de los modelos en detección IN con solo cirugía .....	119



## LISTA DE TABLAS (Continuación)

	<b>Páginas</b>
Tabla XXXVII. Casos de uso principales.....	128
Tabla XXXVIII. Condensado de los resultados obtenidos en la evaluación preliminar del CDSS. ....	155
Tabla XXXIX. Valores posibles del atributo sensibilidad y germen .....	176
Tabla XL. Distribución de probabilidad condicional del atributo neumonía .....	187
Tabla XLI. Distribución de probabilidad condicional del atributo enfermedades degenerativas. ....	187
Tabla XLII. Distribución de probabilidad condicional del atributo inmunocompromiso. .	187
Tabla XLIII. Distribución de probabilidad condicional del atributo servicio. ....	188
Tabla XLIV. Distribución de probabilidad condicional del atributo sitio de infección. ....	188
Tabla XLV. Distribución de probabilidad condicional del atributo DE_2.....	188
Tabla XLVI. Distribución de probabilidad condicional del atributo edad. ....	188
Tabla XLVII. Información del paciente .....	193
Tabla XLVIII. Datos clínicos del paciente Javier.....	198
Tabla XLIX. Antecedentes quirúrgicos .....	203

## Introducción

---

### I.1. Antecedentes

Desde principios de la década de los 60's, los sistemas de información en hospitales (en lo sucesivo, HIS – por sus siglas en inglés, *Hospital Information System*-), han sido desarrollados para cubrir las funciones médicas y administrativas (Collen, 1999). Sin embargo los primeros sistemas estaban centrados en las actividades administrativas del hospital principalmente en aspectos de facturación y reembolso, además de la admisión del paciente al hospital, administrar las citas y proporcionar servicios a los departamentos auxiliares tales como la farmacia, laboratorio y departamento de radiología. Fueron desarrollados para ayudar a los procesos manuales sin brindar un valor agregado.

En los años 90, llegó a estar claro que el objetivo de reducir gastos, podría ser alcanzado creando sistemas de información que asistan a los médicos en sus actividades cotidianas en el cuidado de la salud, tales como la toma de decisiones para diagnósticos clínicos. El desarrollo de los sistemas de información clínicos (en lo sucesivo, CIS – por sus siglas en inglés *Clinical Information System*-), son una evolución natural de los HISs.

En el curso normal de una actividad clínica los profesionales del cuidado de la salud presentan y manipulan diversas piezas de información. En esta presentación y manipulación –ocurre la toma de decisión- con frecuencia bajo condiciones de stress, sobrecarga cognoscitiva, incertidumbre y en un aumento en el escrutinio. Esto deja que las decisiones

clínicas sean imperfectas y que sea menos óptimo el resultado (Englehardt y Nelson, 2001). Esto nos brinda una motivación para el uso de sistema de apoyo a toma de decisiones en un contexto clínico (en lo sucesivo, CDSS -por sus siglas en inglés *Clinical Decision Support System*-).

Los CDSS “*son sistemas que ayudan directamente en la toma de decisiones médicas, donde las características individuales de los pacientes son usadas para su valoración y posteriormente presentadas a los médicos para su consideración*” (Holroyd, 2007).

El reto a afrontar de las tecnologías de información es como incorporar la ciencia de la toma de decisiones en sistemas que sean exactos, descriptivos, discretos y que se usen como parte de una actividad clínica cotidiana (Englehardt y Nelson, 2001). Por lo tanto los CDSS son una extensión de los sistemas de apoyo a la decisión (DSS) -por sus siglas en inglés *Decision Support System*-).

Las tendencias actuales en la investigación y desarrollo de los sistemas de apoyo a decisión según Shim (2002), son las siguientes:

- **Perspectiva socio-técnica**, para diseñar un sistema influyen tanto un enfoque técnico, como uno social y organizacional, para formular y resolver un problemas de decisión.
- **Minería de datos**, uno de los elementos del DSS es el conocimiento el cual puede ser obtenido utilizando técnicas de minería de datos.
- **Cómputo móvil**, los tomadores de decisión deben de estar apoyados en lugares y momentos cambiantes.

El uso de dispositivos móviles dentro en el cuidado de la salud se ha ido incrementando a partir de los años 90, con el uso de una gran variedad de aplicaciones médicas como la de contener información de medicamentos y recomendación de medicamento, estos

dispositivos tienen un gran potencial para administrar el conocimiento en el cuidado de la salud (Fisher et al., 2003).

El conocimiento se está ampliando, los recursos económicos y las capacidades humanas para aplicarlo siguen siendo finitos. En la medicina cada vez se requieren más técnicas para tratar diversas enfermedades y mejorar la calidad de vida de las personas, aun los nuevos conocimientos no son diseminados rápidamente ni se utilizan con eficacia. En el área médica es necesaria la utilización de tecnologías de información para diseminar el conocimiento y propiciar una ayuda a la toma de decisiones en un contexto clínico (Fox, 2001).

Los CDSS requieren conocimiento, el cual es obtenido a partir de datos en los sistemas de información en hospitales. Grandes cantidades de información son almacenadas en bases de datos médicas, las cuales requieren herramientas especializadas para el almacenamiento, acceso, análisis y un uso efectivo de los datos (lavrač, 2001; Imhoff et al., 2000).

Los métodos tradicionales para el análisis de datos se han vuelto inadecuados para satisfacer las necesidades de la informática médica (Fox, 2001), en donde se deben de usar tecnologías con un enfoque interdisciplinario como es el descubrimiento del conocimiento en las bases de datos, (KDD-por sus siglas en inglés, Knowledge Discovery in Databases-).

El proceso de KDD consiste en lo siguiente, el primer paso es establecer el objetivo, luego seleccionamos el conjunto donde extraeremos el modelo predictivo, en el siguiente paso es la limpieza y transformación de los atributos en el cual buscamos realizar una vista minable, lo siguiente es realizar una análisis de exploración donde seleccionamos el método de minería de datos a utilizar, ya seleccionado el método el siguiente paso es obtener el modelo predictivo, posteriormente evaluamos y consolidamos este conocimiento (Fayyad et al., 1996). , este proceso es iterativo en el cual en el paso de la evaluación, puede haber un posible retorno.

Las grandes cantidades de información médica son recursos potenciales para descubrir nuevo conocimiento a través de técnicas de minería de datos. En la medicina se sabe (lavrač, 1999; lavrač, 2001) que las técnicas comúnmente usadas para la extracción de reglas son CN2 (Clark y Bosswell, 1991), ID3 (Quinlan J.R 1986) y *C4.5* (Quinlan J.R 1993), las cuales tienen la propiedad de representar el conocimiento en forma *si-entonces*, en cambio si se desea métodos más robustos como lo son las redes bayesianas, redes neuronales y los *k* vecinos más cercanos, pero carecen de la propiedad de representar las relaciones de las variables convirtiéndose en cajas negras. La minería de datos es un campo que incluye la estadística, visualización, máquina de conocimiento y KDD, que ayudan encontrar relaciones y patrones escondidos en los datos (lavrač, 2001)

El proceso del KDD, es una gran promesa para el desarrollo de bases de conocimiento requeridas en sistemas de apoyo a la toma de decisiones en un contexto clínico, Permitiendo que los datos “hablen” (Englebardt y Nelson, 2001).

La integración del proceso del KDD en las actividades del día a día se está convirtiendo en algo habitual. Los negocios de la distribución y la publicidad dirigida han sido tradicionalmente las áreas en las que más se han empleado el proceso del KDD ya que han permitido reducir costos o aumentar la receptividad de las ofertas. Existen muchas áreas en la cual se pueden encontrar diversas aplicaciones como lo son las financieras, científicas (medicina, farmacia, astronomía), políticas económicas, demográficas, educación procesos industriales.

Se han realizado diversos esfuerzos por hacer CDSS, uno de ellos lo presenta Mangalampalli et. al. (2006) el cual es un sistema basado en reglas especializado en ginecología y que utiliza el sistema médico homeopático, ellos utilizaron redes neuronales para la obtención de reglas.

Por otro lado Dahlstrom et. al. (2006) realizó un estudio con una perspectiva socio-técnica para identificar requerimientos sociales y organizacionales para la implementación

de un CDSS en una clínica de reumatología y que utilice técnicas de minería de datos para la obtención de conocimiento.

Como se puede observar en este punto del documento, los temas de interés de esta tesis son los siguientes: CDSS, sistemas de soporte a toma de decisiones médicas y técnicas de minería de datos para la extracción de conocimiento, en la siguiente sección se detallará sobre estos temas.

## **I.2. Planteamiento del problema**

Durante los últimos 30 años se han llevado a cabo numerosos intentos para desarrollar DSS que ayuden a los médicos a tomar decisiones sobre sus pacientes. Como menciona Mark (2002), en un prestigioso libro de medicina, desde el punto de vista conceptual las computadoras son muy atractivas por que pueden manejar fácilmente un cúmulo grande de información, para establecer predicciones precisas de los resultados, simular todo el proceso de decisión y aportar una guía algorítmica en los procesos de toma de decisiones clínicas. Entonces, a partir de este nicho de aplicación importante y ampliamente estudiado en este trabajo de tesis nos enfocaremos en los DSS en ambientes clínicos.

González et al. (2005) mencionan que para tomar decisiones los trabajadores del cuidado de la salud requieren de una correcta combinación entre gente, lugar, conocimiento y recursos que están distribuidos por toda la organización. Lo cual implica un amplio grado de movilidad y colaboración en el que se requiere de información y conocimiento relevante en el momento y el lugar oportuno para poder tomar decisiones.

Es necesario el desarrollo de un CDSS a partir de un estudio cualitativo para la obtención de requerimientos, con un enfoque socio-técnico y que se tome en consideración que los médicos son trabajadores de conocimiento con necesidades de movilidad.

Es necesario conocer la percepción que tienen los usuarios de sistemas móviles de CDSS basado en técnicas de minería de datos.

### **I.3. Objetivo general**

El objetivo general de este trabajo de tesis es **“explorar cómo las técnicas de minería de datos pueden apoyar a un proceso de toma de decisiones en un contexto clínico, a través del desarrollo y evaluación de un prototipo que utilice cómputo móvil”**. Para alcanzar esto se proponen los siguientes objetivos específicos.

#### **I.3.1. Objetivos específicos**

- Identificar un escenario clínico real donde se toman decisiones y que cuente con una memoria organizacional de tal forma que se puedan aplicar técnicas de minería de datos. Además de que en este escenario sea conveniente la utilización de cómputo móvil.
- Analizar el proceso de toma de decisiones del escenario clínico real.
- Modelar los resultados del análisis del escenario clínico desde una perspectiva socio-técnica, para recopilar requerimientos de un sistema tecnológico embebido en un contexto organizacional, que permita tomar decisiones en este contexto.
- Seleccionar y utilizar técnicas de minería de datos así como una metodología para la extracción de conocimiento en bases de datos para la obtención de modelos predictivos.
- Realizar un diseño de procesos donde se evalúen los requerimientos para el diseño e implementación de un sistema móvil de apoyo la toma de decisiones con el modelo predictivo obtenido de técnicas de minería de datos.
- Evaluar la percepción de utilidad que tendrán los usuarios al CDSS que estén basados en técnicas de minería de datos.

## **I.4. Contribución al conocimiento**

Dentro de la revisión de la literatura, vemos que existe la necesidad de una aplicación que fuese desarrollada a partir de un estudio cualitativo para la obtención de requerimientos, con un enfoque socio-técnico y que se tome en consideración que los médicos son trabajadores de conocimiento con necesidades de movilidad.

Esta investigación se enfoca especialmente en la evaluación de utilidad que tienen los usuarios de sistemas móviles de CDSS basados en técnicas de minería de datos.

Otras contribuciones de este trabajo es que se obtendrán lineamientos de diseño que pueden servir de guía/ayuda en el diseño y desarrollo de herramientas de CDSS que estén basados en reglas obtenidas de técnicas de minería de datos, así como un sistema que permita la implementación reglas en un dispositivo móvil.

## **I.5. Metodología de investigación**

Con el propósito de cumplir con los objetivos planteados, hemos definido una metodología de trabajo compuesta de cinco etapas con sus correspondientes actividades y productos. En la *Figura 1*. Se puede observar la metodología propuesta, y las salidas en cada una de las etapas.

La descripción de cada una de ellas se presenta a continuación.

### **I.5.1. Revisión de literatura**

En esta fase de estudio material bibliográfico respecto a los CDSS, proceso de extracción de conocimiento y técnicas de minería de datos y cómputo móvil aplicado al contexto médico. Con el objetivo de conocer y fundamentar ideas sobre las áreas estudiadas

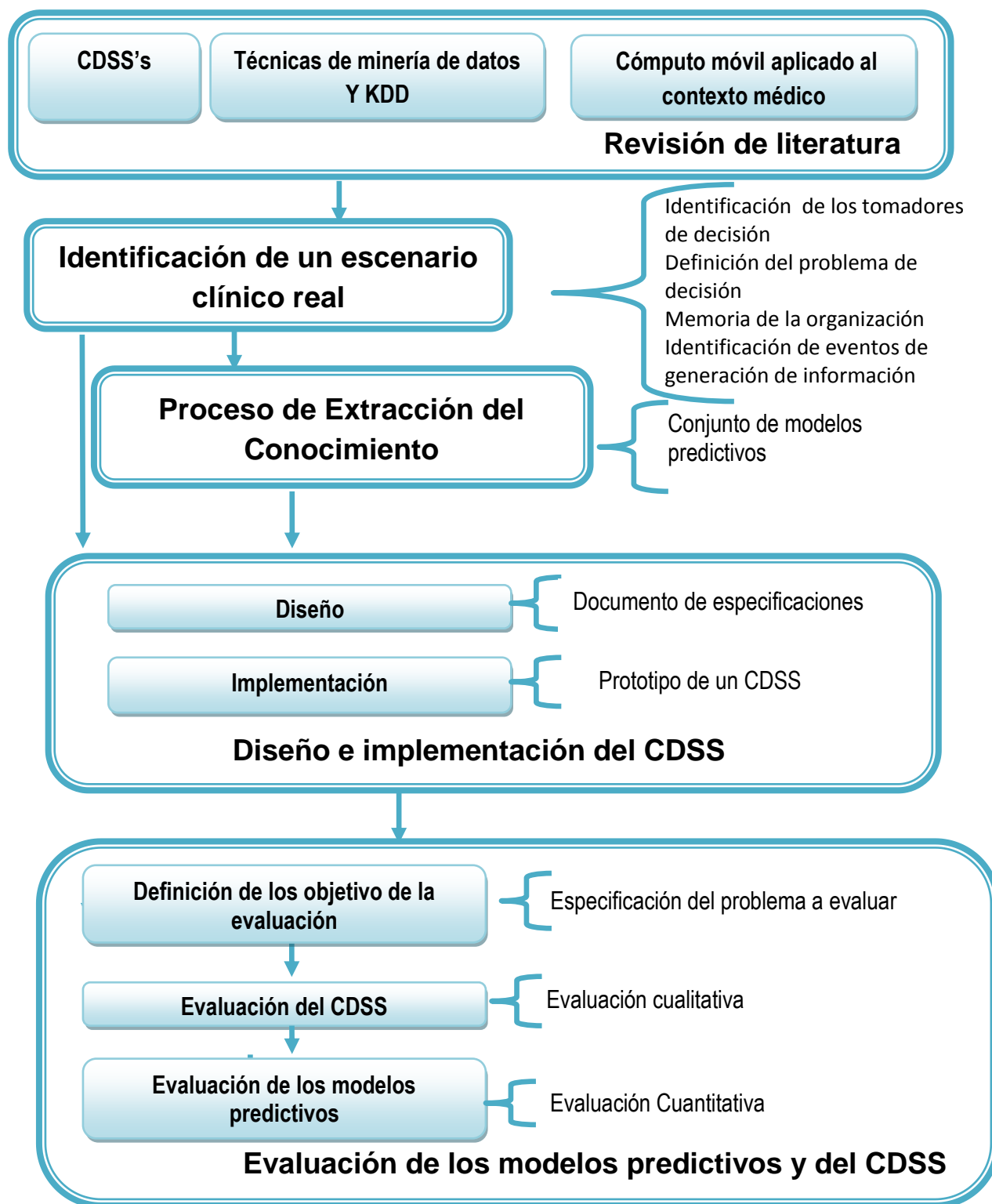


esto se puede ver en la sección II.5 donde se describe sobre los sistemas de ayuda a la toma de decisiones en el contexto medico, en la sección II.6 describe el computo móvil aplicado al contexto médico y en el capítulo III se describe el proceso de extracción de conocimiento y las técnicas de minería de datos para obtener modelos predictivos.

### **I.5.2. Identificación de un escenario clínico real**

Se identifico un escenario clínico real en el hospital con los siguientes requerimientos: se tomen decisiones, tengan una memoria organizacional y que tengan necesidades de movilidad, entonces el escenario que cumplía con estos requerimientos es el proceso de detección y vigilancia de infecciones nosocomiales (PDVIN), el primer paso fue la captura del proceso. En la captura del proceso se realizaron las siguientes actividades: 12 entrevistas, 4 sesiones de observación. Esto se ve reflejado en el Capítulo IV.

En base a esas actividades se obtuvo una descripción textual del proceso, con el cual se modelo con graficas ricas, diagramas de flujo de datos y diagramas de rol actividad. Con esos productos se identificaron al jefe de departamento de epidemiologia y a la epidemióloga como los tomadores de decisión, adicionalmente se identifico el formato de registro de caso de infección nosocomial (FRCIN) como la memoria organizacional la cual es la base para la extracción de conocimiento asimismo la aplicación del KDD y las técnicas de minería de datos para obtener modelos predictivos.



*Figura 1. Metodología de investigación.*

### **I.5.3. Uso de una metodología para la extracción del conocimiento en bases de datos**

Teniendo identificados la base de conocimiento y el caso de estudio se procedió a inicio con el proceso de extracción de conocimiento. El primer paso fue la definición del objetivo el cual es obtener modelos predictivos para la detección oportuna de infecciones nosocomiales. Con esto en claro, se procedió a realizar una selección inicial de atributos a partir de la FRCIN.

Con esta selección inicial se desarrollo una aplicación para la creación de una vista minable, posteriormente se realizo un análisis de exploración de los diversos métodos de minería de datos para el tratamiento de información médica en base a este análisis se eligieron los siguientes métodos: *Naive Bayes* y arboles de decisión (C4.5).

En base a esos métodos elegidos se inicio un proceso de trasformación y generalización de datos sobre la vista minable. Con la vista minable trasformada se aplicaron los métodos minería de datos para obtener modelos predictivos. Finalmente estos modelos de evaluaron para probar su robustez debido a que posteriormente serán implementados en la próxima fase de la metodología. El proceso de extracción de conocimiento se describe en el capítulo V.

### **I.5.4. Diseño e implementación del CDSS**

En esta fase de la metodología se dividirá en dos etapas las cuales se detallarán a continuación:

1. **Diseño del prototipo de CDSS:** Tomando como base el documento de requerimientos resultante de la fase de identificación de un escenario clínico real y los modelos predictivos resultantes de la anterior fase se procedió al diseño de un

prototipo de CDSS que cubriera con los requerimientos anteriores, dando como resultado un documento de especificaciones de los dos sistemas móvil y de escritorio el cual contiene los diseños arquitectónicos, funcionales y de comportamiento basados en el estándar UML.

- 2. Implementación:** La implementación se realizó mediante el documento de requerimientos, especificaciones y las reglas obtenidas del algoritmo de minería de datos seleccionado en el proceso del KDD en las etapas anteriores dando como resultado un prototipo para el sistema móvil y para la versión de escritorio debidamente depurados y probados.

Al finalizar esta etapa se procesará a la siguiente etapa la cual consistió en la evaluación del prototipo del CDSS. Estas etapas del proceso se ven reflejadas en el Capítulo VI.

### **I.5.5. Evaluación de los modelos predictivos y del CDSS**

A partir del diseño e implementación se evaluó la percepción de utilidad que tiene el usuario sobre las técnicas de minería implementadas en el sistema de apoyo a toma de decisiones en un contexto clínico. Asimismo se realizó una evaluación de los diferentes modelos predictivos con un conjunto de datos que no se usó para la obtención del modelo predictivo. Para realizar esta evaluación, se propone utilizar la siguiente metodología:

**Definición de objetivos de la evaluación.** En esta etapa se definirán los objetivos de esta evaluación.

**Evaluación de los modelos.** En esta fase se probará la robustez de los modelos obtenidos en términos del porcentaje de instancias clasificadas correctamente, precisión y alcance de cada una de las clases.

**Evaluación del CDSS.** Con los objetivos de la evaluación definidos se procedió a realizar las actividades directamente con los usuarios del CDSS para ver cuales la percepción de utilidad que tienen los médicos tomadores de decisión, en esta evaluación se obtuvieron resultados cualitativos.

## **I.6. Contenido del documento de tesis**

En esta sección se detalla el contenido y organización de este trabajo. En este primer capítulo se presenta una introducción general para familiarizar al lector con el problema abordado, con los objetivos del mismo y la metodología.

En el capítulo II se explicarían los conceptos de los sobre teoría de decisión, sistemas de ayuda a la toma de decisiones y los sistemas de ayuda en la toma de decisiones en el contexto médico a si como el trabajo previo de estos sistemas.

En el capítulo III se explican el proceso de extracción de conomiendo donde se hace énfasis en el paso de minería de datos. En esa sección se explica los métodos para la extracción de conocimiento y obtención de modelos predictivos, para finalizar el Capítulo se explican las medidas de evaluación de los modelos.

En el capítulo IV describe el caso de estudio sobre infecciones nosocomiales realizado en un hospital público de la localidad donde se realizo la captura del PDVIN para posteriormente modelar el proceso y obtener los requerimientos con una perspectiva socio-técnica.

En el capítulo V. se describe el proceso de extracción de conocimiento para la obtención de modelos predictivos, para la detección oportuna de infección nosocomial.

En el capítulo VI se muestran las características de diseño del sistema de apoyo a toma de decisiones clínicas propuesto a partir de los requerimientos obtenidos en el caso de estudio, donde se detalla la funcionalidad requerida, el diseño arquitectónico del sistema el de comportamiento y la funcionalidad más relevantes del prototipo. Asimismo se describe la implementación del prototipo funcional desarrollado en base al diseño del CDSS.

En el capítulo VII se describe el diseño del experimento realizado con el fin de obtener la percepción de utilidad asimismo se desarrollo una evaluación de los modelos predictivos para ver que tan robustos son sobre un conjunto de datos que no se utilizaron para el entrenamiento.

Finalmente en el capítulo VIII se describen las conclusiones y aportaciones de este trabajo de tesis, se menciona el trabajo a futuro que se propone para continuar con esta investigación.

## Capítulo II

---

### Sistemas de ayuda a la toma de decisiones

---

#### II.1. Introducción

En este capítulo se presentan conceptos del marco teórico y tecnología para el apoyo a la toma de decisiones en la que se basa este trabajo de tesis. El objetivo es entonces de introducir conceptos necesarios, así como enfocarnos en los aspectos teóricos que estamos interesados en explorar en esta tesis.

Primero presentamos una breve introducción a la toma de decisiones y clasificamos los problemas de decisión, enseguida definimos la toma de decisiones como un proceso organizacional y los sistemas de cómputo que dan soporte a la toma de decisiones, después, mostraremos un tipo de especial de sistema de toma de decisiones en el contexto clínico. Y para finalizar se presenta el trabajo previo relevante sobre los sistemas de toma de decisiones en el contexto médico.

#### II.2. Toma de decisiones

La toma de decisiones es parte central en toda actividad humana, donde el concepto más cercano a éste es el pensamiento (Patel et al, 2002). Por ejemplo cuando queremos comprar algún artículo y tenemos opciones de compra, intuitivamente pensamos cual es el artículo que mejor satisface nuestras necesidades. Según Marakas (2003) una decisión que puede hacerse sin razonamiento no es en realidad una decisión, es decir, que la toma de decisiones implica un cierto grado de razonamiento. Sin embargo, para tomar una decisión no sólo es necesario el razonamiento, debido a que ésta no sólo implica hacer una elección en base a

un conocimiento previamente adquirido, sino también involucrarse con el ambiente para de éste modo adquirir y distinguir el conocimiento útil para tomar una mejor decisión.

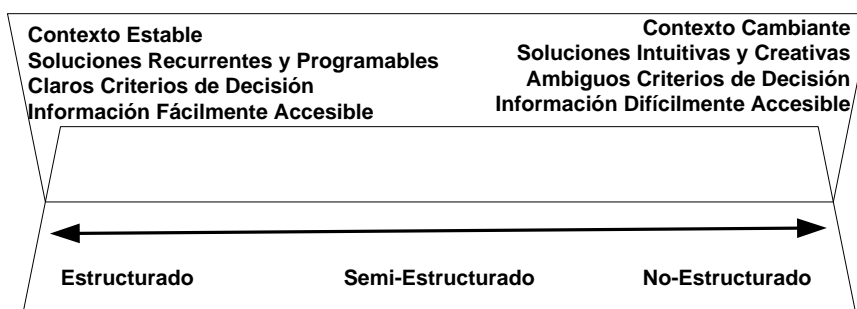
Por ejemplo para adquirir un producto, típicamente se buscan varias tiendas o incluso marcas para ver quien ofrece el mejor precio o servicio. En el caso de comprar vía Internet, incluso podemos consultar las experiencias de otros clientes antes de comprar el producto en el estamos interesados.

Tomar una decisión es, a partir de un conocimiento, analizar en un problema las opciones disponibles dentro de una situación específica, para discernir que es conveniente hacer (Ofstad, 1961). En este sentido, la toma de decisiones no solo implica una perspectiva técnica, en base a métodos racionales y lógicos, donde se analizan cuantitativamente las opciones, sino también implica una perspectiva social y organizacional, basada en modelos de interacción y organización social, donde se analizan cualitativamente estas opciones.

Los problemas de decisión son algo cotidianos, sin pensar tomamos decisiones rutinarias, como por ejemplo qué canal de televisión ver, como vestirse, etc. Sin embargo, también existen problemas de decisión que no son rutinarios. Por ejemplo dentro de un contexto de hospital, cuando se decide el tratamiento de un paciente, implica consecuencias serias por lo cual se deben analizar cuidadosamente las alternativas relevantes al problema de decisión.

Gorry y Scott-Morton (1971), se basaron en el marco de trabajo definido por Simon (1960) que clasificaba los problemas en programables y no programables para proponer un marco de trabajo para clasificar los problemas de decisión en tres tipos: estructurados, no estructurados y semi-estructurados, los cuales se muestra en la figura 1 y se describen brevemente a continuación (Marakas, 2003):



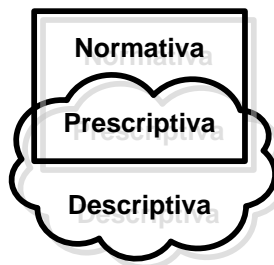


*Figura 2. Clasificación continua de los problemas de decisión, propuesta por Gory y Scott-Morton (1971), extendida de la clasificación discreta de Simon (1960)*

- **Estructurados.** Se refieren a aquellos problemas recurrentes, para los cuales las soluciones son programables, debido a que las decisiones se toman bajo criterios claros y contextos estables. Problemas donde la información y conocimiento necesario para tomar la decisión es fácilmente accesible. Por ejemplo, decidir el nivel de inventario de una fábrica en la cual se puede programar la producción y saber exactamente el nivel de inventario necesario para trabajar.
- **No estructurados.** Estos problemas en cambio son complejos, que requieren soluciones intuitivas y creativas que no pueden ser programables, debido a que las decisiones se toman bajo criterios ambiguos y en contextos cambiantes. Por ejemplo, decidir sobre la foto de la portada de una revista, donde se necesita de la experiencia, intuición y creatividad del editor para decidir que foto poner en la portada de las alternativas disponibles.
- **Semi-estructurados.** Poseen características tanto estructuradas como no estructuradas. En donde es fácilmente accesible parte de la información y conocimiento necesario para tomar la decisión, pero la otra parte requiere intuición y creatividad humana para obtenerla. Por ejemplo, el diagnóstico de una enfermedad, donde se tienen indicadores físicos o de laboratorio que dan pistas a los médicos, sin embargo, cada paciente es diferente y reacciona en forma diferente ante una misma enfermedad. Por este motivo los médicos tienen que hacer uso de su intuición y razonamiento para determinar el diagnóstico de una enfermedad.

## II.2.1. Perspectivas de toma de decisiones

La toma de decisiones es importante porque mediante un buen juicio, se puede elegir la solución más apropiada al problema que se presente. En muchos casos se presentan múltiples soluciones viables para un mismo problema, pero ¿cómo determinar cuál es la solución más apropiada?; dependiendo si el problema de decisión es estructurado, no estructurado o semi-estructurado, la toma de decisiones se puede clasificar a través de la forma en que son tomadas las decisiones como: normativa, descriptiva y prescriptiva (Dillon, 1998), como se muestra en la *figura 3* y se describe brevemente a continuación:



*Figura 3. Perspectivas de toma de decisiones*

- **Normativa**, describe la forma en que deberían tomarse las decisiones, en base a disciplinas como economía, estadística, matemáticas, e investigación de operaciones. Esta clasificación se refiere a la perspectiva técnica, ya que se basa en métodos analíticos y cuantitativos, útiles para problemas de decisión estructurados.
- **Descriptiva**, se refiere a la forma en que se toman las decisiones, en base a disciplinas como psicología, sociología, teoría organizacional y ciencias del comportamiento. Esta clasificación se refiere a la perspectiva social y organizacional, debido a que se basa en métodos cualitativos, los cuales son útiles para problemas de decisión no estructurados.
- **Prescriptiva**, se da como resultado de la integración de las dos primeras clasificaciones. Describe la forma en que se pueden y deben tomar las

decisiones, en base a todas las disciplinas antes mencionadas. Esta última clasificación está enfocada en una perspectiva socio-técnica, la cual es útil en problemas semi-estructurados.

En esta sección definimos la toma de decisiones y clasificamos los problemas de decisión según sus características, para poder establecer las perspectivas según la forma en que pueden ser abordados los problemas de decisión. Enseguida analizaremos el proceso de toma de decisiones, el cual nos ayudará a entender cómo es que se conforma una decisión, pero tomando en cuenta las perspectivas de toma de decisiones mencionadas en esta sección.

### **II.3. Proceso de toma de decisiones**

Definir un proceso para llevar a cabo la actividad de toma de decisiones es importante, porque de esta forma la toma de decisiones tiene una estructura, ya que algunas veces las decisiones se toman sin tenerla, simplemente porque el tomador de decisiones confía en sus habilidades, llegando a resultados inconsistentes (DeSanctis y Gallupe, 1987). Como se mencionó anteriormente la toma de decisiones puede realizarse bajo una perspectiva normativa, descriptiva o prescriptiva, dependiendo del contexto en el que el problema de decisión se presenta. Enseguida analizaremos el proceso de toma de decisiones primero cuando se presentan problemas de decisión, no estructurados y después para problemas de decisión semi-estructurados.

#### **II.3.1. Proceso de toma de decisiones en un ciclo de inteligencia organizacional**

El proceso de toma de decisiones no es una actividad que se realiza en un solo momento, si no que es un proceso reflexivo que implica conocimiento y experiencia de los tomadores de decisión (Simon, 1960). En las organizaciones el conocimiento juega un rol relevante

para el éxito, donde vale más una gota de conocimiento que un mar de información. El conocimiento involucra creatividad, juicio, habilidades y experiencia de las personas que le dan significado a la información que manejan. Por lo que, para que el proceso de toma de decisiones satisfaga los objetivos organizacionales con éxito, se requiere del apoyo de otras actividades relacionadas con el conocimiento organizacional. Choo (1999) distingue tres procesos estratégicos dentro de las organizaciones que definen el ciclo de inteligencia organizacional (véase *Figura 4*) involucrado en la toma de decisiones, estos son percepción, creación de conocimiento y toma de decisiones, los cuales describimos brevemente a continuación:



*Figura 4. Ciclo de inteligencia organizacional.*

- **Percibir el ambiente.** En este proceso se busca que los miembros lleguen a una concepción compartida de lo que es la organización, para que estos estén consientes de su ambiente y puedan adaptarse a los cambios que se presenten en el mismo. Típicamente resulta complicado identificar los cambios en el medio ambiente por ser ambiguos y estar sujetos a complicaciones, pero a partir de la experiencia se pueden predecir dichos cambios.
- **Crear conocimiento.** En este proceso los miembros de la organización crean, organizan y procesan información para generar nuevo conocimiento. El reto es

que el conocimiento y la experiencia están dispersos por toda la organización, con frecuencia lo poseen individuos o grupos de trabajo. Por un lado es difícil aprovechar el conocimiento y por otro es también difícil cuestionar las prácticas existentes y proponer nuevas alternativas.

- **Tomar decisiones.** En este proceso los miembros de la organización realizan una selección racional de las opciones factibles, resultados probables y examinan las consecuencias para la organización, en base en los objetivos organizacionales. El reto es que no dominen intereses individuales o de grupo sobre los objetivos organizacionales. La organización tienen que mantener una apariencia de conducta racional, en base a la confianza interna y la legitimidad externa.

Según Simon (1960) en el proceso de toma de decisiones, primero se debe buscar la ocasión para tomar la decisión, después buscar las posibles alternativas de acción y finalmente seleccionar la alternativa de acción más apropiada. Pero según Choo (1999) es fundamental que estos pasos ocurran en base al aprendizaje de las experiencias de la organización.

Esta clasificación es importante dado que dependiendo de las perspectivas decisión podemos determinar las herramientas tecnológicas apropiadas para el apoyo a la toma de decisiones.

#### **II.4. Sistemas de apoyo a la toma de decisiones**

La toma de decisiones es una actividad compleja debido a que involucra a personas que tienen diferentes roles (conjunto de actividades que son asignadas a humanos o máquinas – que realizan un proceso elemental-, como una unidad de responsabilidad funcional) y por consecuencia diferentes perspectivas del problema, incertidumbre y riesgos al tomar las decisiones, pérdidas y ganancias como consecuencia de la decisión, tiempos limitados en realizarlas y factores subjetivos involucrados (Pacheco-Soto, 2004).

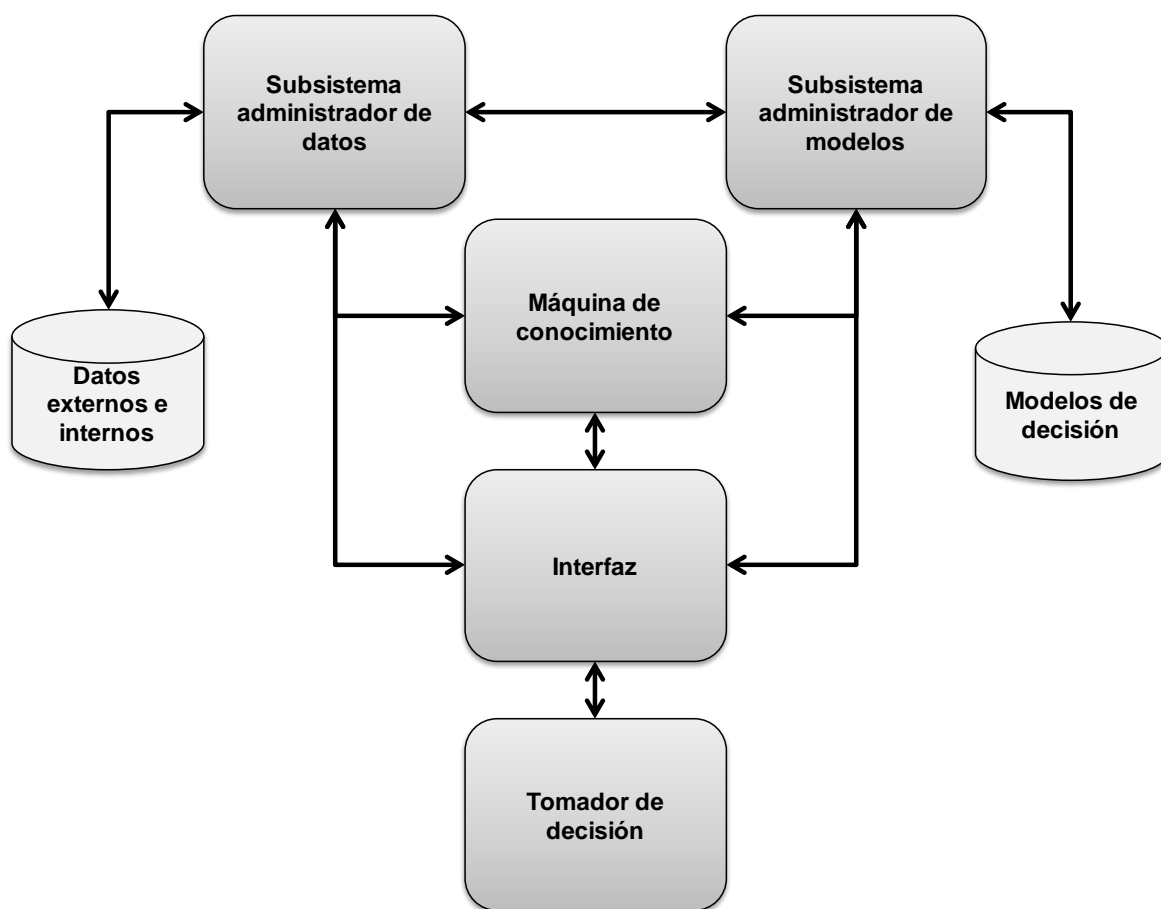
Buscando facilitar el manejo de esta complejidad surgen los sistemas de ayuda a la toma de decisiones (en lo sucesivo, DSS –por sus siglas en inglés *Decision Support System*-), los cuales han evolucionado significativamente desde los principios de 1970. A continuación se presenta una definición. “*Un DSS es un sistema basado en computadora, el cual soporta el proceso de toma de decisiones, al ayudar a los tomadores de decisiones a formar y explorar las implicaciones de sus juicios y de esta forma, tomar decisiones basadas en el entendimiento*” (French, 1998)

Los DSS han evolucionado desde entonces para convertirse en “*sistemas de decisión, que asisten en las actividades de toma de decisión proveyendo un conjunto organizado de herramientas que pretenden imponer una estructura en una porción del problema, con el fin de mejorar la efectividad de la decisión*” (Marakas, 2003).

Los DSS tienen el beneficio de que extienden las habilidades de los tomadores de decisión para procesar la información y conocimiento. Adicionalmente hacen posible afrontar problemas de mayor complejidad; acortan el tiempo con el que se toma la decisión; mejoran la confiabilidad del proceso de decisión y por lo tanto la decisión que se toma. También ayudan a introducir nuevas estrategias para pensar en un problema o contexto de decisión, creando ventajas estratégicas o competitivas contra otras organizaciones. Finalmente generan evidencia para apoyar una decisión o confirmar una conjetura existente (Marakas, 2003).

Los DSS, también tienen limitaciones como el que no pueden ser diseñados para incluir habilidades de los tomadores de decisión humanos tales como la creatividad, imaginación e intuición. El poder de los DSS, está limitado por el poder de procesamiento de los dispositivos de cómputo. Los DSS, no pueden evitar las acciones de malos tomadores de decisiones; adicionalmente los DSS, no son universales, típicamente están diseñados para enfocarse en un escenario de aplicación en particular, por lo que es preciso la integración de múltiples DSS para enfrentar escenarios de decisión complejos (Marakas, 2003).

En otro sentido los DSS comparten ciertas características genéricas. Típicamente están formados por cinco elementos básicos: administración de datos, administración del modelo, máquinas de conocimiento, interfaz de usuario y el mismo tomador de decisión (Marakas, 2003). Los elementos involucrados en los DSS están relacionados con las capacidades que se requieren en un contexto de decisión particular, es decir, en ocasiones podría los DSS no tener uno o varios elementos (exceptuando el tomador de decisión y la interfaz) para una aplicación en particular. Enseguida en la *Figura 5* se muestra el flujo de interacción de los elementos que conforman los DSS, dando después una breve descripción de los mismos.



*Figura 5. Elementos de los DSS.*

- **Subsistema de administración de datos.** Tiene la función de recuperar, almacenar y organizar los datos relevantes en un contexto de decisión en particular. Este subsistema le proporciona de forma transparente los datos necesarios al subsistema administrador de modelos, a la máquina de conocimiento o a la interfaz.
- **Sistema de administración de modelo.** Ofrece las capacidades analíticas, en base a modelos cualitativos o cuantitativos. Este se encarga de administrar transparentemente estos modelos, de ser necesario le solicita datos al subsistema de administración de datos. Sin embargo, típicamente mediante la máquina de conocimiento este elemento entrega los modelos de decisión para que ahí se mezclen los datos con los modelos. O bien, este elemento le proporciona los modelos de decisión al tomador de decisión a través de la interfaz.
- **Máquina de conocimiento.** Aquí se mezclan los datos y los modelos para llevar a cabo el reconocimiento del problema y la generación de soluciones intermedias o finales. En principio este elemento es el que brinda a los tomadores de decisión mediante la interfaz el apoyo al contexto de decisión. Es así que este elemento debe incluirse en un DSS, porque es el corazón de este, debido a que este elemento tiene interacción con la mayoría de los otros elementos.
- **Interfaz (diálogo-comunicación).** Proporciona el medio para que el tomador de decisión controle y maneje fácilmente los DSS y finalmente obtenga y visualice el resultado o recomendación del apoyo a la decisión.
- **Tomador de decisión.** Es una parte integral de los DSS, debido a que para manejar los DSS es de gran importancia las habilidades, conocimientos, motivaciones, patrones de uso y roles en la organización del tomador o tomadores de decisión.

Como hemos mencionando anteriormente, el tomador de decisión es un elemento más de los DSS. Más de una persona puede participar controlado, manejado o aportando al conocimiento a los DSS. Enseguida presentaremos, las tecnologías disponibles para proporcionar apoyo múltiples participantes en la toma de decisión.



La tendencia actual en la investigación y desarrollo de los DSS según Shim et al. (2002) incluyen los siguientes elementos: perspectiva socio-técnica, minería de datos, cómputo ubicuo, múltiples tomadores de decisión, proporcionar servicios de aplicaciones, habilidad de almacenar y analizar múltiples escenarios de solución y el uso de agentes inteligentes.

En esta propuesta de tesis se utilizara la perspectiva socio-técnica ya que el análisis y diseño de un DSS debe de tener en consideración al usuario. La minería de datos utiliza técnicas muy robustas para extraer patrones o reglas las cuales pueden ser implementadas en los DSS's, el uso de dispositivos móviles cubre la necesidad que tienen los usuarios de tomar decisiones en cualquier momento y lugar, a continuación se detallan estas tres tendencias:

**Perspectiva socio-técnica.** En el pasado la perspectiva técnica de desarrollo de bases de datos y modelos dominaba la formación de problemas de decisión, pero esto resultó insuficiente para obtener decisiones eficientes, por que los médicos no encontraban útiles los desarrollos en base a perspectivas netamente técnicas o incluso les perjudicaba como menciona Laplante et al. (2006). Actualmente las perspectivas personal y organizacional ayudan cualitativamente a formular problemas que incluyen factores no cuantificables que afectan en gran medida la funcionalidad de los DSS.

**Minería de datos.** Para aprovechar el conocimiento organizacional registrado en las bases de datos, los trabajadores del conocimiento requieren herramientas de minería de datos. Estas se caracterizan por buscar anomalías sin estar consientes del significado de los datos, para encontrar información y conocimiento, a partir de algoritmos de inteligencia artificial y estadística. La aplicación inteligente de la minería de datos juega un rol significativo para ayudar a las organizaciones en trasformar grandes cantidades de datos en un valioso conocimiento corporativo e inteligente.

**Cómputo ubicuo.** Con el desarrollo de dispositivos móviles y las redes inalámbricas, en este ambiente, los compañeros de trabajo pueden verdaderamente colaborar en cualquier lugar y en cualquier tiempo, sin la necesidad de estar físicamente atado a una computadora. Esto asegura la conectividad de los miembros del grupo de trabajo y equipos virtuales los cuales pueden tener un acceso a una mayor robustez en el soporte a la decisión. Otro beneficio de la interactividad inalámbrica es la mejora de la habilidad de los trabajadores de conocimiento para recolectar múltiples perspectivas de los problemas de decisión.

El reto actual de los DSS en las organizaciones es asistir a la persona que lo requiere, en lugar y momento apropiado y de forma explícita. Un tipo de DSS son los CDSS, los cuales explicaremos en la siguiente sección.

## **II.5. Sistemas de apoyo a la toma de decisiones en un contexto clínico**

La mejora del proceso de cuidado de salud depende de una óptima decisión clínica. Las decisiones en un ambiente clínico pueden ser realizadas por sistemas que faciliten, incrementen o expandan la habilidad clínica para trabajar con datos e información. En el sentido ideal, *“un sistema de apoyo a toma de decisiones clínicas, es un conjunto de bases de conocimientos que ofrecen herramientas para una completa integración de componentes de flujo de trabajo de un registro computarizado de pacientes y un depósito completo y exacto de información clínica”* (Perreault y Metzger, 1999).

Una definición más moderna es la que brinda Osherooff (2004), en la cual los CDSS son vistos como cualquier sistema que *“proveen a los trabajadores del cuidado de salud, conocimiento clínico e información relacionada con el estado del salud del paciente, filtrada en forma inteligente o presentada en un momento adecuado para mejorar el cuidado del paciente”*. Desde la perspectiva médica, de hecho cualquier sistema que presenta información sobre medicamentos o cualquier otra información del paciente son

vistos como un CDSS, los cuales representan un gran apoyo para la práctica médica, pero en sentido estricto no es un DSS.

Un CDSS imita el proceso humano de toma de decisiones. Los humanos observan los aspectos de la salud (identificando patrones), observación de patrones para generar nuevos patrones (reconocimiento de patrones) y posteriormente hacen un plan de acción basado en el diagnóstico (interpretación del patrón). Este proceso ilustra que puede ser usado para desarrollar un CDSS en el cuidado de la salud (Englebardt y Nelson, 2001).

El reto de resolver el problema de la toma de decisiones al cuidado de la salud cae en la investigación y desarrollo de sistemas basados en conocimiento, que incrementan la habilidad humana de analizar, resolver problemas, curar, diagnosticar y estimar pronósticos relacionados con la condición humana (Englebardt y Nelson, 2001).

Los CDSS no son nuevos en el cuidado de la salud, en realidad la primera evidencia de aplicaciones de CDSS fue en 1970, existen muchas razones por las cuales no han conseguido una amplia aceptación en el cuidado de la salud, en (Englebardt y Nelson, 2001; Kaplan, 2001) mencionan algunas de ellas:

- Estrecho alcance de las aplicaciones.
- Desconfianza por las decisiones clínicas hechas por el sistema.
- Inhabilidad para incorporar nuevos descubrimientos.
- Falta de integración con sistemas existentes.

Mucha de la renuencia a usar CDSS es que no encaja en la forma natural del proceso de cuidado y requiere esfuerzo adicional para individuos que están muy ocupados (Coiera, 2003).

El reto a afrontar de las tecnologías de información es como incorporar la ciencia de la toma de decisiones en sistemas que sean exactos, descriptivos, discretos y que se usen como parte de una actividad clínica cotidiana (Englebardt y Nelson. 2001).

Tipos de sistemas de ayuda a toma de decisiones en un contexto clínico.

Coiera (2003) nos presenta que existen muchos tipos diferentes de tareas clínicas en donde los CDSS pueden ser aplicados. Los cuales son descritos a continuación:

- **Recordatorios y alertas.** Estos sistemas son usados para notificar al personal clínico tareas importantes que necesitan estar hechas antes de que los eventos ocurran.
- **Asistencia al diagnóstico.** Cuando el paciente es un caso complejo, raro o la persona que está haciendo el diagnóstico es simplemente inexperta, un sistema experto puede ayudar en el diagnóstico basado en los datos del paciente.
- **Terapia crítica y planeación.** Los sistemas de terapia crítica pueden mirar inconsistencias, errores y omisiones en el plan de tratamiento, pero no asisten en la generación de un plan. Por otro lado los sistemas de planeación, que tienen mayor conocimiento acerca de la estructura del protocolo de tratamiento, pueden ser usados como base para diseñar el tratamiento, en base a un registro electrónico con la información específica de la condición del paciente.
- **Sistemas de ayuda para la toma de decisiones en la prescripción médica.** Una de las tareas más comunes en el contexto clínico es la prescripción de medicamento, y los PDSS (por sus siglas en inglés, *Prescribing Decision Support System*), pueden asistir en comprobar las interacciones entre las drogas, error de dosis, y si está conectada a un registro electrónico de pacientes, puede prescribir contradicciones como alergias.
- **Reconocimiento de imágenes e interpretación.** Muchas imágenes clínicas pueden ser interpretadas automáticamente, desde radiografías a imágenes más complejas.

- **Sistemas de información expertos en laboratorio.** Una de las más exitosas áreas en donde los sistemas expertos son aplicados es en los laboratorios clínicos. Los practicantes pueden pasar por alto una patología que ha sido revisada en un reporte impreso de laboratorio, en cambio sistema de cómputo puede generar un reporte completo y automáticamente interpretar los resultados.

Como se puede observar hasta este punto existen muchos tipos y aplicaciones para un CDSS dentro de las diferentes aéreas y servicios de un hospital. En la siguiente sección se describirá el computo móvil aplicado al contexto médico.

## **II.6. Cómputo móvil aplicado al contexto médico**

Los profesionistas del cuidado de la salud se enfrentan continuamente al reto de lograr una correcta combinación de gente, lugar, conocimiento y recursos para poder llevar a cabo sus responsabilidades (Bardram y Bossen. 2005). Los profesionistas del cuidado de la salud en hospitales involucran un alto grado de movilidad, debido a que los pacientes, equipo, recursos y gente están distribuidos en diferentes localizaciones físicas (Gonzáles et al, 2005).

Un estudio etnográfico realizado por Gonzáles et al. (2005) con el objetivo de entender el comportamiento móvil en los hospitales, donde encontraron que los médicos pasan en movimiento el 48% de su tiempo. Actualmente la tecnología del cómputo móvil ha tenido avances que han ido ofreciendo cada vez mayor poder de cómputo en dispositivos realmente portátiles. Un ejemplo de esto son los Asistentes Personales Digitales (PDA, por sus siglas en ingles *-Personal Digital Asistant-*), conocidas también como agendas electrónicas.

En la siguiente sección mostraremos el trabajo previo que hemos identificado en la bibliografía y que es relevante para este trabajo de tesis.

## II.7. Trabajo previo

En este trabajo de tesis consideremos del trabajo previo los siguientes aspectos: CDSS basados en reglas obtenidas a partir de técnicas de minería de datos, estudios de requerimientos bajo una socio-técnica y sistemas móviles para la ayuda de toma de decisiones.

En (Berlin, 2006), realizaron una clasificación de diferentes sistemas de CDSS así como sus tipos de diseño, función y uso. Estos sistemas que funcionan en un ambiente complejo como es el cuidado de la salud, donde el desafío es diseñar, implementar y evaluar estos sistemas.

Se han realizado diversas evaluaciones a los CDSS, en base a experimentos de laboratorio o ensayos clínicos aleatorios (RCTs por sus siglas en inglés *Randomized Controlled Clinical Trials*) para establecer que tan bien realizan su trabajo los sistemas o los médicos en un ambiente controlado. Otros enfoques para la evaluación de CDSS, son los estudios de campo, simulación, pruebas de productividad, estudios cognitivos, técnicas de grabación de video y sonido y análisis socio técnico que raramente se observa en la literatura (Kaplan, 2001).

Los tipos de evaluación que se han identificado son principalmente dos tipos. El primero estudia la exactitud y el rendimiento del CDSS, los cuales tienen el objetivo de validar el conocimiento de los CDSS. El segundo tipo de evaluación concierne directamente al cuidado del paciente, los cuales son sistemas de recordatorios, alertas y diagnóstico (Kaplan, 2001).

Se han realizado esfuerzos por tratar de realizar una taxonomía de los diversos CDSS. Para realizar esta taxonomía se estudiaron 58 CDSS y los clasificaron de acuerdo a 26 parámetros, de los cuales sólo se mencionarán algunos de ellos: contexto, conocimiento y fuente de información, método de razonamiento, soporte a la decisión, reparto de la

información y flujo de trabajo. Ellos muestran que el diseño de un CDSS es altamente variable en cuanto su función y uso, de acuerdo a lo anterior el 76% de los sistemas CDSS están basados en reglas, otros se basan en redes neuronales (3%), modelos probabilísticas (3%) (Berlin et al., 2006).

Un CDSS basado en reglas, especializado en ginecología y que utiliza el sistema médico homeopático, para desarrollar este proyecto utilizaron redes neuronales y sistemas expertos, basándose para el diagnóstico clínico los síntomas primarios y secundarios del paciente (Mangalampalli et al., 2006). Sin embargo, los diseñadores no tomaron en consideración las necesidades de movilidad que tienen los profesionistas del cuidado de la salud, tampoco muestra un estudio para conocer cómo las decisiones son tomadas a fin de obtener los requerimientos con una perspectiva socio-técnica, se puede observar la necesidad de realizar la evaluación del sistema para conocer la aceptación que obtuvo por parte de los médicos.

Un trabajo realizado bajo una perspectiva socio técnica es el de Dahlstrom et al (2006), donde llevaron a cabo un estudio para identificar los requerimientos sociales y organizacionales para implementación de un CDSS que utiliza técnicas de minería de datos para una clínica de reumatología. Este es un trabajo que utiliza técnicas cualitativas como son la observación, estudio de sombra, cuestionarios y entrevistas. A pesar que realizaron toda la fase de análisis de requerimientos en base a técnicas cualitativas les faltaron las fases de diseño y reestructuración del proceso en base a la implementación del CDSS, así como una evaluación para demostrar que el CDSS cumple con los requisitos propuestos.

El reto en cuanto al desarrollo de sistemas de toma de decisiones al cuidado de la salud cae en la investigación y desarrollo de sistemas basados en conocimiento, que incrementan la habilidad humana de analizar, resolver problemas, curar, diagnosticar y estimar pronósticos relacionados con la condición humana (Englehardt y Nelson, 2001).

Los CDSS no son nuevos en el cuidado de la salud, en realidad la primera evidencia de aplicaciones de CDSS fue en 1970, existen muchas razones por las cuales no han conseguido una amplia aceptación en el cuidado de la salud, en (Englehardt y Nelson, 2001; Kaplan, 2001) mencionan algunas de ellas: estrecho alcance de las aplicaciones, desconfianza por las decisiones clínicas hechas por el sistema, inhabilidad para incorporar nuevos descubrimientos y falta de integración con sistemas existentes.

Mucha de la renuencia a usar CDSS es que no encajan en la forma natural del proceso de cuidado y requiere esfuerzo adicional para individuos que están muy ocupados (Coiera, 2003).

El reto a afrontar de las tecnologías de información es como incorporar la ciencia de la toma de decisiones en sistemas que sean exactos, descriptivos, discretos y que se usen como parte de una actividad clínica cotidiana (Englehardt y Nelson. 2001). Debido a que los profesionistas del cuidado de la salud son altamente dinámicos y con necesidades de movilidad, es necesario introducir estos dispositivos que brinden soporte a sus necesidades. A continuación se hablara del trabajo previo de este tipo de sistemas.

En la Universidad de Ottawa, Canadá utilizaron un CDSS basados en cómputo móvil. Ellos justificaron el uso de cómputo móvil en la necesidad de seguir a los trabajadores en el lugar donde ellos toman las decisiones. Es para protocolos de emergencia abdominales en una sala de emergencias de un hospital pediátrico. Se basa en seguir un conjunto de reglas de decisión para llegar a determinar qué hacer con el paciente, ya sea enviarlo a cirugía o dejarlo en observación o darlo de alta, utilizando para ello información histórica del hospital (Michalowski et al., 2003). Sin embargo no realiza una evaluación cualitativa para conocer el proceso de adopción del usuario final así como un levantamiento de requerimientos con una perspectiva socio-técnica.



El fundamento de cualquier CDSS es el conocimiento médico contenido en un conjunto de reglas y hechos, este conocimiento debe de ser adquirido a partir de la información a continuación se explicara como el conocimiento es adquirido a partir de bases de datos.

## **II.8. Resumen**

En este capítulo se presentaron conceptos sobre la toma de decisiones así como de DSS en esa sección se describen las tendencias actuales en la investigación y desarrollo de estos sistemas adicionalmente se describen los CDSS's y una extensión de ellos los cuales exploran la movilidad.

Para finalizar este capítulo se describieron de manera breve el trabajo previo relacionado con los CDSS's y los CDSS's móviles. El siguiente capítulo describe el proceso de extracción de conocimiento en bases de datos así como las técnicas de minería de datos para la obtención de modelos predictivos.

**Descubrimiento de conocimiento en bases de datos**

---

**III.1. Introducción**

En este capítulo presenta conceptos para entender el proceso de extracción de conocimiento, minería de datos y métodos de minería de datos con el objetivo de introducir conceptos necesarios que estamos explorando en esta tesis.

Primero se presenta una introducción al proceso de descubrimiento de conocimiento, enseguida se mostrara sus diferentes pasos de los cuales daremos especial énfasis al de minería de datos y sus técnicas de extracción de conocimiento.

**III.2. Descubrimiento de conocimiento en bases de datos**

Una gran amplia variedad de campos de información está siendo acumulada a un ritmo dramático. Existe la urgente necesidad de teorías computacionales y herramientas que asistan a los humanos en la extracción de información útil (conocimiento) de forma rápida desde grandes volúmenes de información digital. Estas teorías y herramientas son tema de un campo llamado descubrimiento de conocimiento de bases de datos (KDD-por sus siglas en inglés, Knowledge Discovery in Databases-) (Fayyad et al., 1996).

El método tradicional de convertir información en conocimiento se apoya en el análisis manual y en la interpretación, por ejemplo en la industria del cuidado de la salud es común que los especialistas analicen tendencias y cambio en la información del cuidado médico

formando reportes que sirven como base para la toma de decisiones y planeación en la administración en el cuidado de la salud (Fayyad et al., 1996).

Para estas aplicaciones el sondeo manual de los registros es lento, costoso y altamente subjetivo. Es un hecho que el volumen de las bases de datos se está incrementando de tal forma que el análisis manual es impráctico (Fayyad et al., 1996).

El KDD ha evolucionado, y sigue evolucionando, desde la intersección de campos de investigación como máquina aprendizaje, patrón reconocimiento, bases de datos, estadísticas, inteligencia artificial, adquisición de conocimiento para sistemas expertos, visualización de datos, y computación de alto rendimiento (Fayyad et al., 1996; Hernández et al., 2004).

El objetivo consiste en extraer conocimiento de alto nivel desde información de bajo nivel en el contexto de un gran conjunto de datos (Fayyad et al., 1996).

La minería de datos es una componente del KDD, el cual depende de las técnicas de aprendizaje máquina, reconocimiento de patrones y estadísticas para encontrar patrones en la información, es un paso importante del proceso del KDD (Fayyad et al., 1996; Sumathi y Sivanandam, 2006).

La fuerza motriz del KDD son las bases de datos, de hecho el problema de manipulación efectiva de datos cuando no se ajustan en la memoria principal es de fundamental importancia para el KDD. Las tecnologías de bases de datos están logrando eficiencia en el acceso a datos, operaciones de ordenamiento y agrupamiento así como la optimización de consultas constituyen para ampliar los algoritmos para grandes conjuntos de datos. Los algoritmos de minería, estadísticas, reconocimiento de patrones y aprendizaje máquina suponen que los datos se encuentran en memoria principal.

En Fayyad et al. (1996) se define el KDD como el “proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos”. El KDD es el proceso global de describir conocimiento útil desde las bases de datos, mientras que la minería de datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos (Hernández et al., 2004).

Por otra parte Witten y Frank (2005) definen la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos, es decir, la tarea fundamental de la minería de datos es encontrar un modelo inteligible a partir de los datos.

### **III.2.1. Proceso de descubrimiento de conocimiento en bases de datos**

El proceso hace énfasis en las bases de datos como fuente primaria de información, el proceso consiste en muchos pasos (uno de ellos la minería de datos). El proceso del KDD se centra en el proceso global de descubrir conocimiento de la información incluyendo cómo los datos se almacenan y acceden, cómo se pueden ampliar los algoritmos para que se ejecute de forma eficiente sobre un conjunto masivo de datos, cómo los resultados pueden ser interpretados y visualizados, así como una eficiente interacción humano-máquina (Fayyad et al., 1996; Sumathi y Sivanandam, 2006).

Donde se consideran los datos como un conjunto de hechos, un patrón es una expresión en algún lenguaje describiendo un sub-conjunto de datos o modelo aplicable al sub-conjunto. Por lo tanto, la extracción de patrones designa un modelo adecuado para los datos así como una estructura para los datos; o en general hacer una descripción de alto nivel de los datos (Fayyad et al., 1996).

Fayyad et al. (1996) menciona que el termino proceso implica que el KDD se compone de muchos pasos, los cuales involucran preparación de los datos, búsqueda de patrones, evaluación y refinamiento del conocimiento, todos repetidos en múltiples iteraciones.

Además menciona que por no trivial se entendemos que la búsqueda o inferencia; de tal forma que no sea un simple cálculo de valor promedio de un conjunto de números.

Asimismo los patrones descubiertos deben de ser validos para nuevos datos con cierto grado de certeza, deseamos que los patrones sean de utilidad para el usuario o tarea, finalmente los patrones deben de ser entendibles.

Cualquier proceso de KDD debe constar de los siguientes pasos, los cuales se representan de forma grafica en la *Figura 6* A continuación describe el proceso de extracción de conocimiento.

El primer paso del KDD es el de entender el dominio de aplicación y el conocimiento previo relevante e identificar la meta del proceso del KDD desde el punto de vista del usuario. Posteriormente se selecciona el conjunto de datos en el cual el conocimiento será descubierto.

Teniendo identificados el conjunto de atributos iniciales se procede a realizar la limpieza y procesamiento previo de los datos, unos de los problemas es la presencia de valores que no se ajustan al comportamiento general de los datos, estos datos anómalos pueden representar errores en los datos o pueden ser valores incorrectos que son diferentes a los demás. La presencia de los datos faltantes puede ser un problema pernicioso que puede conducir a resultados poco precisos. Un procesamiento previo de los datos es la construcción de información aplicando alguna operación a los datos ya existentes.

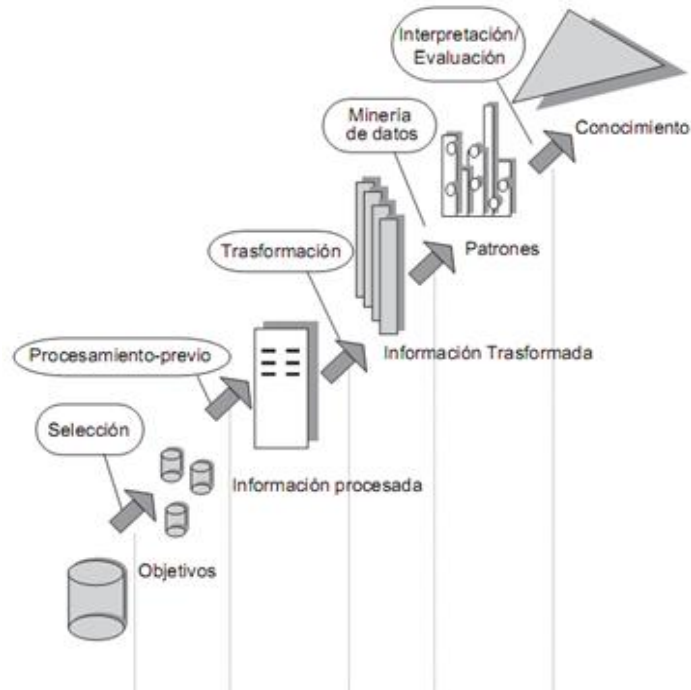
Después del procesamiento previo de los datos se inicia con el proceso de reducción y proyección donde se encuentra rasgos para representar los datos que dependan de la meta de la tarea. Con métodos efectivos de transformación o reducción, el número de variables a consideración puede ser reducido.

Con el conjunto de datos transformado se inicia con el análisis de exploración y selección del modelo de minería de datos y seleccionar los métodos que serán usados para la búsqueda de patrones. Este paso incluye decidir cual modelo puede ser apropiado para las metas del KDD.

Con la selección del modelo de minería de datos se inicia la búsqueda de reglas o patrones de nuestro interés en una forma particular de representación, tales representaciones incluyen reglas de calificación o de árboles, regresión y clasificación.

Cuando se tienen el conjunto de reglas o patrones estos son interpretados y evaluados para un posible regreso a las anteriores fases del proceso de extracción de conocimiento

Para finalizar el conocimiento adquirido debe de ser consolidado, en un sistema o en la creación de un reporte en el cual sea utilizado el conocimiento adquirido.



**Figura 6. Proceso de extracción de conocimiento el cual se destaca el de minería de datos (Fayyad et al., 1996.)**

Como se puede observar el proceso del KDD consta de varios pasos de cuales se destaca el de minería de datos a continuación se describirá el proceso las operaciones que se pueden realizar para el procesamiento previo y transformación de la información.

### **III.2.2. Procesamiento previo y transformación de la información**

En los proyectos de minería de datos, hay una excesiva prisa en la aplicación de técnicas de análisis para la extracción del conocimiento oculto en los datos. Para utilizar dichas técnicas y herramientas, hay que desarrollar una de las partes fundamentales del proyecto y de las que se lleva más tiempo, que es en la fase de preparación previa a la aplicación de los algoritmos de análisis (Bramer, 2007).

Las bases de datos reales son altamente susceptibles a datos perdidos e inconsistencias las cuales tienen origen desde múltiples fuentes heterogéneas, la baja calidad de los datos brindará una baja calidad de los resultados de la minería (Hernández et al., 2004).

Existen numerosas técnicas para el procesamiento previo de la información. Las técnicas de limpieza de datos pueden ser aplicadas para remover el ruido y corregir las inconsistencias en la información. La integración de datos puede ser usada para fusionar los datos de diferentes bases de datos tales como las bodegas de bases de datos. La transformación de los datos como la normalización puede mejorar la exactitud y eficiencia de los algoritmos de la minería de datos. La reducción de los datos puede reducir el tamaño de los datos eliminando atributos redundantes (Han y Kamber, 2006).

La información puede ser reducida por generalización con la utilización de conceptos de jerarquías, donde los conceptos de bajo nivel como *ciudades* a la que pertenece un paciente puede ser remplazada por un concepto de alto nivel tales como la *región* a la cual pertenece la ciudad. La discretización es otra forma de reducción que es muy útil para la generación automática de jerarquías de información numérica (Han y Kamber, 2006).

El proceso de discretización consiste en transformar los valores de una variable continua en un conjunto de valores discretos. Un algoritmo de discretización divide el dominio de la variable continua en un número finito de intervalos disjuntos que cubran completamente el dominio. Luego se asigna a cada observación de la variable continua el valor correspondiente al intervalo que lo contiene. De acuerdo con López de Castilla (2005) los métodos de discretización se dividen en:

- **Métodos Locales o Globales.** Los métodos globales utilizan todas las observaciones para el proceso de discretización; a diferencia de los métodos locales que sólo utilizan un subconjunto de las mismas.
- **Métodos Supervisados o No Supervisados.** Los métodos de discretización pueden utilizar la información de la clase a la que pertenece cada observación (métodos supervisados) para mejorar la calidad de la discretización o no tomar en consideración esta información (métodos no supervisados).
- **Métodos Estáticos o Dinámicos.** Los métodos estáticos son aquellos que discretizan cada variable independientemente de las demás. Los métodos dinámicos son aquellos que discretizan todas las variables simultáneamente tratando de utilizar las dependencias existentes entre ellas.
- **Métodos separación o agrupamiento.** Los métodos separación comienzan con una lista vacía de puntos de corte. Estos se van agregando dividiendo los intervalos conforme el proceso de discretización se realiza. Los métodos de agrupamiento comienzan con una lista completa de todos los valores continuos de la variable como puntos de corte. Estos se van eliminando juntando los intervalos conforme el proceso de discretización se realiza.



Hasta este punto se mostraron de forma breve en que consiste el procesamiento previo de los datos y algunos métodos para ello, a continuación mostraremos

A continuación se describirá el paso de la minería de datos cuáles son las técnicas de minería de datos relevantes para este trabajo de tesis en el contexto de extracción de conocimiento en ambientes médicos.

### **III.3. Minería de datos**

La cantidad de datos en el mundo y en la vida de las personas ha estado creciendo en las últimas décadas, gran parte de esta información es histórica, es decir, representa transacciones que se han producido. Además de su función como registro de las decisiones de la organización, la información histórica es útil para explicar el pasado, entender el presente y predecir información futura (Hernández et al., 2004).

La minería de datos es la aplicación de los algoritmos para la extracción de modelos predictivos entonces objetivo fundamental de la minería de datos es encontrar modelos inteligibles a partir de grandes cantidades de datos.

La minería de datos clínica es un área de investigación multidisciplinaria, usando metodología de extracción de conocimiento, inteligencia artificial, estadística, tecnología de bases de datos, sistemas expertos, y visualización de la información. Actualmente las herramientas de la minería están disponibles para la exploración de las bases de datos, ejemplo de estas técnicas son reglas de asociación y clasificación de árboles. El potencial para el diseño del DSS usando bases de datos clínicas, responde a la necesidad de diagnósticos efectivos (Dahlstrom et al., 2006).

En el problema de decisión médica es importante que los CDSS brinden una explicación para justificar sus decisiones, especialmente cuando se presenta un nuevo problema, por lo tanto la interpretación del conocimiento es una importante propiedad de estos sistemas que inducen soluciones a partir de información de casos médicos del pasado, algunos métodos

de minería de datos tienen la propiedad de inducir representaciones simbólicas a partir de los datos, como los son los árboles de decisión, reglas de inducción, reglas de asociación y clasificación (lavrač, 1999).

Se han realizado esfuerzos en la obtención de reglas para el diagnóstico médico. Tsumoto (2004), caracteriza los métodos de obtención de reglas a partir de técnicas de minería de datos, posteriormente hace una evaluación de diferentes métodos de obtención de reglas para el diagnóstico médico de dolores de cabeza, meningitis, y enfermedades cerebro vasculares, demostrando que las reglas obtenidas a partir bases de datos con información médica, representan correctamente el proceso de decisión de los expertos.

Dentro de las aplicaciones que tiene la minería de datos son los diagnósticos médicos, detección de tarjetas de crédito fraudulentas, predicción del comportamiento de compra del cliente, predecir intereses personales para los usuarios de web y optimización del proceso de manufactura (Mitchell, 1999).

Ejemplos de Aplicaciones de minería de datos es el análisis de riesgo para la aprobación de crédito de un cliente, otro ejemplo es para mejorar la decisión médica identificando futuros riesgos en el embarazo, especialmente a las que requieren cesaría, basándose en un conjunto de datos de 9714 registros de mujeres embarazadas (Mitchell, 1999).

Por lo tanto, dos grandes retos de la minería de datos son: trabajar con grandes volúmenes de datos y usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. Este conocimiento puede ser en forma de relaciones, patrones o reglas inferidas de los datos. Estas reglas constituyen el modelo de los datos analizados (Hernández et al., 2004).

Los modelos pueden ser para la predicción o descripción. Los modelos para la predicción pretenden estimar valores futuros o desconocidos de las variables de interés. Los modelos de descripción, en cambio identifican patrones que explican o resumen los datos

(Hernández et al., 2004; Fayyad et al. 1996). A continuación se describen las tareas de la minería de datos.

### **III.3.1. Tareas de minería datos**

Dentro de la minería de datos se pueden distinguir diversos tipos de tareas, cada una puede considerarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos. Esto significa que cada tarea tiene sus propios requisitos y el tipo de información obtenida con una tarea puede diferir mucho de lo obtenido de la otra tarea (Hernández et al., 2004).

Tal como se describió anteriormente las tareas pueden ser descriptivas o predictivas. Entre las tareas predictivas encontramos la clasificación y la regresión mientras que el agrupamiento, las reglas de asociación secuenciales y las de correlación son tareas descriptivas. A continuación daremos una breve introducción a cada una de ellas.

La clasificación es quizás la tarea más utilizada. En ella, cada instancia pertenece a una clase, la cual se indica mediante el valor de un atributo que se conoce como clase de la instancia. Este atributo puede tomar diferentes valores discretos, cada uno de los cuales corresponde a una clase. El resto de los atributos de la instancia se utilizarán para predecir la clase. El objetivo es predecir la clase de nuevas instancias de las que se desconoce la clase (Hernández et al., 2004).

La regresión es una tarea predictiva que consiste en aprender una función real que asigna a cada instancia un valor real. Esta es la principal diferencia respecto a la clasificación; el valor a predecir es numérico. El objetivo en este caso es minimizar el error entre el valor predicho y el valor real (Hernández et al., 2004).

El agrupamiento es una tarea descriptiva la cual consiste en obtener diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos (Hernández et al., 2004).

Las correlaciones son una tarea descriptiva que se usa para examinar el grado de similitud de los valores de dos variables numéricas.

Las reglas de asociación son también una tarea descriptiva, muy similar a las correlaciones que tiene como objetivo identificar relaciones no explícitas entre atributos categóricos. A continuación se presentan los métodos de minería de datos.

### **III.3.2. Métodos de minería de datos**

En esta sección se muestran los métodos y algoritmos para la obtención de modelos predictivos a partir de la información médica. Estos métodos de aprendizaje maquina presentados pueden ser clasificados en métodos simbólicos y sub-simbólicos. Los métodos símbolos son aquellos que el conocimiento puede ser representado de forma explícita ya sea en forma de reglas o arboles de decisión, en cambio los métodos sub-simbolicos son aquellos que carecen de esa capacidad convirtiéndose en cajas negras. Ejemplos de métodos simbólicos son las reglas de inducción métodos tales como aprendizaje de reglas de *si- entonces* , arboles de regresión y decisión, programas lógicos y aprendizaje basado en casos. Ejemplos de métodos sub-simbólicos son: aprendizaje bayesianos, métodos de aprendizaje basados en instancias, redes neuronales (lavrač, 1999).

Para los médicos es importante que un sistema de soporte a decisión sea capaz de justificar las decisiones, especialmente cuando se enfrenta a un nuevo problema los usuarios requieren de una considerable justificación y explicación (lavrač, 1999).

Entonces la interoperatividad del conocimiento inducido es una importante propiedad de los sistemas que inducen soluciones de información de casos médicos resueltos en el pasado. Los métodos simbólicos de minería de datos tienen esta propiedad de inducir representación simbólicas desde los datos (tales como los arboles de decisión). Por otro lado tenemos los métodos sub-simbólicos que típicamente carecen de esta propiedad en la cual esconden su uso en situaciones en las cuales una explicación es requerida (lavrač, 1999).

Sin embargo cuando la exactitud de la clasificación es el principal criterio de aplicación, los métodos sub simbólicos pueden ser muy apropiados y conseguir mayor exactitud que los clasificadores simbólicos (lavrač, 2001).

### **III.3.2.1. Reglas de inducción**

Las reglas de asociación son altamente populares en la representación de patrones de datos. Estos patrones pueden servir para conocer el comportamiento general del problema que generan la base de datos, y de esta manera tener más información que pueda asistir en la toma de decisiones. Una regla de asociación es una proposición probabilística sobre la ocurrencia de ciertos estados en una base de datos. Entonces, dado un conjunto clasificado de ejemplos, un sistema de reglas de inducción construye un conjunto de reglas en la forma si condiciones entonces conclusión. La condición parte de una regla que es la conjunción de atributos de la forma  $A_i = valor$  para atributos discretos y  $A_i < valor$  O  $A_i > valor$  para atributos continuos. La conclusión es parte de la clase asignada  $C = c_i$ . Un ejemplo de reglas de inducción es en el dominio de diagnostico de enfermedades reumáticas asignados en el diagnostico de cristal inducido sinovitis para pacientes mayores de 46 años que tienen más de tres articulaciones dolorosas y psoriasis como manifestación en la piel, entonces teniendo los atributos de sexo, edad, numero de articulaciones y tipo de manifestación en la piel la regla se representa como sigue:

“**Si** sexo = masculino Y Edad > 46 Y Número de articulaciones dolorosas > 3 Y Manifestación en la piel es de psoriasis **Entonces**

    Diagnostico = Cristal inducido sinovitis

**Fin Si**”

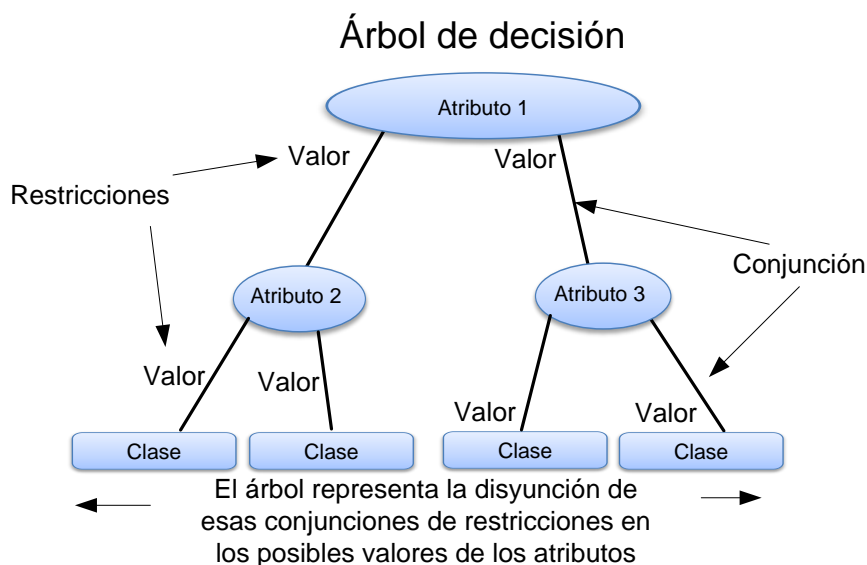
Es bien conocido que los algoritmos de la familia CN2 (Clark y Boswell, 1991) para aprendizaje de reglas. Cubriendo un sistema de reglas para cada caso  $c_i$ . Cuando la reglas de la clase  $c_i$  se están construyendo, los ejemplos son positivos y el restos de los ejemplos son negativos El algoritmo CN2 trabaja de la siguiente manera: construye una regla que correctamente es clasificada con algunos ejemplos, remueven los ejemplos positivos cubiertos por la regla y repiten el proceso hasta que no sigue habiendo más ejemplos. Para construir una regla sencilla el CN2, comienza con una regla vacía (SI ENTONCES) y selecciona una clase que tenga una consecuencia. El antecedente de esta regla que satisface todos los ejemplos del conjunto de entrenamiento. Progresivamente el algoritmo CN2 va agregando condiciones hasta que satisface todos los ejemplos, el antecedente de la clase  $c_i$ , para permitir manejar los datos imperfectos, algunas reglas si- entonces pueden ser reglas imprecisas que no fueron clasificadas correctamente por el conjunto de entrenamiento. El algoritmo CN2 es frecuentemente usado para analizar información médica (lavrač, 2001).

### III.3.2.2. Árboles de decisión

El aprendizaje de arboles de decisión es una de las técnicas de inferencia inductiva más usadas. Se trata de un método para aproximar funciones de valores discretos, capaz de expresar hipótesis disyuntivas y robusto al ruido en los ejemplos de entrenamiento. La familia de algoritmos para la inducción de arboles de decisión que incluyen ID3 (Quinlan J.R, 1986) y C4.5 (Quinlan J.R, 1993). Estos algoritmos llevan a cabo su búsqueda de hipótesis en un espacio completamente expresivo. El sesgo inductivo en este caso, consiste en la preferencia por árboles pequeños, sobre arboles grandes. Un árbol aprendido puede representarse también como un conjunto de reglas *si-entonces*, más fáciles de entender para un usuario (Mitchell, 1997).

### III.3.2.2.1. Representación de los árboles de decisión

Cada nodo del árbol está conformado por un atributo, las ramas que salen de los nodos, corresponden a los posibles valores del atributo correspondiente. Un árbol de decisión clasifica a un ejemplar, filtrándolo de manera descendente, hasta encontrar una hoja, que corresponde a la clasificación buscada.



**Figura 7. Árbol de decisión.**

Cada rama que va de la raíz del árbol a una hoja, representa una conjunción de tales restricciones y el árbol mismo representa la disyunción de esas conjunciones, entonces en general, un árbol de decisión representa una disyunción de conjunciones de restricciones en los posibles valores de los atributos de los ejemplares, esto se puede observar en la *Figura 7*.

### III.3.2.2.2. Problemas apropiados para los árboles de decisión

Aun cuando se han desarrollado diversos métodos para la inducción de árboles de decisión, y cada uno de ellos ofrece diferentes capacidades, en general estos algoritmos son apropiados para solucionar problemas de aprendizaje conocidos como problemas de

clasificación. De acuerdo con Mitchell (1997) estos problemas presentan las siguientes características:

**Ejemplares representados por pares atributo-valor.** Los ejemplares del problema están representados como un conjunto fijo de atributos, por ejemplo *Cielo* y sus valores, por ej. Soleado El caso más sencillo es cuando cada atributo toma valores de un pequeño conjunto discreto y cada valor es disjunto, por ejemplo {Soleado, Nublado, Lluvia}. Existen extensiones para trabajar con atributos de valores reales, por ejemplo Temperatura expresado numéricamente.

**La función objetivo tiene valores discretos.** El árbol de decisión de la figura 1, asigna una clasificación binaria, por ejemplo SI o NO a cada ejemplar. Un árbol de decisión puede ser extendido fácilmente, para representar funciones objetivos con más de dos valores posibles. Una extensión menos simple consiste en considerar funciones objetivo de valores discretos.

**Se necesitan descripciones disyuntivas.** Como se mencionó, los árboles de decisión representan naturalmente conceptos disyuntivos.

**Ruido en los ejemplos de entrenamiento.** El método es robusto al ruido en los ejemplos de entrenamiento, tanto errores de clasificación, como errores en los valores de los atributos.

**Valores faltantes en los ejemplos.** El método puede usarse aún cuando algunos ejemplos de entrenamiento tengan valores desconocidos para algunos atributos.



### III.3.2.2.3. Algoritmo básico de aprendizaje de arboles de decisión

Este procedimiento debido a Quinlan (1979), desarrollo a partir del sistema de aprendizaje basado en objetos (Hunt et al., 1966), constituye un método sencillo y eficaz para identificar arboles mínimos de clasificación de unidades de experiencia previamente asignadas a determinadas clases.

Por lo tanto, se trata de un sistema de obtención de reglas. El lenguaje de descripción, tanto de los antecedentes, como de los consecuentes es del tipo atributo valor.

El árbol de clasificación se obtiene por enumeración sucesiva de los valores asignables a cada atributo en forma sucesiva. Así, se fija como raíz el atributo A y se obtienen como vértices siguientes, los correspondientes a sus distintos valores:  $A_1, A_2, \dots, A_n$ . En cada vértice  $A_i$  se obtienen los siguientes, a partir de los valores del siguiente atributo  $R: R_1, R_2, \dots, R_n$ . y así sucesivamente. A cada vértice se asigna el subconjunto de la experiencia de partida, que satisface la conjunción de afirmaciones atributo-valor en el camino hasta la raíz (Cuenca et al. 1986).

Evidentemente, cuando el subconjunto asignado a un vértice, pertenece a única clase de las prefijadas, no es preciso continuar su expansión, ya que se trata de un nodo en que se identifica ya una clase.

Así, por ejemplo, se parte de un conjunto de experiencias, cuyos elementos se denominan:

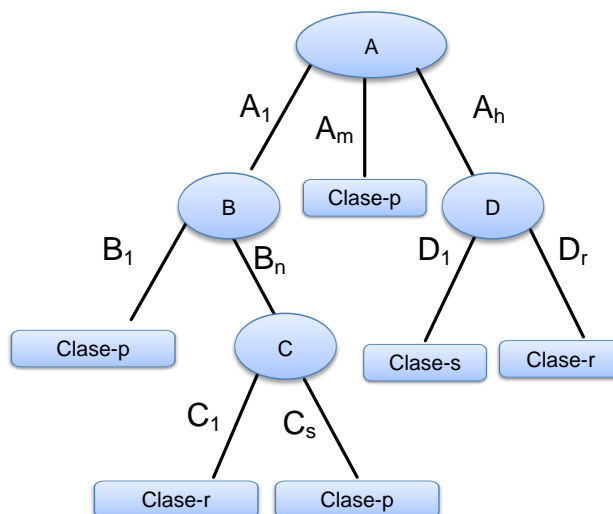
$$A_m^j, B_n^j, C_s^j, \dots, Clase_p^j \quad (1)$$

Siendo:

$A_m^j, B_n^j, C_s^j, \dots$  Los valores de los atributos  $A, B, C, \dots$  en el elemento  $j$  del conjunto de experiencias.

$Clase_p^j$  La clase asignada,  $p$ , en el mismo elemento  $j$ .

En la *Figura 8*, se muestra un posible árbol de clasificación según las clases  $p$ ,  $r$ ,  $s$  el cual puede tener la siguiente forma:



*Figura 8. Árbol de clasificación según las clases  $p$ ,  $r$ ,  $s$ .*

En cada rama se incluyen los valores de los atributos con representantes en el conjunto de experiencias (así en el atributo  $A$ , por ejemplo, se han considerado solamente el  $A_1$  y  $A_m$  por ejemplo).

De acuerdo con el árbol de la *Figura 8*, la experiencia podría describirse con las siguientes reglas de asignación de clases:

**Si  $A = A_1$  y  $B = B_1$  Entonces**

Clase =  $p$

**Fin si**

**Si  $A = A_1$  y  $B = B_n$  y  $C = C_1$  Entonces**

Clase =  $r$

**Fin si**

**Si**  $A = A_1$  y  $B = B_n$  y  $C = C_s$  **Entonces**

Clase =  $p$

**Fin si**

**Si**  $A = A_m$  **Entonces**

Clase =  $p$

**Fin si**

Dado que los elementos del conjunto de experiencias de partida son conjunciones en número finito de pares atributo-valor, este proceso necesariamente termina, una vez prefijado el orden de consulta de los atributos en cada rama. El interés reside en definir un criterio que permita un orden de consulta que produzca un árbol mínimo, lo que se correspondería con un número mínimo de reglas con la estructura (numero de premisas) más simple posible. Para ello Quinlan propone como regla de selección del atributo a clasificar en cada vértice aquel con mayor contenido de información respecto del conjunto de experiencias a clasificar en el vértice. La medida a emplear para el contenido información es la entropía (Cuenca et al. 1986). La cual se explica en la siguiente sección.

#### *III.3.2.2.4. Entropía y ganancia de información*

El algoritmo ID3 utiliza la ganancia de información como medida de selección de un atributo. Esta medida está basada en el trabajo de Claude E. Shannon y Weaver (1948). sobre teoría de información, el cual estudia el contenido de información. Sea un nodo en el que existen un subconjunto de de experiencias a clasificar según los tipos  $1, 2, \dots, n$  y en el que es posible utilizar los atributos  $A, B, C$ , con los valores  $A_1, A_2, \dots, A_n; B_1, B_2, \dots, B_s; C_1, C_2, \dots, C_r$ .

La entropía en el nodo, se aproxima la probabilidad de una clase por su frecuencia, es:

$$H = - \sum_{i=1}^n P_i \log_2 P_i \quad (2)$$

Donde

$$P_i = \frac{\text{n elementos en el conjunto de experiencias en clase } i}{\text{n total de elementos en el conjunto de experiencias}} \quad (3)$$

Es decir la función  $H$  mide el número óptimo de bits necesarios para almacenar los valores de la  $clase_i$ , en otras palabras mide la cantidad de información contenida en la  $clase_i$  (Han y Kamber, 2006). La entropía es siempre no negativa, la entropía es cero si solo la  $clase_i$  es perfectamente predecible, es decir si la  $clase_i$  tiene probabilidades de uno (López de Castilla, 2005).

Si se clasifica según el atributo  $A$ , se producirán  $h$  subconjuntos de experiencia asociados a cada nodo en cada modelo, en los cuales se puede calcular la entropía parcial, siendo  $HP$  la entropía parcial,  $HP_i (i = 1, h)$  de la misma forma. Si se estima la nueva entropía por el promedio:

$$H(A) = \sum \text{prob}(A_i) HP_i \quad (4)$$

El incremento de información aportado por la clasificación según  $A$ , se evalúa por la disminución de la entropía producida respecto de la inicial, es decir:

$$\Delta I = H - H(A) \quad (5)$$

El criterio propuesto por Quinlan es seleccionar como siguiente atributo es clasificar según el árbol aquel con mayor  $\Delta I$ . Por lo tanto, el proceso consiste en obtener de manera similar  $\Delta I_a, \Delta I_b, \Delta I_c, \dots$  y seleccionar aquel con máximo valor (Cuenca et al. 1986).

Puede emplearse este método para la obtención de reglas con grado de deducibilidad parcial, si se establece como criterio de nodo terminar aquel en que algún atributo-valor de clasificación tiene una frecuencia relativa superior a un umbral específico.

### III.3.2.3. Aprendizaje Bayesiano

Los clasificadores bayesianos son clasificadores estadísticos. Ellos pueden predecir por medio de probabilidades si una tupla pertenece a una clase en particular.

Los clasificadores bayesianos se basan en el teorema de Bayes que más adelante describiremos. Estudios comparativos de han demostrado que el clasificador Bayesiano simple conocido como clasificador *Naive Bayes*, tiene un rendimiento comparado con los arboles de decisión y las redes neuronales. Los clasificadores bayesianos han exhibido alta exactitud y velocidad cuando se aplica a grandes bases de datos (lavrač, 1999).

El clasificador bayesiano simple supone que el valor de un atributo sobre una clase es independiente sobre los valores de otros atributos. Esta suposición es llamada independencia condicional de la clase. Esto hace que los cálculos involucrados sean sencillos. Las redes bayesianas son modelos gráficos que a diferencia del clasificador bayesiano simple permite la representación de dependencias entre los sub-conjuntos de atributos, las redes bayesianas también son usadas para la clasificación (Han y Kamber. 2006). A continuación se describe el teorema de Bayes en el cual se basa el aprendizaje Bayesiano.

Sea  $X$  una tupla de información. En términos bayesianos,  $X$  es considerado evidencia. La cual es descrita por un conjunto de  $n$  atributos. Sea  $H$  alguna hipótesis tal que la tupla de información  $X$  pertenece a una clase  $C$  específica.

Para problemas de clasificación queremos determinar  $P(H|X)$  que es la probabilidad posterior (*probabilidad posteriori*) de  $H$  condicionada sobre  $X$ . en contraste  $P(H)$  es la

probabilidad a priori de  $H$  sobre un ejemplo. En otras palabras queremos determinar a qué clase  $C$  pertenece la tupla  $X$  (Han y Kamber. 2006).

La probabilidad posterior  $P(H|X)$ , esta basada sobre mayor información que la probabilidad a priori  $P(H)$ ,  $P(H)$  es independiente de  $X$ .  $P(X)$  es la probabilidad a priori de  $X$ . Similarmente  $P(X|H)$  es la probabilidad posterior de  $X$  condicionada sobre  $H$ .

El teorema de Bayes propone una forma de calcular la probabilidad posterior  $P(H|X)$ , desde.  $P(H)$ ,  $P(X|H)$  y  $P(X)$ . El teorema de Bayes es el siguiente:

$$P(H|X) = \frac{P(X|H) P(H)}{P(X)} \quad (6)$$

En la siguiente sección mostraremos como el teorema de Bayes es usado en la clasificación

### III.3.2.3.1. Clasificador Naive Bayes

El clasificador Bayesiano simple trabaja de la siguiente forma:

1. Sea  $D$  un conjunto de tuplas de entrenamiento y que tengan asociadas una clase, como es usual cada tupla es representada por un vector de atributos de  $n$  dimensiones,  $X = (x_1, x_2, \dots, x_n)$ , describiendo  $n$  mediciones hechas sobre las tuplas desde los  $n$  atributos respectivamente  $A_1, A_2, \dots, A_n$ .
2. Se supone que hay  $m$  clases,  $C_1, C_2, \dots, C_m$ . Las cuales representan la hipótesis  $H$  entonces dado una tupla  $X$ , el clasificador podrá predecir que  $X$ , pertenece a la clase con la más alta probabilidad posterior condicionada sobre  $X$ . Esto es el clasificador Bayesiano *Naive Bayes* predecirá que la tupla  $X$  pertenece a la clase  $C_i$  si y solo si

(7)

$$P(C_i|X) > P(C_j|X) \text{ para } 1 \leq j \leq m, j \neq i$$

3. Entonces se maximiza  $P(C_i|X)$  para la clase  $C_i$  esta maximización es llamada la estimación de la hipótesis más probable, usando el teorema de Bayes.

$$P(C_i |X) = \frac{P(X|C_i) P(C_i)}{P(X)} \quad (8)$$

4. Como  $P(X)$  es constante para todas las clases, solo  $P(X|C_i) P(C_i)$  necesitan ser maximizadas. Si la probabilidad a priori no es conocida, entonces es común suponer que todas las clases son iguales, esto es  $P(C_1) = P(C_2) = \dots = P(C_m)$ , y por ello se maximiza  $P(C_j|X)$ . De otra manera se maximiza  $P(X|C_i) P(C_i)$ . la probabilidad a priori de las clases pueden ser estimadas por  $P(C_i) = |C_{i,D}| / |D|$ , donde  $|C_{i,D}|$  es el numero de entrenamiento de las tuplas de la clase  $C_i$  en  $D$ .
5. Dado un conjunto datos con muchos atributos, tendría un costo computacional muy alto calcular  $P(C_j|X)$ . entonces para reducir calculo en la evaluación de  $P(C_j|X)$ . Naive supone la independendencia condicional de la clase. Esto supone que los valores de los atributos son condicionalmente independientes uno de otro. Por lo tanto:

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i) \quad (9)$$

$$= P(x_1|C_1) \times P(x_2|C_2) \times \dots P(x_n|C_n)$$

6. De esta forma es más sencillo calcular  $P(x_1|C_1) \times P(x_2|C_2) \times \dots P(x_n|C_n)$  sobre un conjunto de entrenamiento. Donde  $x_k$  se refiere al valor de atributo  $A_k$  de la tupla  $X$ .
7. Para predecir a que clase pertenece  $X$ ,  $P(X|C_i) P(C_i)$  es evaluado para cada clase  $C_i$ , el clasificador predice a que clase  $C_i$  pertenece la tupla  $X$  si y solo si

$$P(X|C_i) P(C_i) > P(X|C_j) P(C_j) \text{ para } 1 \leq j \leq m, j \neq i \quad (10)$$

A continuación se mostrara una extensión del clasificador *Naive Bayes* el cual es las redes bayesianas.

#### III.3.2.3.2. Redes Bayesianas

Las redes bayesianas un formalismo que en los últimos años ha demostrado su potencialidad como modelos de representación del conocimiento con incertidumbre. Este formalismo nació como una aportación de diferentes campos de investigación: teoría de toma de decisiones, estadística e inteligencia artificial. El éxito de numerosas aplicaciones variadas como en la medicina, la recuperación de información, la visión artificial, la fusión de información, la agricultura etc., avalan este formalismo (Han y Kamber. 2006).

Las redes bayesianas representan conocimiento cualitativo del modelo mediante un grafo dirigido acíclico, este conocimiento se articula en la definición de relaciones de independencia/dependencia entre las variables que componen el modelo. Estas relaciones abarcan desde una independencia completa hasta una dependencia funcional entre variables del modelo. El hecho de utilizar una representación grafica para la especificación de un modelo hace que las redes bayesianas son un herramienta muy atractiva en su uso como representación del conocimiento. A continuación se brinda la definición formal de una red bayesiana (Witten y Frank, 2005).



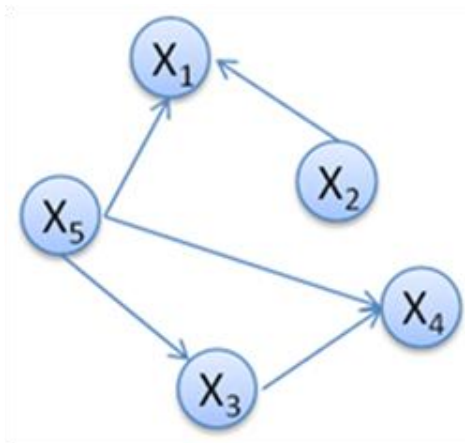
Una red bayesiana para un conjunto de variables aleatorias  $X = (X_1, X_2, \dots, X_n)$ , es un par  $B = (G, \Theta)$ , donde  $G$  es un grado aciclico dirigido, cuyos nodos se encuentran en correspondencia uno a uno con las variables  $X$ ,  $\Theta$  es un conjunto de funciones de probabilidad  $P(X_i|Pa(X_i))$  de  $i$  hasta  $n$ . Se usa  $Pa(X_i)$  denotar a los padres del nodo  $X_i$  en el grado  $G$ .

La codificación de las relaciones de independencia condicional expresadas en una red bayesiana hace que la distribución de probabilidad se pueda almacenar mucho más eficiente y local a cada una de las variables del modelo

$$P(X) = P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i|Pa(X_i)) \quad (11)$$

Las funciones de probabilidad local  $P$  son las funciones de probabilidad correspondientes a los términos de la ecuación (11). Un ejemplo de red bayesiana se presenta en la siguiente *Figura 9*. La función de probabilidad conjunta representada en esta red es:

$$P(X_1, X_2, \dots, X_5) = P(X_1|X_2, X_5)P(X_2)P(X_3|X_5)P(X_4|X_3, X_5)P(X_5) \quad (12)$$



**Figura 9. Red Bayesiana**

Una red bayesiana, una vez construida constituye un dispositivo potente para el razonamiento probabilístico, sin embargo nos queda la tarea de la construcción del modelo. Por lo tanto el problema de aprendizaje de las redes bayesianas se puede definir como sigue: dado un conjunto de datos  $D$ , encontrar el grafo aciclico dirigido  $G$  que mejor represente el conjunto de dependencias/independencias presentes en los datos. Para solucionar este problema desde el punto de vista bayesiano, es necesario calcular la probabilidad posterior de una red bayesiana en concreto, dado el conjunto de datos conocidos, esto es,  $P(G|D)$ . Una vez que sabemos cómo calcular esta probabilidad, tendremos la medida de educación de cada grafo (red bayesiana) a los datos de partida y por consiguiente podremos comparar para quedarnos con la mejor, entre distintas redes bayesianas. Se ha demostrado que el problema del aprendizaje, tal como se ha definido, es un problema NP-Duro (Chikering et al. 1996)

Para resolver el problema, se utiliza métrica para medir la optimalidad de una red bayesiana a continuación se presenta el criterio de información bayesiana.

*Criterio de información Bayesiana.* También conocido como criterio de información de Schwarz (1978) no requiere la evaluación de la función de probabilidad a priori. Consecuentemente, se trata de un criterio práctico usado en circunstancias apropiadas. Cuando se aplica a redes con distribución multinomial para las variables este criterio, la cual tiene la siguiente expresión:

$$f_i(X_i|Pa(X_i):D) = \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \frac{N_{ijk}}{N_{ij}} - 1/2 Dim(B) \log \|B\| \quad (13)$$

Donde  $Dim(B)$  representa una función para calcular el número de parámetros libres que se necesitan especificar en una red bayesiana, esto es,  $Dim(B) = \sum_{i=1}^n (r_i - 1)q_i$ , (número de entradas independientes en las tablas de probabilidad) .

Una vez especificado el problema de aprendizaje se puede identificar que es un problema de optimización, esto es, se tiene una métrica definida para calcular la adaptación de los datos a una red bayesiana por lo tanto se debe de encontrar la solución que maximice la métrica. Pero como este problema de optimización no tiene una solución exacta, por consiguiente, se han planteado en la literatura específica del tema de algoritmos de búsqueda específicos o adaptación de algoritmos metaheurísticos para la resolución aproximada del aprendizaje estructural (Sumathi y Sivanandam, 2006)

Los algoritmos de búsqueda más utilizados para el aprendizaje, basados en métodos voraces y locales: K2(Cooper et al. 1992), B (Buntine 1994), de los cuales solo estudiaremos el algoritmo K2.

#### *III.3.2.3.2.1. Algoritmo K2*

Este algoritmo utiliza un esquema voraz en su búsqueda de soluciones candidatas cada vez mejores y parte de las variables de entrada están ordenadas, de esta forma que los posibles padres de una variables aparecen en el orden antes que ella misma. El proporcionarle al algoritmo un orden entre las variables y este tan solo tenga que buscar el mejor conjunto de padres posible de entre las variables predecesoras en el orden. La búsqueda del conjunto se hace de forma voraz.

El algoritmo parte de que el conjunto de padres para cada variable es el conjunto vacío. Posteriormente, y siguiendo el orden preestablecido, pasa a procesar cada variable  $X_i$ , calcula la ganancia que se produce en la medida utilizada al introducir una variables  $X_j$  como padre de  $X_i$  de entre todas sus predecesoras, esto es, para todo  $j < i$  y se queda con la que produce la mejor ganancia. Desde el enfoque bayesiano, esta ganancia se calcula como la razón de la métrica vista en el punto anterior de una red bayesiana con la variable  $X_i$  sin padres y una red bayesiana donde  $X_i$  tenga como padre a  $X_j$ . Una vez fijado el primer padre de  $X_i$  entonces se prosigue de igual forma, pero esta vez se mide la razón de una red

bayesiana con  $X_i$  y el primer padre fijado en una red bayesiana donde se le introduce un segundo padre no insertado previamente. En la *Figura 10* podemos encontrar el pseudocódigo de este algoritmo.

```

Algoritmo K2(X: nodos (variables ordenadas), D: Datos)
Para cada  $X_i$  ( $i = 0$  hasta  $n$ )
     $Pa(X_i) = \emptyset$ 
Fin para cada
Para cada  $X_i$  ( $i = 0$  hasta  $n$ )
     $Z := verdadero$ 
    Mientras  $Z$ 
        sea  $X_i$  el nodo tal que  $i < j$  y  $X_j$  no pertenezca a  $Pa(X_i)$ 
        que maximiza  $f_i(X_i | Pa(X_i) \cup X_j : D)$ .
    Si  $f_i(X_i | Pa(X_i) \cup X_j : D) > f_i(X_i | Pa(X_i) : D)$  Entonces
         $Pa(X_i) := Pa(X_i) \cup X_j$ 
    Si No
         $Z := falso$ 
    Fin si
Fin mientras
Fin para cada
Fin del algoritmo

```

*Figura 10. Pseudocódigo del algoritmo K2*

#### III.3.2.4. Aprendizaje basado en instancias

El aprendizaje basado en instancias, trata de resolver el problema a partir de información extraída de un conjunto de ejemplos. Los métodos basados en vecindad reciben su nombre del hecho de que la predicción se basa fundamentalmente en la utilización del conjunto de ejemplos (vecinos) al dato que hay que procesar, o en caso más general porque la *distancia* entre cada ejemplo y el dato en cuestión es esencial en el proceso (Witten y Frank, 2005).

Como se aprende de ejemplos, casos o datos conocidos. Entonces la idea es actuar de la misma forma que en las situaciones anteriores *parecidas o similares*, si en estas se tuvo

éxito. En consecuencia parece una regla bastante clara y natural que puede ser aplicada a multitud de tareas de minería de datos (Hernández et al., 2004)

En la tarea de clasificación, se puede asignar una clase a un nuevo caso, observando las clases de casos similares. Igualmente, en la tarea de agrupamiento, asignaremos un nuevo ejemplo al grupo donde estén los individuos similares. En el caso de la regresión, el valor predicho para un nuevo ejemplo no puede distar mucho de los valores obtenidos para ejemplos similares. Se entiende por similitud, dando lugar al concepto matemático equivalente de *distancia*, posteriormente se tiene que explotar dicha similitud (Sumathi y Sivanandam, 2006).

La manera de formalizar el concepto de similitud es a través de métricas o medidas de distancia. Una medida de distancia es aquella que se aplica sobre un conjunto de instancias o ejemplos, tales que todos los atributos son numéricos, por ejemplo, se puede definir la distancia euclidiana como sigue, donde los puntos o instancias  $x$  e  $y$  de dimensión  $n$ .

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (14)$$

El concepto de distancia no solo existe para valores numéricos también se puede utilizar cuando los ejemplos están formados por atributos nominales. Por ejemplo para atributos nominales se suele utilizar la función delta es decir  $\delta(a, b) = 0$  si y solo si  $a = b$  y  $\delta(a, b) = 1$  en caso contrario. En esta función se define la distancia de manera sencilla:

$$d(x, y) = \omega \sum_{i=1}^n \delta(a, b) \quad (15)$$

Donde  $\omega$  es un factor de reducción (por ejemplo  $1/n$ ), a continuación se explicara en qué consiste el algoritmo de los  $k$  vecinos más cercanos

En el algoritmos de los  $k$  vecinos más cercanos la calcificación se hace para cada nuevo caso en particular. Cuando un nuevo caso aparece, se genera un círculo con centro en dicho punto y un radio fijado prefijado como parámetro del sistema. Se calcula el número de ejemplos que caen dentro del círculo y se etiqueta al nuevo caso como perteneciente a la clase más numerosa dentro del círculo (Hernández et al., 2004).

Si se aumenta o disminuye el radio, el numero de ejemplos que caen dentro del circulo puede variar, y la predicción realizada también. Por lo tanto, este parámetro es crítico. También se puede observar que el concepto de círculo viene determinado por el concepto de distancia.

El método espera hasta la aparición de un nuevo dato a clasificar para la utilización del conjunto e ejemplos. Cuando el dato está disponible se recurre a los ejemplos para realizar la calcificación, se crea una regla local al data que acaba de llegar, se realiza la clasificación, y se abandona dicha regla. Si en ese momento se tendría que clasificar otro dato más, los cálculos realizados para la clasificación del dato anterior seria inservibles y se tendría que realizar de nuevo todo el proceso. Este método tiene que almacenar los ejemplos de entrenamiento, ya que son utilizados una y otra vez.

#### *III.3.2.4.1. Selección de atributos utilizando métodos de aprendizaje basado en instancias.*

Los atributos pueden ser seleccionados utilizando métodos de aprendizaje basado en instancias. En la mayoría de los dominios algunos atributos son irrelevantes y otros relevantes, entonces el próximo paso es conocer cuales atributos son relevantes para el resultado. En algunos esquemas el peso específico de la clase en algún atributo puede ser más importante para una clase que para otra. Entonces una descripción es producida por cada clase que distingue entre sus miembros de las otras clases. Esto conduce al problema

de que una nueva instancia puede ser asignada a varias clases diferentes o a ninguna de las clases. Se han propuesto soluciones heurísticas para resolver esta situación.

Una de ellas es la medida de distancia en la cual se incorporaron pesos  $w_1, w_2, \dots, w_n$  en cada dimensión:

$$\sqrt{w_1^2(x_1 - y_1)^2 + w_2^2(x_2 - y_2)^2 + \dots + w_n^2(x_n - y_n)^2} \quad (16)$$

En caso de que una clase especifique pesos, habrá un conjunto aparte de pesos para cada clase.

Todos los pesos de los atributos son actualizadas después de cada instancia del conjunto de entrenamiento es clasificada y la instancia más similar (o la instancia mas similar de cada clase) es usada como base para actualizar. Entonces en el conjunto de entrenamiento la instancia  $x$  y el ejemplo más similar  $y$ . Para cada atributo  $i$ , la diferencia  $|x_i - y_i|$  es la medida de contribución de ese atributo para la decisión. Si la diferencia es pequeña entonces el atributo contribuye de manera positiva, mientras si la medida es grande entonces contribuye de manera negativa. La idea básica es actualizar el peso  $i$  en base al tamaño de la diferencia. Si la clasificación fue correcta el peso asociado se incrementa y si no es correcta el peso asociado se decremento, al final del proceso se eligen los atributos con peso positivo y se descartan los que tengan peso negativo

### III.3.3. Evaluación del conocimiento adquirido

Los métodos de aprendizaje permiten construir modelos o hipótesis a partir de un conjunto de datos, o evidencia. Es necesario evaluar la calidad de las hipótesis de la manera más exacta posible.

Por lo tanto, la etapa de evaluación de modelos es crucial para la aplicación real de las técnicas de minería de datos. Entonces el objetivo del aprendizaje automático supervisado es el aprendizaje de una función  $f$ , considerando un espacio de posibles hipótesis  $H$  (Mitchell, 1997). Para ello los algoritmos de aprendizaje emplean normalmente una evidencia o muestra  $S$  formada por ejemplos de la función objetivo  $f$  de acuerdo con una distribución  $D$ . Esta distribución define la probabilidad de encontrar cada instancia de la evidencia en el espacio de todas las posibles evidencias. A continuación se mostrará una técnica de evaluación basada en precisión la cual es la validación cruzada.

### **III.3.3.1. Evaluación mediante validación cruzada**

Este método consiste en dividir el conjunto de evidencia en  $k$  subconjuntos disjuntos de similar tamaño. Entonces, se aprende una hipótesis utilizando el conjunto formado por la unión de  $k - 1$  subconjuntos y el subconjunto restante crea para calcular un error de muestra parcial. Este procedimiento se repite  $k$  veces, utilizando siempre un subconjunto diferente para estimar el error de muestra parcial. El error de la muestra final se calcula como la media aritmética de los  $k$  errores de muestra parciales. De esta manera recoge la media de los experimentos con  $k$  subconjuntos de prueba independientes (Hernández et al., 2004).

Una ventaja de la validación cruzada es que la varianza de los  $k$  errores de muestra parciales, permite estimar la variabilidad del método de aprendizaje con respecto a la evidencia. Comúnmente, se suelen utilizar 10 particiones (Witten y Frank, 2005).

### **III.3.3.2. Medidas de evaluación de un clasificador**

La exactitud y precisión de un clasificador sobre un conjunto de prueba, el porcentaje de las tuplas correctamente clasificadas por el clasificador, se le conoce como tasa de reconocimiento del clasificador, esto refleja cómo el clasificador reconoce las tuplas de que pueden corresponder a varias clases. A si la tasa de error de un clasificador  $M$  es  $1 -$



$Acc(M)$ , donde  $Acc(M)$  es la exactitud y precisión de  $M$ . Esta tasa de error es optimista comparado con la verdadera tasa de error. A continuación se presentaran otras herramientas para medir la exactitud, precisión de la tasa de error del clasificador.

Una herramienta muy utilizada para analizar el comportamiento del calificador sobre diferentes clases es la matriz de confusión. Dado  $m$  clases, una matriz de confusión es una matriz del tamaño de  $m \times m$ . Para una celda  $CM_{i,j}$  indica el numero de tuplas de la clase  $i$  han sido clasificadas en la clase  $j$ . Para que un clasificador tenga buena exactitud la mayoría de las tuplas deben de estar representadas en la diagonal principal de la matriz de confusión desde  $CM_{1,1}$  hasta  $CM_{m,m}$ , con el resto de las celdas cercanas a cero.

Entonces dadas dos clases se utilizan los términos de verdaderos positivos para referirse a las tuplas positivas que son correctamente clasificadas mientras que los verdaderos negativos para referirse a las negativas tuplas que son correctamente clasificadas. Las falsas positivas son las tuplas negativas que son clasificadas de forma incorrecta, las falas negativas son las tuplas positivas que son clasificadas de forma incorrecta. Esto se puede ver en la *Tabla I*.

*Tabla I. Matriz de confusión para las tuplas positivas y negativas*

	$C_1$	$C_2$
$C_1$	Verdaderos positivos (VP)	Falsos negativos (FN)
$C_2$	Falsos positivos (FP)	Verdaderos negativos (VN)

Un clasificador puede estar solamente reconociendo los verdaderos positivos o los verdaderos negativos, entonces es necesario conocer que tan bien el clasificador puede reconocer los verdaderos positivos de los falsos negativos, las medidas de precisión y alcance. La precisión mide la probabilidad de que si un sistema clasifica una tupla en una cierta clase, la tupla realmente pertenezca a esa clase, por otro lado, el alcance mide la probabilidad de que si una tupla pertenece a cierta clase, el sistema lo asigne a la clase.

Para estimar las medidas de precisión y alcance se utilizan las siguientes expresiones:

$$\textit{Precisión} = \frac{VP}{VP + FP} \quad (17)$$

$$\textit{Alcance} = \frac{VP}{VP + FN} \quad (18)$$

### **III.4. Resumen**

En este capítulo se presento un bosquejo general sobre el proceso de extracción de conocimiento. Al inicio se presento la definición del proceso y los pasos que lo componen. Posteriormente se mostro el paso de minería de datos donde se muestran los métodos de minería de datos para el tratamiento de información médica.

Para finalizar el capítulo se mostraron diversos esquemas de evaluación del los modelos predictivos. El siguiente capítulo se describe el proceso de detección y vigilancia epidemiológica donde se entenderá el dominio de aplicación y posteriormente iniciar con el proceso de extracción de conocimiento.

---

## **Proceso de detección y vigilancia de infecciones nosocomiales**

---

### **IV.1. Introducción**

Con el objetivo de obtener información directamente de los dueños del proceso, acerca del proceso de decisión se estudiamos en esta tesis, realizamos un estudio del contexto de aplicación en el lugar de trabajo de un grupo de trabajadores de conocimiento en el área clínica. En esta área de estudio empleamos métodos de investigación de campo cualitativos para la obtener información sobre los requerimientos, con una perspectiva socio-técnica. Todo lo anterior para ayudarnos a entender el contexto del proceso de decisión de los trabajadores de conocimiento.

El caso de estudio consistió en analizar el trabajo de vigilancia epidemiológica del hospital el cual es la observación sistemática, activa y continua de la ocurrencia y distribución de las infecciones en un hospital y de los eventos que aumentan o disminuyen el riesgo de que la infección ocurra.

Las infecciones nosocomiales (IN) son un conjunto heterogéneo de enfermedades infecciosas cuyo denominador común es el haber sido adquiridas en un hospital. Son causadas por gérmenes hospitalarios, en general se consideran que se presentan posteriores a las 72 horas de ingreso al centro asistencial. Una IN prolonga el tiempo de estancia hospitalaria, incrementado la mortalidad, eleva los costos de atención y afecta la calidad de vida del individuo durante la recuperación de la enfermedad de base. La infección no está presente ni incubándose al momento del ingreso del paciente al hospital.

La organización y estructura para la vigilancia y control de infecciones nosocomiales es el Comité de Detección y Control de Infecciones Nosocomiales (CODECIN) y la Unidad de Vigilancia Epidemiológica Hospitalaria (UVEH).

El CODECIN es el Grupo técnico multidisciplinario que participa en la detección y análisis de casos y es responsable en coordinación con el médico tratante de investigar, prevenir y proponer medidas de control, que permitan disminuir las infecciones hospitalarias.

Por otro lado la UVEH es la Instancia técnico-administrativa que efectúa las actividades de vigilancia epidemiológica incluyendo la referida a las infecciones nosocomiales (IN). Está conformada por: un epidemiólogo y una auxiliar de enfermería en salud pública, La UVEH realizará la vigilancia de los padecimientos considerados como infecciones nosocomiales conforme a lo establecido en la norma NOM-045-SSA2-2005.

La función del CODECIN es analizar la información sobre la vigilancia epidemiológica de IN y proponer al personal directivo de la unidad, acciones de mejora, para fortalecer la vigilancia epidemiológica.

## **IV.2. Estudio del proceso de detección y vigilancia de infecciones nosocomiales**

El enfoque de este estudio se hizo para entender como apoyar el proceso organizacional del hospital. Para lo cual se estableció los requerimientos de la investigación que pudiera apoyar al proceso organizacional, con un sistema de ayuda a la decisión en el contexto médico.

- Diseñada bajo una perspectiva socio-técnica
- Implementada sobre computo móvil

- Que utilice la memoria de la organización como base de conocimiento

Posteriormente se identificaron los requerimientos del caso de estudio los cuales son los siguientes:

- Necesidad de toma de decisiones
- Que tenga necesidades de movilidad
- Que cuente con una memoria organizacional

Después de revisar varios procesos hospitalarios, se identificó el proceso de detección y vigilancia epidemiológica (PDVIN). Donde existe la toma de decisión y satisfacen la mayoría de las características del caso del estudio que estamos buscando.

La siguiente actividad consistió en observar dentro de la organización con una perspectiva socio-técnica, el contexto del proceso al que le daremos apoyo, para realizar una ingeniería de requerimientos. En este sentido utilizamos técnicas cualitativas de investigación de campo, con el fin de observar como los trabajadores interactúan en el flujo de trabajo, y que documentos y tecnologías emplean en ellos. Además de familiarizarse con los conceptos del dominio del proceso. Para lo que nos apoyamos en las siguientes técnicas cualitativas:

**Entrevistas:** consistió en realizar 12 entrevistas al personal de la unidad de vigilancia epidemiología del hospital con el objetivo de entender el proceso de vigilancia y detección de infecciones nosocomiales. Y conocer cuáles son los artefactos de información que manejan e interactúan para la toma de decisión

**Observaciones informales:** se realizaron 4 observaciones informales del contexto de un grupo de médicos durante las actividades cotidianas dentro de la organización. Con el fin de verificar la información obtenida de las entrevistas.

En la primera entrevista que se realizó con el jefe de departamento de epidemiología nos describió a grandes rasgos en qué consistía el proceso de detección y vigilancia de infecciones nosocomiales y nos entregó los siguientes documentos: la norma oficial mexicana PROY-NOM-045-SSA2-2007, para la vigilancia epidemiológica, prevención y control de las infecciones nosocomiales y el documento donde explica el procedimiento para realizar la vigilancia de infecciones nosocomiales en las unidades médicas hospitalarias del instituto mexicano del seguro social.

El PDVIN se realiza de acuerdo a la norma oficial mexicana y a las entrevistas tiene consta de tres pasos principales los cuales son los siguientes: detección, información y control, esto se puede ver en la *Figura 11*, que componen el proceso del PDVIN.



**Figura 11.** El PDVIN, consta de tres sub-procesos en el cual se muestran los elementos de información utilizados y actividades que tiene cada uno.

La norma oficial mexicana PROY-NOM-045-SSA2-2005, contiene las definiciones utilizadas en el dominio de aplicación, adicionalmente contiene los criterios de diagnóstico para una infección nosocomial. A continuación se mostrara los productos obtenidos de este

caso de estudio como lo es la descripción textual del proceso y las diferentes técnicas de modelado de proceso como lo es la grafica rica, Diagramas Rol Actividad (RAD, por sus siglas en ingles) (Ould, 1995) y los Diagrama de Flujo de Datos (DFD).

#### **IV.2.1.Descripción textual del proceso**

El proceso de detección inicia desde la admisión del paciente la cual puede ser de dos formas un servicio hospitalario programado o de urgencia. Los ingresos de urgencia pueden ser primera vez o reingreso. El ingreso del paciente genera la hoja de ingreso y contiene la información con sus respectivos datos y especifica si el paciente es de nuevo ingreso o reingreso. Un reingreso es cuando un paciente es dado de alta en el hospital y que posteriormente se complico su estado de salud.

El paciente es atendido en unos de los servicios del hospital los cuales son: ginecología y obstetricia, cirugía y trauma, pediatría, medicina interna, posteriormente estos servicios del hospital realizan un censo de pacientes, excepto el de ginecología y obstetricia debido a que es una de los servicios más dinámicos del hospital dando como resultado una estancia muy corta, menor de dos días.

El censo de pacientes se encuentra en el área de servicios del hospital al igual que el expediente médico, el proceso de detección de infecciones nosocomiales puede ser de tres formas: el censo, por el expediente médico, preguntar a jefe de piso.

El censo puede ser una de las formas de identificar un caso sospechoso de infección nosocomial, así como realizar un recorrido por los diferentes servicios del hospital para revisar a los pacientes e identificar algún factor de riesgo, así como preguntar a los jefes de servicio, jefes de piso, enfermeras y médicos tratantes por los pacientes. Si el médico tratante, jefe de piso o enfermera detecta caso sospecho de infección nosocomial deberá de

reportarlo a la UVEH por medio del anexo 3 pero esta situación no se presenta ya que el personal médico no realiza esta actividad.

La forma de detección de una infección nosocomial a partir del censo es la siguiente todas las mañanas los jefes de los servicios como es medicina interna, pediatría, cirugía y traumatología realizan un censo de pacientes que se encuentran hospitalizados, y la cual contiene la siguiente información: nombre del paciente, número de cama, número de cédula, edad, diagnóstico, días de estancia, servicio, pendientes, cada división tiene un formato del censo diferente, esta información está disponible en digital.

Con esta información la UVEH puede identificar algunos factores de riesgo como puede ser la sobre-estancia en el hospital, la edad, el diagnóstico y los servicios que se le han realizado al paciente, de los cuales se pueden clasificar en una paciente con cirugía o sin cirugía.

Si encuentran a un paciente que presenta algunos factores de riesgo entonces revisan su expediente médico para ver a más a detalle el historial clínico del paciente al revisar las notas médicas, cirugías, tratamientos, exámenes de laboratorio, si es nuevo ingreso o reingreso y diagnósticos de enfermedades que ha padecido el paciente.

Con toda esta información es suficiente para determinar e identificar los factores de riesgo. Si el paciente presenta algunos factores de riesgo se convierte en un caso sospecho, cuando un caso es sospechoso se llena la forma de registro de infección nosocomial (FRCIN) ayudándose del censo para los datos iniciales y posteriormente utiliza el expediente médico para obtener la información sobre los diagnósticos, el historial quirúrgico, factores de riesgo, información sobre los análisis clínicos realizados y algunos tratamientos.

Posteriormente se dirige donde se encuentra el paciente para consultar, entrevistar, validar la información con el expediente con los síntomas presentados en paciente, además



de explorar otros factores de riesgo. Asimismo en algunos casos es necesario realizar exámenes de laboratorio por medio de la forma MF-8/2000 o una solicitud de consulta radiología por medio de la forma 4-30-2 y en base a los criterios de diagnóstico de cada infección determinan si el caso es confirmado o descartado.

Cuando un paciente es confirmado se inicia con el sub-proceso de información el cual consiste en vigilar la evolución relacionada con la infección nosocomial hasta el egreso del paciente en la cual puede ser por los siguientes motivos: transferencia, defunción, mejoría y alta voluntaria.

Cabe de aclarar que el trabajo de la UVEH es el de identificar la infección nosocomial y en ningún momento tratarla ya que la responsabilidad del paciente es del médico tratante, a lo mas realiza una sugerencia en el esquema de tratamiento de la infección por medio de una nota medica anexándola al expediente médico.

Otra forma de detección de casos de infección nosocomial es realizar un recorrido por las diferentes servicios de hospital para ver si un paciente presenta datos clínicos de infección nosocomial así como preguntar al personal médico ya sea los jefas de piso, jefes de servicio y médicos internos, las preguntas realizadas por la UVEH son de carácter informal.

A partir de los casos confirmados se realiza el informe mensual de infecciones nosocomiales el cual es entregado y presentado ante CODECIN y enviado a autoridades de salud superiores.






Todos los análisis realizados al paciente son anexados al expediente médico ya sea que fueran solicitados por la UVEH o el médico tratante, en caso de que el paciente presente una infección nosocomial la UVEH realiza una nota médica sobre el expediente médico del paciente especificando la infección. En la siguiente sección se mostrara el modelado del proceso en base a descripción textual

### IV.3. Modelado del caso de estudio

Una de las formas más efectivas para entender los procesos existentes, es presentarlos en un diagrama. Existen distintas técnicas diagramáticas para el modelado de procesos. En nuestro caso hemos decidido utilizar la técnica conocida como Gráfica Rica, desarrollada por Peter Checkland como parte de la Metodología de Sistemas Suaves (Monk and Howard 1998) y (Checkland and Scholes. 1999).

Una gráfica rica permite obtener una visión global del proceso, así como representarlo a distintos niveles de detalle. Mediante esta técnica es posible mostrar a los actores o personas involucradas en el proceso, así como su propósito en el mismo, sus deseos, miedos e inquietudes. También es posible mostrar los conflictos o acuerdos existentes, así como los detalles del medio ambiente en el que se desarrolla el proceso. La *Tabla II* presenta ejemplos de los elementos de la simbología de la de gráfica rica.

*Tabla II. Simbología de la grafica rica.*

Símbolo	Significado
	Rol
	Actividad
	Artefacto
	Conexiones
	Acciones de los roles

**Roles (agentes)**, que se representan con el ícono que se considere más apropiado.

**Actividades**, son tareas o interacciones de los roles y son representadas por nubes.

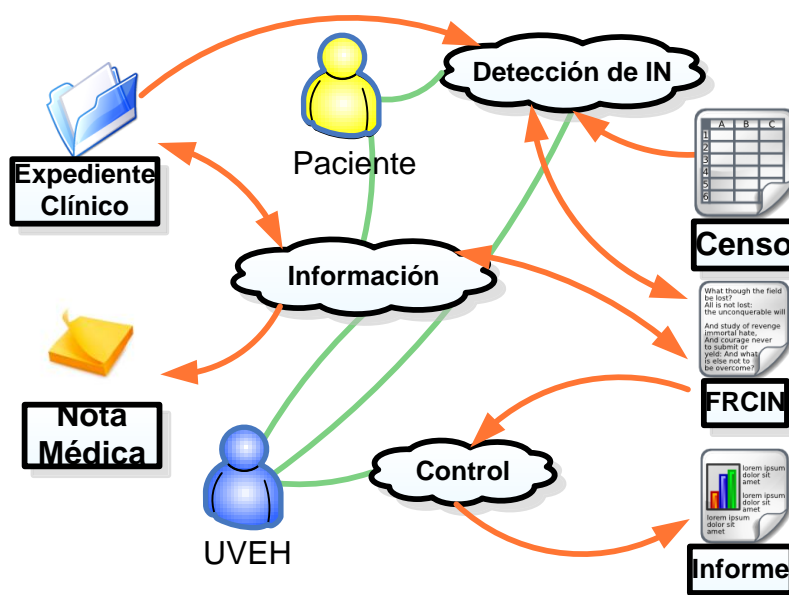
**Artefactos**, salidas o entradas de las actividades, los cuales se pueden representar por cajas o íconos (lo que se considere más apropiado).

**Conexiones**. Existen dos tipos de conexiones o relaciones principales entre los elementos de una gráfica rica: los roles se conectan con las actividades mediante líneas, mientras que los artefactos lo hacen mediante flechas, cuya dirección indica si son entradas, salidas o son modificados por dicha actividad

Acciones de los roles, pueden ser representadas por nubes de pensamiento. También pueden servir para expresar las inquietudes, miedos, preocupaciones, molestias, etc. de las personas que ejecutan dichos roles.

### IV.3.1. Grafica rica global del proceso de detección y vigilancia de IN

En la *Figura 12* se muestra la grafica rica global del proceso de detección y vigilancia de infecciones nosocomiales y cómo interactúan los artefactos con las principales actividades del proceso.

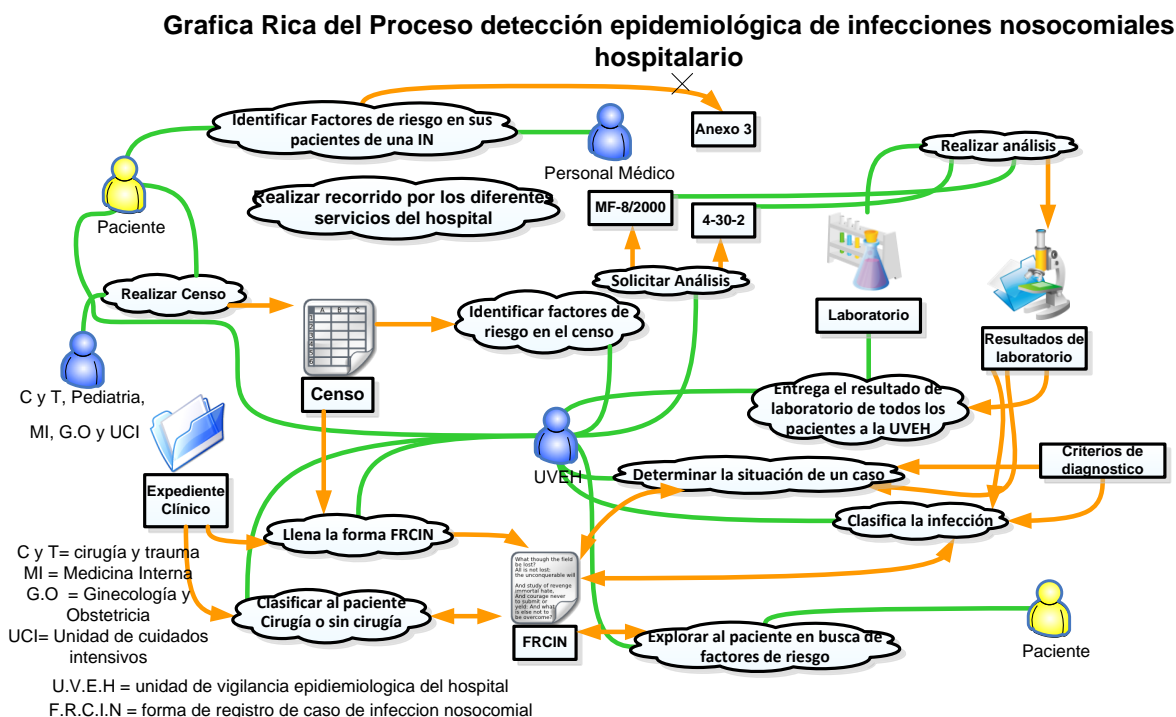


*Figura 12. Grafica rica global del proceso de detección y vigilancia de infecciones nosocomiales.*

En la *Figura 12* se muestra el proceso global del PDVIN y los principales actividades del y con qué artefactos de información interactúan como lo es el FRCIN, Censo expediente médico, la nota medica y el informe. En la siguiente grafica rica se muestra el sub-proceso de detección.

### IV.3.2. Grafica rica del sub-proceso de detección de IN

En la *Figura 13*. Se muestran todas las actividades realizadas para la detección de una infección nosocomial.



*Figura 13. Grafica rica para el sub proceso de detección epidemiológica de infecciones nosocomiales*

En la graficar rica de la *Figura 13* es muy importante debido a que muestra los artefactos y las actividades relevantes para la toma de decisión. Como la búsqueda de factores de riesgo del paciente, el llenado del FRCIN e identificación de factores de riesgo en el censo. La siguiente grafica se muestra el sub-proceso de información.

### IV.3.3. Grafica rica del sub-proceso de información de IN

En la *Figura 14*. Se muestra la grafica rica para el proceso de información, en el cual inicia con la detección de un caso de IN, en ella se ve que interactúa con el expediente clínico con el FRCIN y el paciente.

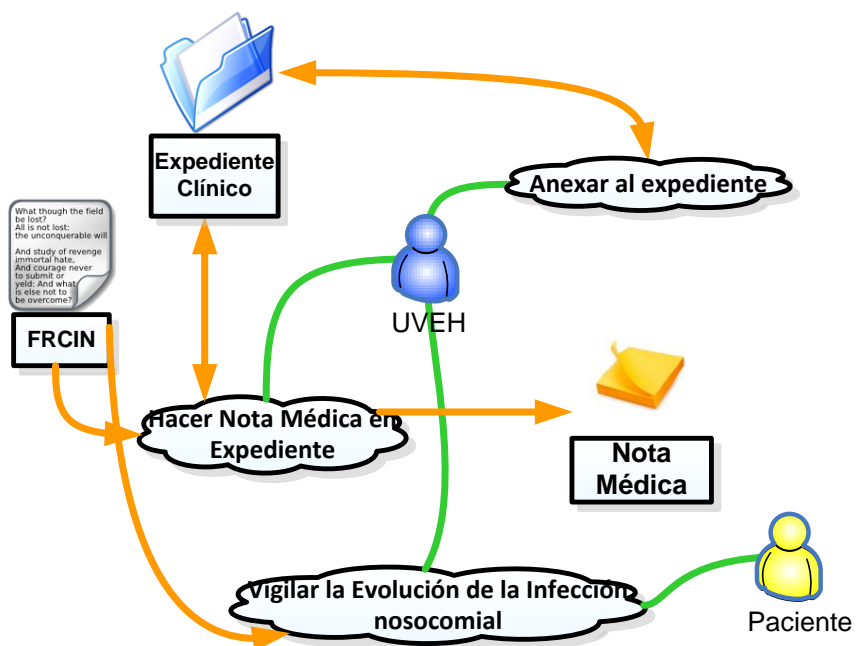


Figura 14. Grafica rica para el sub-proceso de información de infecciones nosocomiales.

A continuación se presenta la grafica rica del sub-proceso de control.

#### IV.3.4. Grafica rica del sub-proceso de control de IN

En Figura 15 se muestra la grafica rica para el proceso de control el cual consiste en la realización del informe mensual.

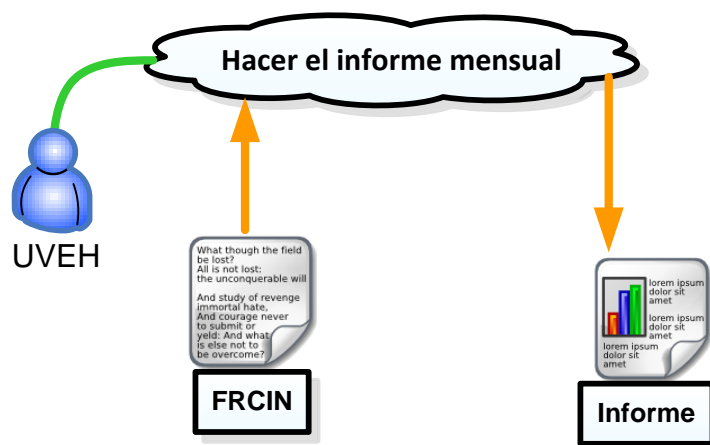
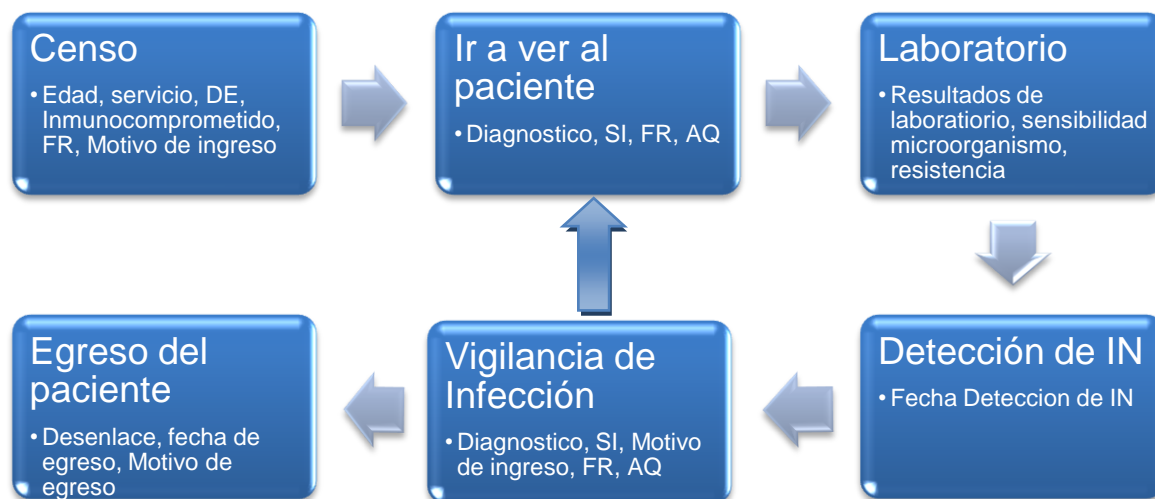


Figura 15. Grafica rica ara el sub-proceso de control de infecciones nosocomiales.

Debido a que la información del paciente es muy dinámica, ya que se va generando conforme a su estadía en el hospital, se realizó una grafica para representar los elementos de información conforme a los eventos que se generan.

#### IV.3.5. Grafica de generación de información de acuerdo a eventos



*Figura 16. Grafica de generación de información de acuerdo a evento durante todo el proceso de detección y vigilancia de infecciones nosocomiales.*

En la *Figura 16*, se muestran que los elementos fundamentales del FRCIN, donde se pueden ver los elementos de información, asimismo muestra que es un ciclo hasta que el egreso del paciente.

#### IV.4. Requerimientos de diseño del CDSS

A partir de la metodología anterior. Obtuvimos los elementos para un diseño de un CDSS. El cual fue presentado de forma preliminar a los dueños del proceso para su validación. A continuación se presentan los requerimientos generales del CDSS.

#### **IV.4.1.Requerimientos del CDSS**

- Contar con mecanismos que permitan organizar, almacenar y recuperar la información relevante para el contexto de decisión.
- Tener un mecanismo administrador de modelo de decisión base que permita un fácil acceso a los diferentes modelos necesarios para el contexto del problema, así como facilitar la entrada de parámetros.
- Contar con mecanismo que permitan ejecutar los modelos para la toma de decisiones.
- Los modelos deben de ser completados con los datos proporcionados por el usuario
- Contar con un mecanismo que permita administrar el acceso a las diferentes bases de conocimiento que se tengan.
- Contar con mecanismos de transformación de datos para que la información sea acorde a la entrada a los modelos de decisión.

Una vez que se conocen los requerimientos básicos que la herramienta a desarrollar debe de tener, a continuación se muestran los requerimientos obtenidos del caso de estudio.

#### **IV.4.2.Requerimientos resultantes del caso de estudio.**

Como ya se ha mencionado, para que los médicos acepten utilizar una herramienta que les apoye a tomar decisiones que reflejen su profesionalismo y responsabilidad, esta requiere se lo más transparente posible y estar embebida dentro de flujo de trabajo.

Entonces a partir del caso de estudio se puede observar que la UVEH tiene necesidad de tomas de decisión para la detección oportuna de IN, al igual de manejo de la información del paciente. Asimismo el personal de la UVEH tiene necesidades de movilidad debido a que tiene que hacer el recorrido de paciente y tiene que registrar todos los síntomas y datos clínicos generados por su estadía en el hospital.

El principal artefacto dentro del flujo de trabajo del proceso de detección y vigilancia de infección nosocomial es el FRCIN el cual se utiliza durante todo el proceso, por esa razón nuestra propuesta tecnológica se basa en ese artefacto. La propuesta consta de tres componentes: sistema FRCIN, CDSS Móvil, y la extracción de conocimiento de la base de datos del sistema FRCIN. A continuación se mostraran los requerimientos que tienen que tiene el CDSS en los sistemas FRCIN y el CDSS móvil.

Los requerimientos de alto nivel que fueron determinados en base al caso de estudio los cuales se muestran a continuación.

#### **Proceso de extracción de conocimiento en bases de datos**

- Se requiere soporte para el manejo de la información histórica utilizada para obtener los modelos predictivos.
- Se requiere una herramienta para el pre-procesamiento de los datos y convertirlos a un formato adecuado para aplicar el proceso de extracción de conocimiento.

#### **Sistema Móvil**

- Que tenga implementados mecanismos para obtener información de pacientes previamente capturados en el sistema FRCIN.
- Que permita capturar toda la información sobre el paciente la cual está basada en la FRCIN
- Que tenga implementados mecanismos para la administración de datos en caso de que no se tenga acceso a la base de datos del sistema FRCIN.
- Que permita agregar a un paciente y en caso de que exista actualizar la información capturada por el sistema móvil en la base de datos del sistema FRCIN
- Que tenga los módulos de valoración de acuerdo a los diversos eventos de generación de información.



### Sistema FRCIN

- Desarrollado para ambiente Web, debido a que el personal médico ya tiene capacitación para el manejo de sistemas de este tipo.
- Que permita almacenar todos los elementos de información contenidos en el FRCIN, cómo lo son los datos básicos del paciente, diagnósticos, antecedentes quirúrgicos, sitios de infección, análisis de laboratorio, factores de riesgo.
- Que permita almacenar la información del paciente conforme se va generando
- Que de soporte en el sub-proceso de control, para crear el informe de infecciones nosocomiales.

A continuación se presenta la propuesta del soporte al PDVIN, tomando en consideración la lista de requerimientos anteriores.

### IV.5. Propuesta de soporte al PDVIN

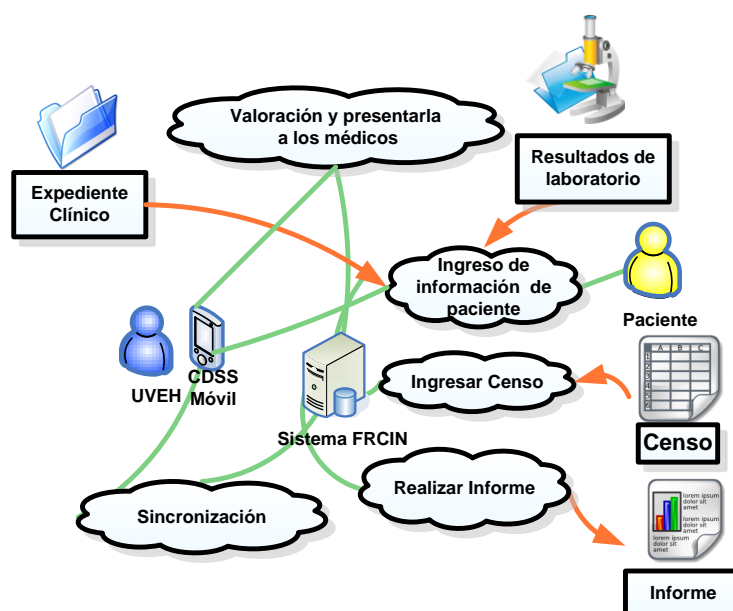


Figura 17. Soporte propuesto al PDVIN

En la *Figura 17* se muestra una propuesta para brindarle soporte al proceso al PDVIN, en ella se muestra la interacción de los dos sistemas por medio de la actividad sincronización, donde se actualizan las bases de datos, por otro lado en la actividad ingreso de información donde se da entrada a los datos clínicos del paciente para posteriormente iniciar la actividad de valoración donde se utilizan estos datos para calcular el riesgo que tiene ese paciente de contraer una infección nosocomial.

Finalmente, para brindarle soporte al sub-proceso de control el sistema FRCIN tiene la opción para realizar el reporte mensual de infecciones nosocomiales.

El FRCIN contiene toda la experiencia y conocimiento la cual es la base para aplicar el proceso de extracción de conocimiento el cual se detallara en el próximo capítulo.

#### **IV.6. Resumen.**

En este capítulo se identifico un escenario clínico real en el hospital con los siguientes requerimientos: se tomen decisiones, tengan una memoria organizacional y que tengan necesidades de movilidad, entonces el escenario que cumplía con estos requerimientos es el PDVIN

Posteriormente se inicio la captura del proceso en el cual se realizaron las siguientes actividades: 12 entrevistas, 4 sesiones de observación. En base a esas actividades se obtuvo una descripción textual del proceso, con el cual se modelo con graficas ricas, diagramas de flujo de datos y diagramas de rol actividad.

Con esos productos se identificaron al jefe de departamento de epidemiología y a la epidemióloga como los tomadores de decisión, adicionalmente se identifico el formato de registro de caso de infección nosocomial (FRCIN) como la memoria organizacional la cual es la base para la extracción de conocimiento.

Adicionalmente se identifico que el llenado del FRCIN es muy dinámico debido a que la información del paciente se genera conforme se encuentra en el hospital, por esta razón se dividió el llenado de la FRCIN en eventos de generación de información.

En base a lo anterior se obtuvo un documento de requerimientos con una perspectiva socio-técnica y se propone un escenario aumentado de tecnología para la detección oportuna de infecciones nosocomiales. En el próximo capítulo se describe el proceso de extracción de conocimiento aplicado a este contexto con el objetivo de obtener modelos predictivos para la detección oportuna de infecciones nosocomiales.

---

## Proceso de Extracción de Conocimiento

---

### V.1. Introducción

En este capítulo se presenta el proceso de extracción de conocimiento aplicado a nuestro caso de estudio presentado en el capítulo IV. El proceso general corresponde a los pasos de extracción del conocimiento en bases de datos de la metodología presentada en el capítulo III. Primero mostraremos los eventos de generación de información en las etapas del proceso a las cuales se aplicara el proceso de extracción de conocimiento y obtener modelos predictivos. Continuaremos con el proceso de procesamiento previo de la información. Se procede a aplicar una metodología de reducción de atributos y se obtiene los diferentes modelos obtenidos por medio de los métodos de *Naive Bayes* y arboles de decisión.

A continuación se presenta el proceso de extracción de conocimiento que seguimos en nuestro trabajo.

### V.2. Proceso de extracción de conocimiento.

La *Figura 18*, muestra el proceso de extracción de conocimiento aplicado al contexto del problema, se inicio con la base de datos obtenida a partir del FRCIN, el siguiente paso es establecer el objetivo del proceso de extracción de conocimiento el cual consiste en obtener modelos predictivos para la detección oportuna de infección nosocomial en las diferentes etapas de generación de información, en base al objetivo se realiza una selección inicial de atributos de la base de datos para extraer el conocimiento.

Para que los métodos de minería de datos sean eficientes es necesario la creación de una vista minable a partir del procesamiento previo de la información, con la vista minable se realiza un análisis de exploración para seleccionar entre los diversos métodos de minería de datos los que generaban modelos predictivos más robustos, de este análisis se eligió el método de *Naive Bayes* y el de árboles de decisión (C4.5). El siguiente paso fue el de dividir la vista minable en conjuntos de datos de acuerdo a los eventos de generación de información.

Teniendo los conjuntos de datos se inicio el proceso de transformación de los datos y generalización de conceptos (ver *Figura 18 A*) lo que se produce el conjunto de datos transformado.

Con este conjunto se inicia el proceso de reducción de atributos (ver *Figura 18 B*) con el objetivo de disminuir la dimensionalidad de la vista minable. Esto consiste en la eliminación de los atributos menos relevantes para la clase objetivo, que es el atributo desenlace.

En el proceso de reducción de atributos aplicamos el método de minería de datos lo que produce un modelo predictivo y su respectiva datos de evaluación, posteriormente se realiza una valoración para conocer la relevancia del conjunto de atributos con respecto al atributo objetivo con el conjunto en base a este análisis se eliminan los atributos menos relevantes y se vuelve aplicar el método de minería de datos para ver si el modelo predictivo resultante mejoro, este proceso es iterativo y finaliza cuando el modelos ya no puede mejoran. Los modelos obtenidos en este modelo son almacenados en un depósito de modelos.

Si al conjunto de datos se le puede aplicar una operación de transformación que puede mejorar el desempeño del modelo predictivo entonces se inicia nuevamente el proceso, en caso de que no, se inicia con el próximo paso, el cual es la selección de modelos definitivos.

Para la selección de los modelos definitivos (ver *Figura 18 C*) se compararon los diversos modelos obtenidos en base a sus datos de evaluación como lo es el porcentaje de instancias clasificadas correctamente, la precisión y alcance de cada una de las clases. En base a esos criterios se eligieron los modelos predictivos. El FRCIN es la base de conocimiento para obtener los modelos predictivos, para adquirir estos modelos se utilizo Weka (Witten y Frank, 2005).

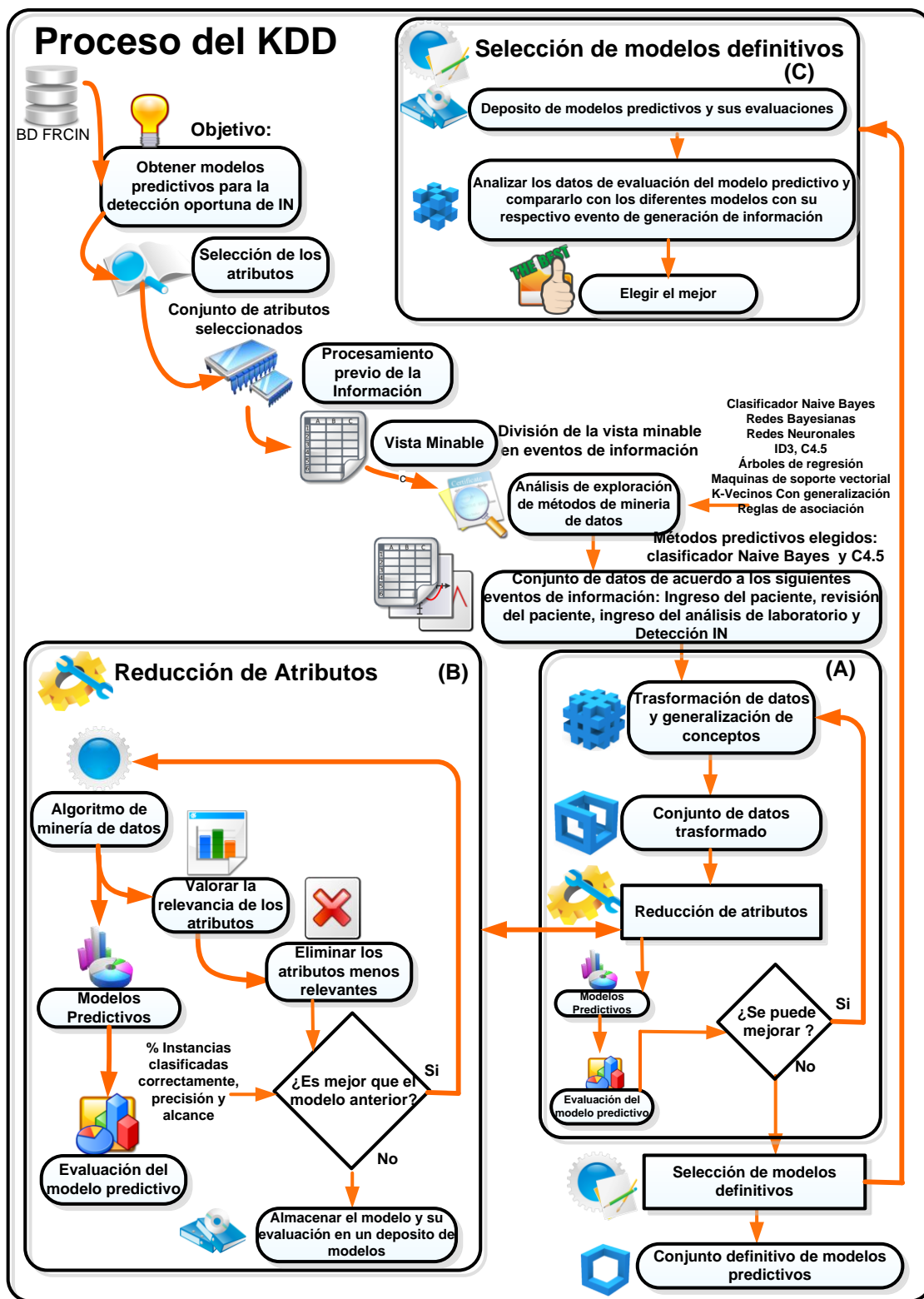


Figura 18. Proceso global de extracción del conocimiento.

### **V.3. Herramienta para la extracción de conocimiento**

Weka es una colección del estado de arte de algoritmos de maquinas de aprendizaje y herramientas para el pre-procesamiento de datos, de tal forma que pueden ser aplicados rápidamente a un conjunto de datos de formas muy flexibles.

Proporciona una extensiva ayuda durante todo el proceso experimental del KDD, incluyendo la preparación de la entrada de datos, evaluación del aprendizaje con esquemas estadísticos y visualización de la entrada y los resultados.

Weka fue desarrollado por la universidad de Waikato en nueva Zelanda, implementado en java y distribuido bajo la licencia de GNU (General public license). Se ejecuta en la mayoría de las plataformas de sistemas operativos como los son Linux, Windows y Macintosh.

A continuación se presenta el proceso de generación de información en el proceso del PDVIN, los cuales se identificaron a partir del caso de estudio.

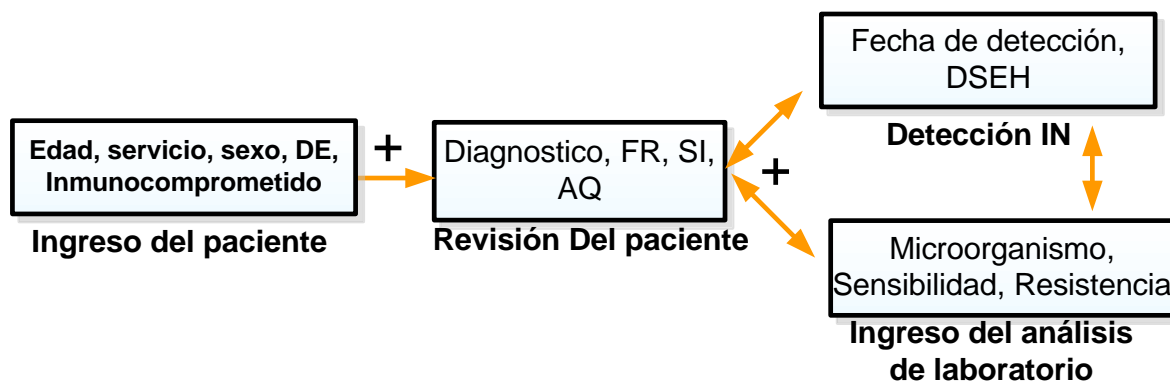
### **V.4. Generación de información del PDVIN.**

En la sección IV.3.5, se mostró el proceso de generación de información del paciente se realiza de forma dinámica y conforme se encuentra en la estadía del hospital, por esta razón se decidió dividir esta generación de información en eventos. Este paso corresponde a los pasos de entender el dominio de la aplicación y selección de atributos de la metodología de extracción de conocimiento y al mismo paso de la *Figura 18*. Los eventos son los siguientes: Ingreso del paciente, Revisión del paciente, Ingreso de análisis de laboratorio, Detección de IN.



El evento de ingreso del paciente, ocurre cuando se tienen disponibles los datos de ingreso del paciente, como lo son: Edad, género, servicio a la que pertenece, días de estancia (DE), diagnóstico, si el paciente es inmunocomprometido y el motivo de ingreso. El evento de revisión del paciente ocurre cuando el médico se dirige hacia el paciente y lo examina, la información que se genera es la siguiente: sitio de infección (SI), Factores de riesgo (FR), Antecedentes Quirúrgicos (AQ). En el evento de ingreso de análisis de laboratorio en el cual ocurre cuando el médico tiene disponible los resultados del laboratorio, que previamente solicitó, en los resultados se dispone la siguiente información microorganismo, sensibilidad, resistencia. En la fase de detección de IN ocurre cuando un caso es confirmado y de esta forma se tiene la fecha de detección de la infección lo que nos da los días de sobre-estancia hospitalaria (DSEH). Los últimos tres eventos pueden llegar a repetirse hasta el egreso del paciente.

En base a estos eventos y conjuntos de atributos se aplicó el proceso de extracción de conocimiento, con el propósito de obtener modelos predictivos que tengan como entrada los datos que se tengan disponibles en cada evento, lo cual permitirá tener una valoración más precisa.



*Figura 19. Eventos de generación de información.*

Para realizar este proceso de extracción de conocimiento se contó con 264 casos del FRCIN, el atributo objetivo es el desenlace, la distribución del mismo es la siguiente: 33 casos pertenecen a la clase finado y 231 pertenecen a la clase de alta por mejoría.

La clase finado solo representa el 12.5% del todo el conjunto de datos, entonces se creó un conjunto especial de datos, para ello se procedió eliminar casos de la clase alta, dejando la siguiente distribución: 21% para la clase finado y 79% para la clase alta, a este conjunto le llamamos conjunto balanceado. En la *Tabla III* se encuentra la información sobre los atributos disponibles en el FRCIN.

*Tabla III. Información en la forma FRCIN*

Atributo	Tipo	Dominio	C
<b>Edad</b>	Numérico	$Edad > 0$ y $edad < 120$	1
<b>Inmunocomprometido</b>	Nominal	{si, no}	1
<b>Sexo</b>	Nominal	{Hombre, Mujer}	1
<b>Servicio</b>	Nominal	{medicina interna, pediatría, cirugía y trauma, ginecología y obstetricia, unidad de cuidados intensivos}	1
<b>Fecha de ingreso (FI)</b>	Fecha	$FI \geq fecha\ actual$	1
<b>Fecha egreso (FE)</b>	Fecha	$FE \geq FI$	1
<b>Fecha reingreso (FRI)</b>	Fecha	$FRI \geq FE$	1
<b>Fecha de reingreso (FRE)</b>	Fecha	$FRE \geq FRI$	1
<b>Fecha Detección de IN (FD)</b>	Fecha	$FD \geq FI$	1
<b>Desenlace</b>	Nominal	{Alta mejoría, finado, alta voluntaria, traslado a otra unidad}	1
<b>Causa de muerte</b>	Texto	{a, ..., z}, {A, ... Z}	1
<b>Sitio de infección (SI)</b>	Nominal	{Infecciones del tracto respiratorio, Infecciones de vías respiratorias bajas, Infecciones cardiovasculares, Diarrea, Infecciones de vías urinarias, Infecciones del sistema nervioso central, Infecciones oculares, Infección de piel y tejidos}	[1,5]

		<i>blandos, Bacteremias, Infecciones de sitio de inserción de catéter, Flebitis, Infección de heridas quirúrgicas, Peritonitis no quirúrgica, Endometritis, Infecciones transmitidas por transfusión o terapia con productos derivados del plasma, Infección transmitida por productos humanos industrializados, Enfermedades exantemáticas, Fiebre postoperatoria, Tuberculosis, Otras_infecciones}</i>	
<b>Motivo de ingreso</b>	Texto	{a, ..., z}, {A, ... Z}	1
<b>Diagnostico</b>	Texto	{a, ..., z}, {A, ... Z}	[1,5]
<b>Factores de riesgo (FR)</b>	Nominal	<i>{Herida Quirurgica, Sobrestancia Hospitalaria, Sondas Vesicales, Venocclisis, Cateter Venoso Central, Cateter Peritoneal, Ventilador Automatico, Edad, Inmunocomprometido, Antibioticoterapia, Inmovilizacion, Cancer, Diabetes Mellitus, Hipertencion Arterial Sistemática, Insuficiencia renal sistemática, Cirugia, Sonda Foley, Obesidad, Otros}</i>	[1,10]
<b>Antecedentes quirúrgicos (AQ)</b>			[0,3]
<b>Fecha</b>	Fecha	<i>fecha &lt; 3 años de última cirugía y fecha &gt; fecha actual</i>	1
<b>Tipo</b>	Nominal	<i>{Electiva, Urgencia, No presente}</i>	1
<b>Grado Contaminación</b>	Nominal	<i>{Limpia, limpia con implante, limpia contaminada, sucia, contaminada, no clasificada }</i>	1
<b>Nombre</b>	Texto	{a, ..., z}, {A, ... Z}	1
<b>Información microbiológica</b>			[0,3]
<b>Tipo de cultivo</b>	Texto	{a, ..., z}, {A, ... Z}	1
<b>Fecha</b>	Fecha	<i>Fecha ≥ fecha actual</i>	1
<b>Germen</b>	Nominal	<i>Se encuentran en el anexo B</i>	[1,5]
<b>Sensibilidad</b>	Nominal	<i>Se encuentra en el anexo B</i>	[1,5]

A continuación se muestra el pre-procesamiento de la información.

### V.5. Procesamiento previo de la información y creación de la vista minable.

El procesamiento previo de la información corresponde al paso con el mismo nombre de la metodología presentada en la sección III.2.1 y con respecto a la *Figura 18*. Para el procesamiento previo y limpieza se desarrolló una aplicación para la creación de una vista minable a partir de la base de datos del sistema FRCIN, la cual consta de 33 entidades de información, las cuales deben estar contenidas en una sola entidad de información y ajustarlo al formato ARFF (-por sus siglas en ingles Attribute Relation File Form-) el cual es el formato que utiliza Weka. La vista minable consta de los siguientes atributos: {*Folio, servicio, DSEH, DE, Edad, sexo, SI1, SI2, SI3, SI4, SI5, Enfermedad1, Enfermedad2, Enfermedad3, Enfermedad4, Enfermedad5, FR1, FR2, FR3, FR4, FR5, FR6, FR7, FR8, FR9, FR10, Diagnostico1, Diagnostico2, Diagnostico3, Diagnostico4, Diagnostico5, Cirugía, Tipo, Contaminación, Microorganismo, FR\_Obesidad, FR\_Tratamiento, FR\_Otros, Neumonía, Desenlace*}. A continuación se presentara el proceso y transformación de la información.

### V.6. Transformación de datos y generalización de conceptos

La teoría transformación de datos y generalización de conceptos se describen en la sección III.2.2. Asimismo para realizar estas operaciones fuimos asistidos por el jefe de departamento de epidemiología del hospital. De este proceso genero varios conjuntos de datos los cuales pueden ver en la *Tabla IV*.

*Tabla IV. Conjuntos de datos generados durante el proceso del pre-procesamiento de la información*

Conjunto	Características
I	<i>Generalización del Desenlace, Discretización de DE, Edad</i>

<b>II</b>	<i>Generalización de conceptos de los atributos SI y FR</i>
<b>III</b>	<i>Trasformación de FR, generalización de conceptos de DE, Edad y Ingreso del atributo cirugía y microorganismo.</i>
<b>IV</b>	<i>Trasformación de SI</i>
<b>V</b>	<i>Mejoría de FR y creación de Enfermedades Degenerativas y neumonía.</i>
<b>VI</b>	<i>Creación de atributos tipo y contaminación</i>
<b>VII</b>	<i>Conjunto de casos de cirugía</i>
<b>VIII</b>	<i>Modifica a los atributos Cat_Edad</i>
<b>IX</b>	<i>Modificación a Cat_DE y Cat_DSEH</i>
<b>X</b>	<i>Conjunto de casos de solo cirugía versión II</i>
<b>XI</b>	<i>Se utilizó el conjunto balanceado y se aplicaron las operaciones de los conjuntos VIII y IX</i>

En la *Tabla IV* se muestra la evolución del conjunto de datos, en el conjunto I se generalizó el atributo desenlace quedando solo dos clases las cuales se pueden observar en la *Tabla V*, adicionalmente se aplicaron técnicas de discretización en los atributos de edad y días de estancia dividiéndolo en cuatro clases, para el conjunto de datos II se aplicaron técnicas de generalización de conceptos al atributos de SI y FR.

**Tabla V. Generalización del atributo desenlace**

<b>Desenlace</b>	<b>Valor:</b>
<b>Bajo riesgo</b>	<i>Alta mejoría, alta voluntaria, traslado a otra unidad</i>
<b>Alto riesgo</b>	<i>Finado o fallecido</i>

Para el atributo SI se generalizó de acuerdo al sistema físico donde se encuentra ubicada la infección, en la *Tabla VI* se muestra el valor generalizado que puede adquirir el atributo SI de acuerdo al valor original de SI, entonces antes del procedimiento el atributo SI podía adquirir 25 valores distintos de los cuales se redujeron a 8.

*Tabla VI Generalización de conceptos del atributo sitio de infección.*

<b>Sitio de infección</b>	<b>Valor:</b>
<b>Sistema respiratorio</b>	<i>Infecciones del tracto respiratorio, Infecciones de vías respiratorias bajas, Tuberculosis</i>
<b>Sistema Cardiovascular</b>	<i>Infecciones cardiovasculares</i>
<b>Sistema digestivo</b>	<i>Tracto digestivo</i>
<b>Sistema genitourinario</b>	<i>Infecciones de vías urinarias, Endometritis</i>
<b>Sistema nervioso central</b>	<i>Infecciones del sistema nervioso central</i>
<b>Organos y tejidos</b>	<i>Infecciones oculares, Infección de piel y tejidos blandos, Infecciones de sitio de inserción de catéter, Flebitis, Infección de heridas quirúrgicas, Peritonitis no quirúrgica, Enfermedades exantemáticas, Infección de heridas quirúrgicas, Infecciones de sitio de inserción de catéter, túnel o puerto subcutáneo</i>
<b>Hemático</b>	<i>Bacteremias, Infecciones transmitidas por transfusión, terapia con productos derivados del plasma, Infección transmitida por productos humanos industrializados, Infección transmitida por productos humanos industrializados (de origen no sanguíneo) o por injertos</i>
<b>Otras infecciones</b>	<i>Otras infecciones</i>

El atributo FR se generalizó en tres conceptos o clases: asociadas al paciente, asociadas al tratamiento y otros, las asociadas al paciente corresponden a las condiciones que tiene que ver con el paciente como los son: Obesidad, Cáncer, Diabetes Mellitus, Hipertensión arterial sistemática, Insuficiencia renal crónica, Inmunocomprometido, Edad; las asociadas al tratamiento hacen referencia a los factores de riesgo asociados al tratamiento como lo

son: Inmovilización, Sobre-estancia Hospitalaria, Antibioticoterapia, Venoclisis, Catéter Venoso Central, Catéter Peritoneal, Sondas Vesicales, Herida Quirúrgica, Ventilador automático, Cirugía, Sonda Foley.

Entonces los atributos de FR solo podían adquirir tres valores: asociadas al paciente, asociadas al tratamiento y otros, por lo tanto los atributos FR podían adquirir mismo valor lo cual no producía una correcta discriminación de esa variable debido a que los 10 atributos de FR podían adquirir el mismo valor, por esa razón se decidió realizar la transformación del atributo FR.

La transformación de FR en el conjunto III, consistió en la creación de tres atributos: FR\_Paciente, FR\_Tratamiento y FR\_otros, Estos atributos corresponden a la generalización que se realizó anteriormente, pueden adquirir un valor de “sí” o “no” en caso de que uno de los atributos de FR tuviera o no tuviera factores de riesgo asociados al paciente, tratamiento y otros. En el conjunto III se realizaron generalización de conceptos en los atributos de edad y DE.

La primera versión de la generalización de conceptos del atributo edad consistió en dividirlo en 4 clases las cuales corresponden a la *Tabla VII*, el atributo DE se generalizó en tres clases las cuales corresponden a la

*Tabla VIII*. Adicionalmente al conjunto III se le agregó el atributo cirugía el cual puede adquirir el valor de “sí” o “no” de acuerdo si el paciente tiene antecedentes quirúrgicos, también se creó el atributo microorganismo el cual puede adquirir los valores de “sí” o “no” de acuerdo si el paciente tiene resultados positivos en sus análisis de laboratorio.

*Tabla VII. Generalización de conceptos del atributo edad.*

Edad	Valor:
Niños	0 a 9

<b>Adolescentes</b>	<i>10 a 19</i>
<b>Adulto</b>	<i>20 a 59</i>
<b>Adulto mayor</b>	<i>Mayor a 59</i>

*Tabla VIII. Generalización de conceptos del atributo DE.*

<b>DE</b>	<b>Valor:</b>
<b>Estancia baja</b>	<i>0 a 4</i>
<b>Estancia media</b>	<i>5 a 9</i>
<b>Sobre-estancia</b>	<i>Mayor a 9</i>

En el conjunto de prueba IV se crearon los atributos: SI\_Sistema\_Respiratorio, SI\_Cardiovascular, SI\_Digestivo, SI\_Genitourinario, SI\_Sistema\_Nervioso, SI\_Organos\_Tejidos, los cuales puede adquirir el valor de “Si” o “No” en caso de presentar una infección en su correspondiente generalización.

En el conjunto de prueba V se observó que los valores del factor de riesgo que están asociadas al pacientes ya están considerados como atributos como es el caso de la edad y inmunocomprometido, entonces se decidió quitarlos y crear dos atributos Enfermedades\_Degenerativas y FR\_Obesidad. También se decidió quitar en los valores asociadas al tratamiento la sobre-estancia debido a que se cuenta con el atributo DE. Entonces los valores asociadas a cada atributo se representan en la *Tabla IX*. En este mismo conjunto se creó el atributo neumonía, a partir del atributo enfermedad, este atributo puede adquirir el valor de “Si” o “No”.

*Tabla IX. Generalización de conceptos del atributo FR.*

<b>Atributo</b>	<b>Valor:</b>
<b>Enfermedades</b>	<i>Cáncer, Diabetes Mellitus, Hipertensión arterial</i>



<b>degenerativas</b>	<i>sistemática, Insuficiencia renal crónica</i>
<b>Factor de riesgo tratamiento</b>	<i>Inmovilizacion, Antibioticoterapia, Venoclisis, Cateter Venoso Central, Cateter Peritoneal, Ventilador Automatico, Sonda Foley, Sondas Vesicales</i>
<b>Factor riesgo obesidad</b>	<i>Obesidad</i>
<b>Otros</b>	<i>Otros factores de riesgo</i>

En el conjunto de VI se ingresaron los atributos tipo y contaminación las cuales corresponden a la información contenida a los antecedentes quirúrgicos del paciente, no se tomaron en consideración los casos en el que el tipo es “*No presente*”, entonces solo se generalizo el atributo contaminación de acuerdo a la *Tabla X*.

*Tabla X. Generalización de conceptos del atributo contaminación.*

<b>Contaminación:</b>	<b>Valor</b>
<b>Limpia</b>	<i>Limpia, limpia con implante</i>
<b>Limpia-contaminada</b>	<i>Limpia- contaminada</i>
<b>Contaminada</b>	<i>Sucia, no clasificada</i>

En conjunto VII solo se encuentran casos de pacientes que tienen en el atributo cirugía un “*Si*”. En el conjunto VIII se realizó una modificación en el atributo Edad debido a que el más del 90% de los casos se encontraban en las clases de “*adulto*” y “*adulto mayor*” se decidió realizar una nueva generalización para una mejor distribución dentro de las clases de edad la cual se puede ver en la *Tabla XI*.

*Tabla XI. Generalización de conceptos del atributo Edad\_2.*

<b>Edad_2:</b>	<b>Valor</b>
<b>Joven</b>	<i>0 a 24</i>

<b>Adulto</b>	<i>25 a 46</i>
<b>Adulto mayor</b>	<i>47 a 69</i>
<b>Tercera edad</b>	<i>Mayor a 69</i>

En el conjunto IX se realizo una nueva generalización en los atributos DE y DSEH por la misma razón que el conjunto anterior, entonces atributo DE se le agrego una clase llamada “*estancia alta*” como se muestra en la *Tabla XII*.

*Tabla XII. Generalización de conceptos del atributo DE\_2.*

<b>DE_2</b>	<b>Valor</b>
<b>Estancia baja</b>	<i>0 a 4</i>
<b>Estancia media</b>	<i>5 a 9</i>
<b>Estancia alta</b>	<i>10 a 15</i>
<b>Sobre-estancia</b>	<i>Mayor a 15</i>

Para el atributo DSEH se le agrego la clase “*Baja*” la distribución de las clases se muestran en la *Tabla XIII*

*Tabla XIII. Generalización de conceptos del atributo DSEH\_2.*

<b>DSEH_2</b>	<b>Valor</b>
<b>Baja</b>	<i>0 a 4</i>
<b>Moderada</b>	<i>5 a 9</i>
<b>Promedio</b>	<i>10 a 15</i>
<b>Elevada</b>	<i>Mayor a 15</i>

En el conjunto X se utilizo el conjunto VII en el cual se las modificaciones realizadas en los atributos Edad, DE y DSEH.

Como se había mencionado anteriormente, el proceso del KDD es un proceso iterativo, entonces para obtener el modelo definitivo se realizaron varias operaciones de pre-procesamiento sobre la vista minable con el objetivo de obtener un modelo predictivo más robusto basado en las estimaciones de rendimiento.

En el conjunto XI se aplicaron las operación de de transformación de atributos y generalización de conceptos las mismas a las aplicadas a los conjuntos VIII y IX. A continuación se presenta el siguiente paso que es la selección de atributos

## **V.7. Reducción de atributos**

En esta parte del KDD se reducirán la dimensionalidad de la vista minable con los siguientes objetivos:

- Eliminar atributos irrelevantes y redundantes.
- Expresar el modelo redundante en función de menos variables.
- Mejorar la calidad del modelo predictivo.

Entonces para reducir la dimensionalidad de los atributos se evaluaron para conocer que tan relevantes son con respecto al atributo desenlace. Las técnicas utilizadas para hacer esta valoración son las siguientes: k-vecinos, Chi-cuadrada y taza de ganancia de información.

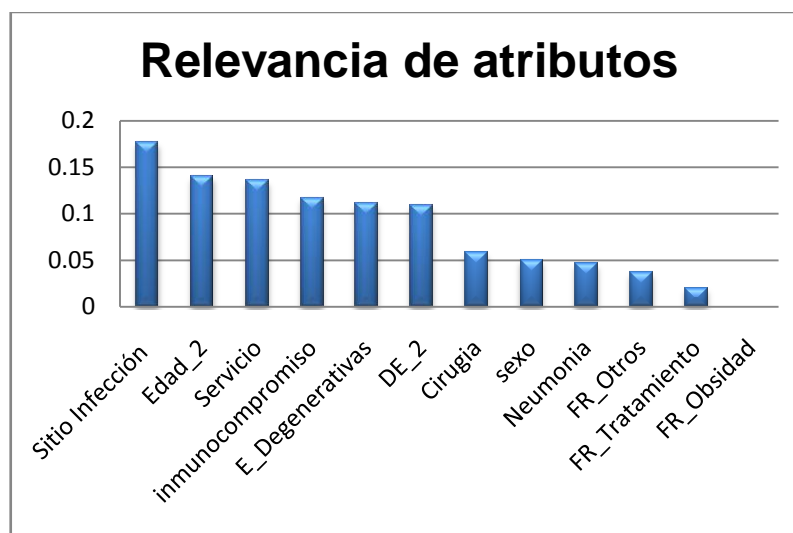
Se utilizo una metodología para la reducción de atributos con los tres métodos, pero el que dio mejor resultados fue el de los k-vecinos (sección III.3.2.4.1). A continuación se describe la metodología para realizar esa reducción de dimensionalidad.

La metodología consiste en lo siguiente: para cada fase de generación de información realizar una valoración con todos los atributos, posteriormente aplicar el método de minería de datos y se obtienen los modelos predictivos. Posteriormente se eliminan los atributos

menos relevantes, y se analiza la evaluación de los modelos para ver su rendimiento buscando la mejor.

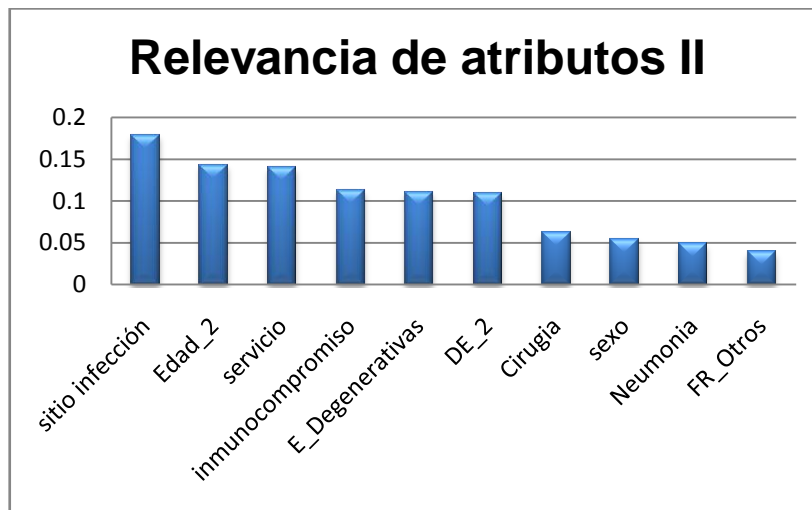
En las medidas de evaluación que se utilizaron son: el porcentaje de instancias clasificadas correctamente, los valores de alcance y precisión de cada una de las clases. Entonces cada vez que se eliminan atributos se aplica el método de minería y se analiza los valores de evaluación, si estos mejoran se sigue con la eliminación de atributo en caso de que no se termina el proceso. A continuación se presentara, como ejemplo la aplicación de esta metodología para la obtención de un modelo en la fase de revisión del paciente con el método *Naive Bayes* en el conjunto IX el cual tiene 81.85% de instancias clasificadas correctamente.

En la *Figura 20* se muestra la relevancia de los atributos de la fase de ingreso del paciente con respecto al atributo desenlace, entonces, como se puede ver los atributos FR\_Obesidad y FR\_Tratamiento, son los menos relevantes para el atributo desenlace.



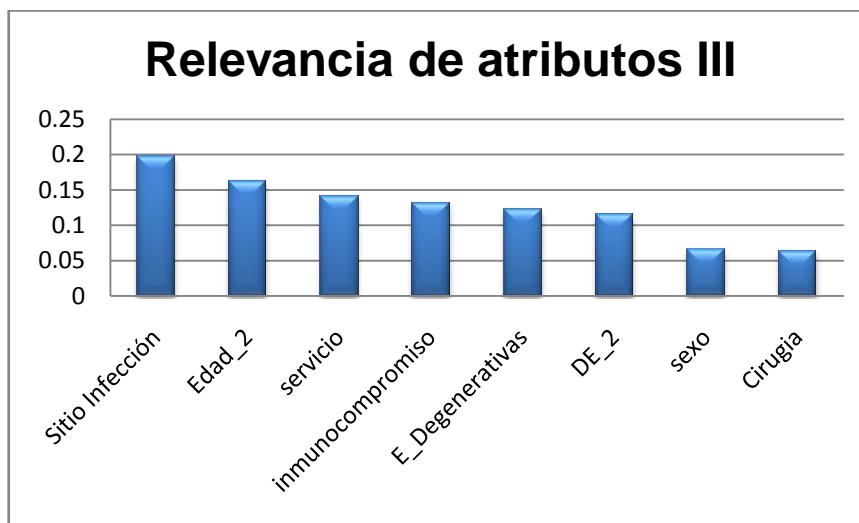
*Figura 20. Relevancia de atributos.*

Por esta razón se procede a eliminarlos y aplicar nuevamente el algoritmo de minería de datos en donde se obtuvo el 82.57%. El nuevo análisis que ahora se presenta en la *Figura 21*.



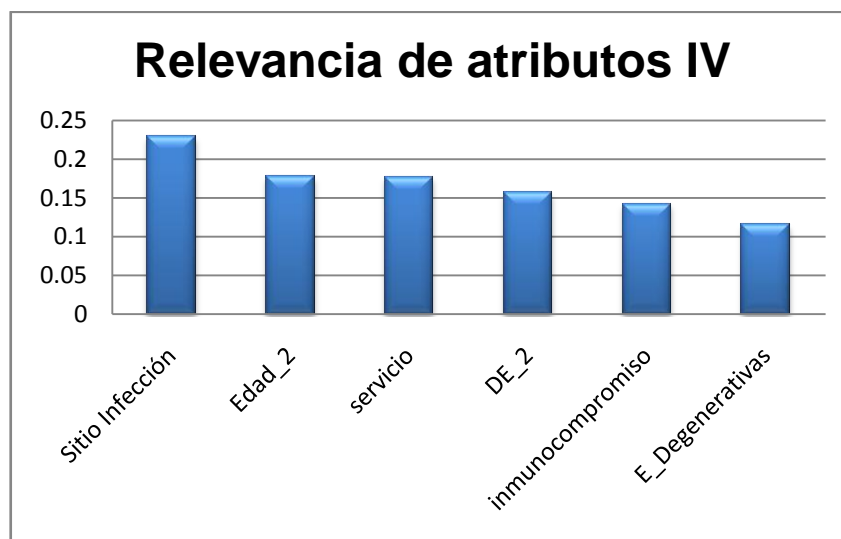
*Figura 21. Relevancia de Atributos II.*

En esta etapa los atributos menos relevantes son: Fr\_otros, neumonía y sexo si los eliminamos nos quedamos con los atributos mostrados en *Figura 22*.



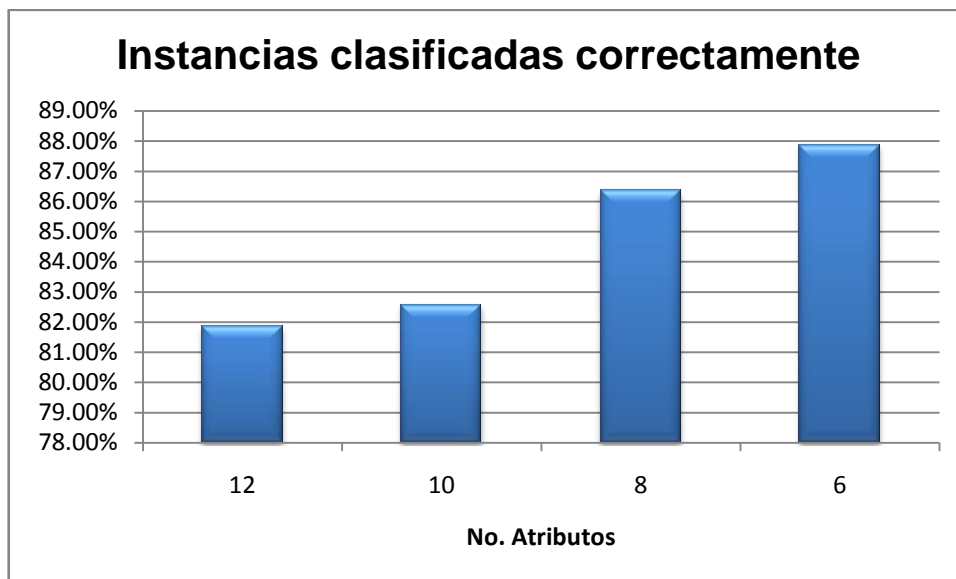
*Figura 22. Relevancia de Atributos III.*

En estos atributos se aplico se aplico el método de minería de datos con en donde se obtuvo un 86.36% de instancias clasificadas correctamente, entonces en la *Figura 22* se puede ver que los dos atributos menos relevantes son cirugía y sexo, por lo tanto se decidió eliminarlos y realizar nuevamente el análisis de relevancia de atributos.



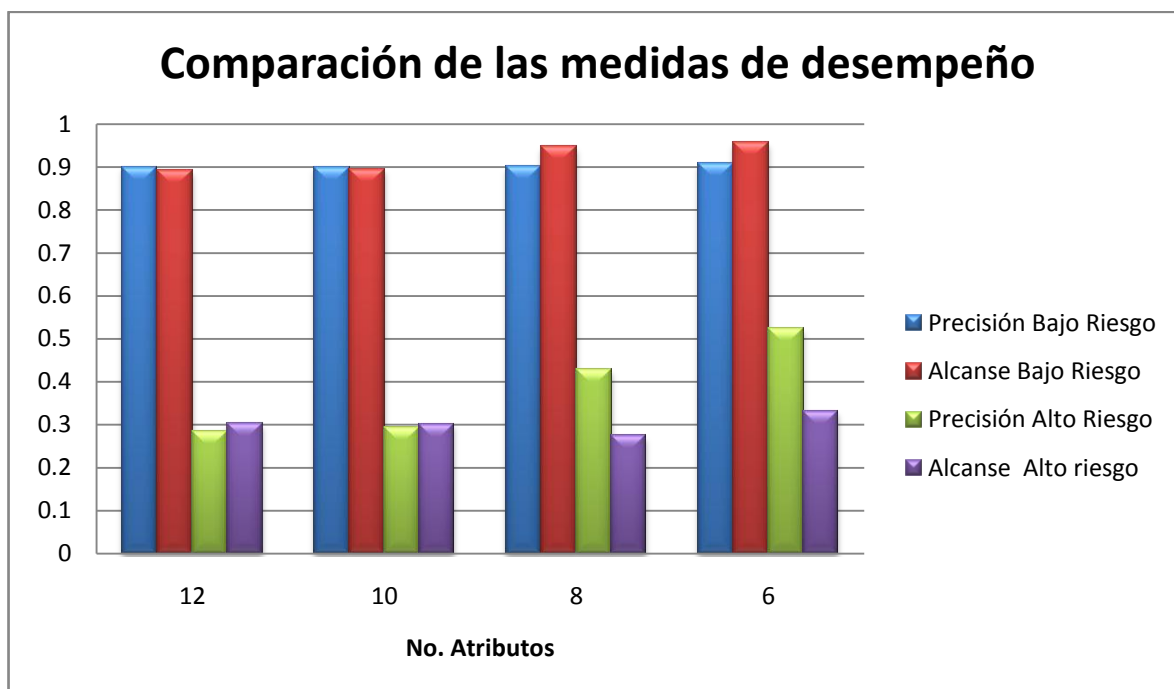
*Figura 23. Relevancia de Atributos IV.*

Al aplicar el método de minería de datos con el conjunto de atributos mostrados en la *Figura 23* se obtuvo un 87.88% de instancias clasificadas correctamente, entonces de acuerdo con la *Figura 23* el atributo menos relevante es el de E\_Degenerativas entonces se procede a eliminarlo, el resultado de instancias clasificadas correctamente después de la eliminación fue de 86.74%. Esto indica que el atributo E\_Degenerativas es relevante para la clasificación y nos indica el fin de la metodología de reducción de atributos.



*Figura 24. Instancias clasificadas correctamente.*

En la *Figura 24* se puede observar el aumento del 8% del porcentaje de las instancias clasificadas correctamente conforme se eliminan los atributos irrelevantes. En la *Figura 25* se muestran las medidas de desempeño conforme se eliminaban atributos, se puede observar que no hay mucha diferencia cuando se tiene doce y diez atributos, pero cuando solo se tiene ocho se puede observar que la precisión en la clase defunción aumenta en un 50%, y cuando se tienen solo seis es aumenta en un 83% con respecto a 10 o 12 atributos.



*Figura 25. Comparación de las medidas de desempeño.*

Se utilizó esta metodología para reducir los atributos redundantes en los diferentes conjuntos de datos de las diferentes fases de generación de información. En la siguiente sección se mostrarán los modelos predictivos obtenidos en las diferentes fases.

## V.8. Análisis de exploración

En esta etapa de exploración se analizaron varios métodos de minería de datos como lo son las redes bayesianas, redes neuronales, árboles de regresión, ID3, C4.5, máquinas de soporte vectorial, K-vecinos de generalización y reglas de asociación. Con este análisis se generaron varios modelos predictivos y se analizaron los datos de evaluación y se eligieron dos tipos de modelos predictivos: los simbólicos y los sub-simbólicos.

Del tipo sub-simbólico se eligió el método de *Naive Bayes* y simbólico se eligió el método de *C4.5*, se eligieron estos métodos ya que cumplían con los objetivos de este



trabajo de investigación el cual es comparar la percepción que tiene los médicos sobre los diferentes métodos de minería de datos, es decir sobre los métodos que pueden mostrar el razonamiento como lo es el árbol de decisión sobre los métodos que son una caja negra como lo es las redes bayesianas. Adicionalmente estos métodos son los que daban mejor resultados de acuerdo a nuestro conjunto de datos.

Adicionalmente, un producto interesante de este análisis de exploración, la de haber obtenido modelos predictivos robustos para la detección de neumonía. Los modelos resultantes de este análisis se pueden ver en el apéndice C.

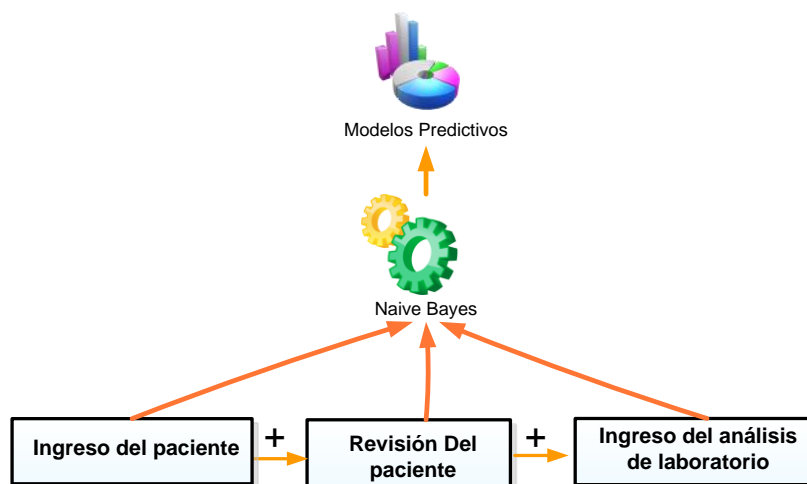
## **V.9. Modelos predictivos definitivos**

En esta sección se presenta los modelos predictivos obtenidos por medio de los modelos de extracción de conocimiento dinámico con *Naive Bayes* (ver *Figura 26*) y con árboles de decisión (*Figura 30*).

A continuación se describirá los modelos basados en el clasificador *Naive Bayes* y posteriormente los modelos basados en arboles de decisión.

### **V.9.1. Modelos predictivos por el clasificador *Naive Bayes***

El modelo de extracción de conocimiento con *Naive Bayes* consiste en las siguientes fases: ingreso del paciente, revisión del paciente y ingreso de análisis de laboratorio. Estas fases se pueden ver la *Figura 26*, en cada fase se genero un modelo predictivo. En bases a los atributos disponibles. A continuación se muestra cada modelo predictivo obtenido en cada fase con el método de *Naive Bayes*.



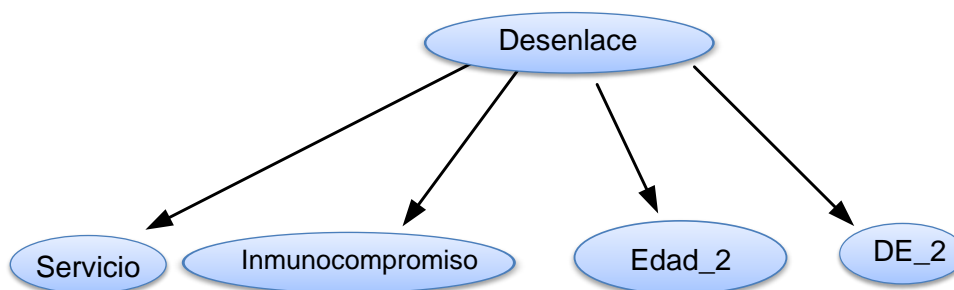
*Figura 26. Modelo de extracción de conocimiento dinámico con el clasificador Naive Bayes.*

#### V.9.1.1. Ingreso del paciente

En la fase de ingreso del paciente se cuentan con los atributos mostrados en la *Figura 19*, los cuales son: edad servicio, sexo, inmunocompromiso y DE, después de aplicar la metodología de reducción se eliminó el atributo sexo. En la *Tabla XIV* se puede ver que el modelo obtenido a partir del conjunto IX tiene un mayor porcentaje de instancias clasificadas correctamente, sin embargo la precisión de la clase alto riesgo es 0, comparada con la precisión de la clase alto riesgo de .6 del conjunto XI, por esta razón se eligió el modelo predictivo del conjunto XI. El modelo resultante se puede ver en la *Figura 27*.

*Tabla XIV. Comparación de modelos en ingreso del paciente*

Porcentaje de instancias clasificadas correctamente	
Conjunto IX	87.00%
Conjunto XI	82.50%



*Figura 27. Clasificador bayesiano en Ingreso del Paciente.*

Cada nodo del clasificador *Naive Bayes* tiene una distribución de probabilidad condicional las cuales dependen del nodo desenlace. En la *Tabla XV* se encuentran la distribución de probabilidad condicional del atributo desenlace

*Tabla XV. Distribución de probabilidad condicional del atributo desenlace.*

Desenlace		
	Bajo Riesgo	Alto Riesgo
	0.78797468	0.21202532

*Tabla XVI. Distribución de probabilidad condicional del atributo inmunocompromiso.*

Inmunocompromiso		
	No	Si
Bajo Riesgo	0.276	0.724
Alto Riesgo	0.10294118	0.89705882

*Tabla XVII. Distribución de probabilidad condicional del atributo Servicio.*

Servicio					
	C.T	G.O	M.I	Pediatría	UCI
Bajo Riesgo	0.44664032	0.02766798	0.47826087	0.04347826	0.00395257
Alto Riesgo	0.21126761	0.01408451	0.71830986	0.01408451	0.04225352

*Tabla XVIII. Distribución de probabilidad condicional del atributo Edad\_2.*

Edad_2				
	Joven	Adulto	Adulto Mayor	3° Edad
Bajo Riesgo	0.13888889	0.25	0.33730159	0.27380952
Alto Riesgo	0.01428571	0.12857143	0.44285714	0.41428571

*Tabla XIX. Distribución de probabilidad condicional del atributo DE\_2.*

	DE_2			
	Estancia Baja	Estancia Media	Estancia Alta	Sobre-estancia
Bajo Riesgo	0.12301587	0.30555556	0.31349206	0.25793651
Alto Riesgo	0.04285714	0.21428571	0.24285714	0.5

A continuación se presenta un ejemplo de uso del clasificador bayesiano

#### V.9.1.1.1. Ejemplo de uso del clasificador bayesiano

Supongamos que se tiene un paciente  $P$  y el paciente presenta inmunocompromiso, y se encuentra en el servicio de medicina interna, tiene 50 años de edad y tiene 10 días de estancia hospitalaria. Entonces con esos datos se puede crear la instancia  $P$  con los siguientes atributos:

$$P = \{inmunocompromiso, servicio, edad, DE\}$$

Sustituyendo con los datos del paciente

$$P = \{si, medicina interna, 50, 10\}$$

Después de aplicar las operaciones de transformación sobre los atributos de edad y DE, queda de la siguiente forma

$$P = \{si, medicina interna, adulto mayor, estancia alta\}$$

Si se desea calcular la probabilidad del paciente de contraer una infección nosocomial ( $p_a$ ) solo se tiene que tomar los valores correspondientes a las tablas. Entonces para el cálculo de  $p_a$  se utiliza la siguiente expresión

$$p_a = 0.89705882 \times 0.71830986 \times 0.24285714 \times 0.24285714 \times 0.21202532 \quad (18)$$

$$p_a = 0.00805791$$

El último término de la expresión (18) corresponde al atributo desenlace de igual forma se precede para calcular la probabilidad del paciente de no contraer una infección nosocomial ( $p_b$ ), entonces  $p_b = .0029503$ . Teniendo estos valores de probabilidad es necesario normalizarlos con la siguiente expresión:

$$P_a = \frac{p_a}{p_a + p_b} \quad (19)$$

$$P_b = \frac{p_b}{p_a + p_b} \quad (20)$$

Entonces la probabilidad del paciente  $P$  de contraer una infección nosocomial es  $P_a = .96467916$  y la probabilidad de no contraer una infección nosocomial es  $P_b = .035320843$

A continuación se presenta el modelo en la fase de revisión del paciente.

#### **V.9.1.2. Fase de Revisión del paciente**

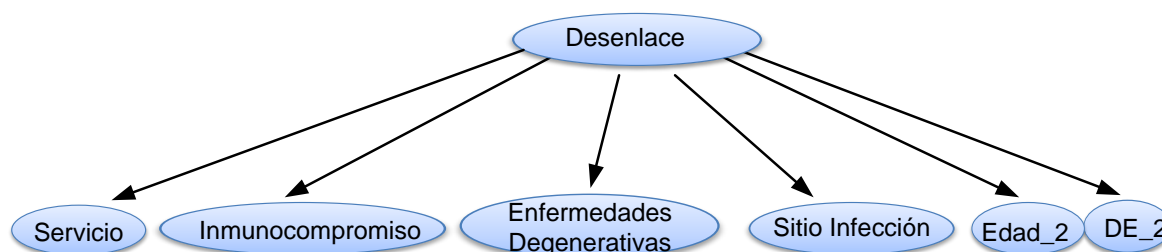
Para obtener un modelo definitivo se utilizaron los conjuntos VIII y IX, en los cuales se aplicó la metodología para la reducción de atributos. Los atributos que se tienen disponibles son los siguientes: servicio, inmunocompromiso, Edad, DE, sexo, cirugía, sitio de infección, FR\_obesidad, FR\_Tratamiento, FR\_otros, Enfermedades\_Degenerativas, Neumonía.

Después de aplicar la reducción de atributos en ambos conjuntos quedaron los siguientes: servicio, inmunocompromiso, enfermedades degenerativas, sitio de infección, edad y DE. Posteriormente se revisaron el porcentaje de instancias clasificadas correctamente, esto se presenta en la *Tabla XX*.

**Tabla XX. Comparación de modelos en revisión del paciente.**

Porcentaje de instancias clasificadas correctamente	
Conjunto VIII	85.98%
Conjunto IX	87.88%

Entonces en la *Tabla XX*, se puede ver que el porcentaje de instancias clasificadas correctamente es mayor en el conjunto IX, por esa razón el modelo final en *Naive Bayes* en la fase de revisión del paciente es la del conjunto IX el modelo resultante se presenta en la *Figura 28*.



**Figura 28. clasificado Naive Bayes en revisión del paciente.**

A continuación se presentan las tablas de distribución de probabilidad condicional:

**Tabla XXI. Distribución de probabilidad condicional del atributo desenlace.**

Desenlace		
	Bajo Riesgo	Alto Riesgo
	0.87358491	0.12641509

**Tabla XXII. Distribución de probabilidad condicional del atributo enfermedades degenerativas.**

Enfermedades Degenerativas		
	No	Si
Bajo Riesgo	0.6012931	0.3987069
Alto Riesgo	0.42647059	0.57352941

**Tabla XXIII. Distribución de probabilidad condicional del atributo inmunocompromiso.**

Inmunocompromiso		
	No	Si

Bajo Riesgo	0.29525862	0.70474138
Alto Riesgo	0.10294118	0.89705882

*Tabla XXIV. Distribución de probabilidad condicional del atributo servicio.*

Servicio					
	C.T	G.O	M.I	Pediatría	UCI
Bajo Riesgo	0.46895075	0.03640257	0.44753747	0.04068522	0.00642398
Alto Riesgo	0.21126761	0.01408451	0.71830986	0.01408451	0.04225352

*Tabla XXV. Distribución de probabilidad condicional del atributo sitio de infección.*

Sitio de infección							
	SR	SC	SD	SG	SNC	SG	SOT
Bajo Riesgo	0.129787	0.0063829	0.002127	0.1255319	0.0021276	0.6914893	0.0063829
Alto Riesgo	0.337837	0.0135135	0.013513	0.1486486	0.0135131	0.2837837	0.0135135

*Tabla XXVI. Distribución de probabilidad condicional del atributo DE\_2.*

DE_2				
	Estancia baja	Estancia media	Estancia alta	Sobre-estancia
Bajo Riesgo	0.1223176	0.28969957	0.29399142	0.29399142
Alto Riesgo	0.04285714	0.21428571	0.24285714	0.5

*Tabla XXVII. Distribución de probabilidad condicional del atributo edad.*

Edad_2				
	Joven	Adulto	Adulto Mayor	3° Edad
Bajo Riesgo	0.13888889	0.25	0.33730159	0.27380952
Alto Riesgo	0.01428571	0.12857143	0.44285714	0.41428571

A continuación se mostrara el modelo obtenido en la fase de ingreso de análisis de laboratorio.

### V.9.1.3. Fase de ingreso de análisis de laboratorio.

Para obtener un modelo definitivo se utilizaron los conjuntos VIII y IX, en los cuales se aplico la metodología para la reducción de atributos. Los atributos disponibles son los

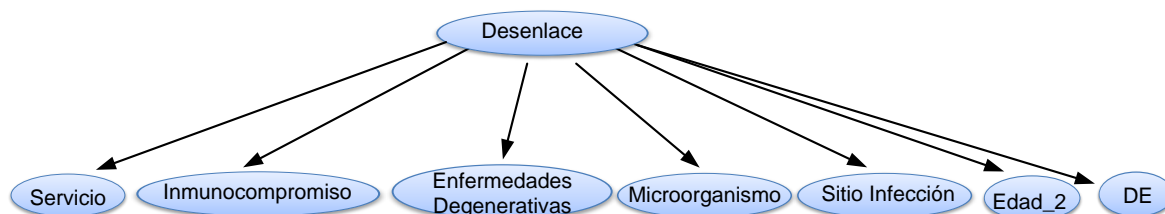
siguientes: servicio, inmunocompromiso, Edad, DE, sexo, cirugía, sitio de infección, FR\_obesidad, FR\_Tratamiento, FR\_otros, enfermedades degenerativas, microorganismo, neumonía.

Después de aplicar la reducción de atributos en ambos conjuntos quedaron los siguientes: servicio, inmunocompromiso, enfermedades degenerativas, microbiología, sitio de infección, edad y DE. Posteriormente se revisaron el porcentaje de instancias clasificadas correctamente, esto se puede ver en la *Tabla XXVIII*.

*Tabla XXVIII. Comparación de modelos en ingreso de análisis de laboratorio.*

Porcentaje de instancias clasificadas correctamente	
Conjunto VIII	84.5%
Conjunto IX	86.36%

Entonces en la *Tabla XXVIII*, se puede ver que el porcentaje de instancias clasificadas correctamente es mayor en el conjunto IX, por esa razón el modelo final con *Naive Bayes* en la fase de ingreso de análisis de laboratorio es la del conjunto IX el modelo resultante se presenta en la *Figura 29*.



*Figura 29. Clasificador Naive Bayes en Ingreso de Análisis de laboratorio.*

A continuación se presentan las distribuciones de probabilidad condicional de los nodos del clasificador bayesiano, en el caso de los nodos de servicio, inmunocompromiso y enfermedades degenerativas, las distribuciones de probabilidad condicional son las mismas que la del modelo anterior.



**Tabla XXIX. Distribución de probabilidad condicional del atributo microorganismo.**

Microorganismo		
	No	Si
Bajo Riesgo	0.22198276	0.77801724
Alto Riesgo	0.30882353	0.69117647

**Tabla XXX. Distribución de probabilidad condicional del atributo DE.**

DE			
	Estancia Baja	Estancia media	Sobre-estancia
Bajo Riesgo	0.1311828	0.29032258	0.57849462
Alto Riesgo	0.04347826	0.2173913	0.73913043

**Tabla XXXI. Distribución de probabilidad condicional del atributo sitio de infección.**

Sitio de infección							
	SR	SC	SD	SG	SNC	SG	SOT
Bajo Riesgo	0.129787	0.0063828	0.002127	0.1255319	0.0021276	0.6914893	0.0063829
Alto Riesgo	0.337837	0.0135135	0.013513	0.1486486	0.0135135	0.2837837	0.0135135

**Tabla XXXII. Distribución de probabilidad condicional del atributo Edad\_2.**

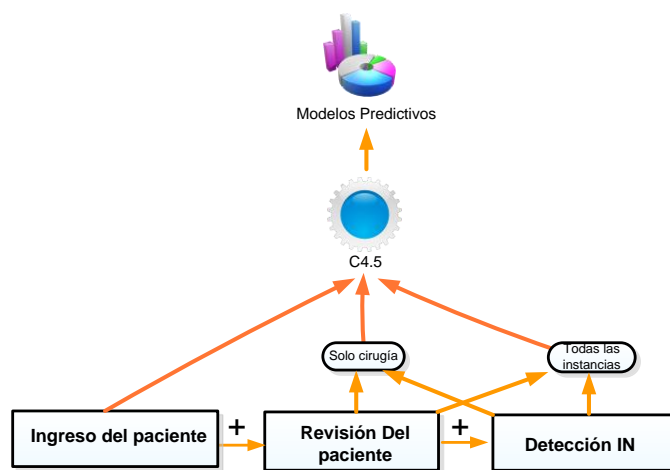
Edad_2				
	Joven	Adulto	Adulto Mayor	3° Edad
Bajo Riesgo	0.13519313	0.2167382	0.36266094	0.28540773
Alto Riesgo	0.01428571	0.12857143	0.44285714	0.41428571

## V.9.2. Modelos predictivos por árboles de decisión.

El modelo de extracción de conocimiento con árboles de decisión consiste en las siguientes fases: ingreso del paciente, revisión del paciente y detección de Infección nosocomial (IN). Como se puede ver en la en la *Figura 30*, las dos primeras fases son iguales a las del modelos de extracción de conocimiento dinámico del Bayesiano sin embargo se creó el conjunto X de datos, el cual consiste en casos de solo cirugías por lo tanto en la fase de revisión del paciente y detección IN se tienen dos modelos uno de ellos a

partir del conjunto de datos de solo cirugías y otro modelos con las instancias disponibles de los conjuntos.

Por otro lado, este modelo tiene el evento de detección de IN, este evento consiste en la valoración del paciente después de la detección de IN. A continuación se muestra los modelos predictivos obtenidos por el método de árboles de decisión en la fase de ingreso de paciente.

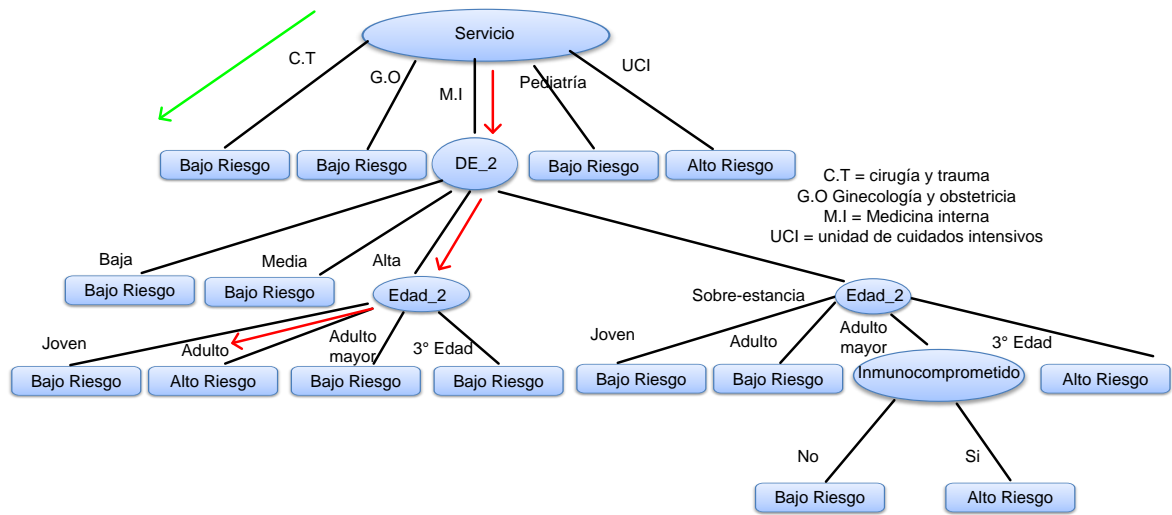


*Figura 30. Modelo de extracción de conocimiento dinámico con árboles de decisión*

### V.9.2.1. Ingreso del paciente

En la fase de ingreso del paciente se cuentan con los atributos mostrados en la *Figura 19*, los cuales son: edad servicio, sexo, inmunocompromiso y DE, después de aplicar la metodología de reducción de atributos sobre el conjunto balanceado se elimino el atributo sexo.

Se hizo la misma comparación realizada en la fase de ingreso de paciente con *Naive Bayes* y produjo los mismo resultado entonces el modelo resultante se puede ver en la *Figura 31*.



**Figura 31. Modelo de árbol de decisión para el ingreso del paciente.**

#### V.9.2.1.1. Ejemplo de uso de árbol de decisión

En la *Figura 31* se observa el árbol de decisión el cual es un modelo de tipo simbólico porque tiene la capacidad de representar el conocimiento de forma explícita, como lo son las reglas en su forma *si-entonces*.

Supongamos que se tiene un paciente  $P$  y el paciente presenta inmunocompromiso, y se encuentra en el servicio de medicina interna, tiene 45 años de edad y tiene 10 días de estancia hospitalaria. Entonces con esos datos se puede crear la instancia  $P$  con los siguientes atributos:

$$P = \{\text{inmunocompromiso}, \text{servicio}, \text{edad}, \text{DE}\}$$

Sustituyendo con los datos del paciente

$$P = \{\text{si}, \text{medicina interna}, 45, 10\}$$

Después de aplicar las operaciones de transformación sobre los atributos de edad y DE, queda de la siguiente forma

$$P = \{si, medicina\ interna, adulto, estancia\ alta\}$$

Con la instancia  $P$ , el árbol de decisión utiliza la siguiente regla, la cual se obtiene de seguir las flechas de color rojo sobre el árbol de decisión de la *Figura 31*

**Si** servicio = medicina interna y DE = Alta y Edad = Adulto **Entonces**

Riesgo = alto

**Fin si**

Por otro lado si el paciente hubiera estado en el servicio de cirugía y trauma, el árbol de decisión utilizaría la siguiente regla:

**Si** servicio = cirugía y trauma **entonces**

Riesgo = bajo

**Fin si**

Esta regla se obtiene a partir del árbol de decisión de la *Figura 31* siguiendo la flecha de color verde.

A continuación se presenta el modelo de revisión del paciente.

#### **V.9.2.2. Revisión del paciente.**

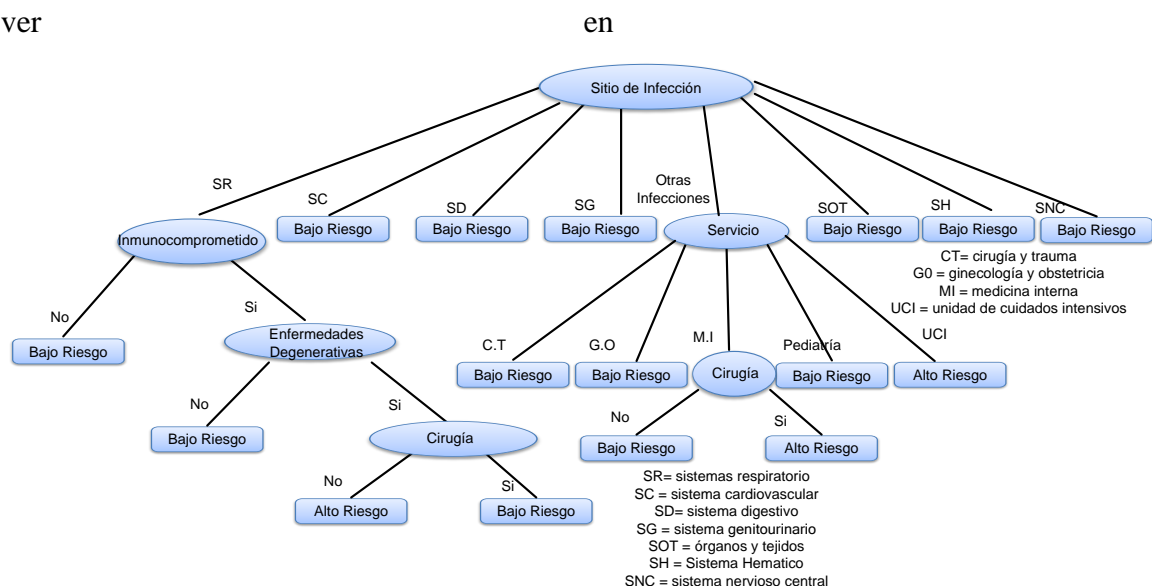
Para obtener un modelo definitivo se utilizaron los conjuntos V, VIII y IX, en los cuales se aplicó la metodología para la reducción de atributos. Los atributos que se tienen disponibles son los siguientes: servicio, inmunocompromiso, Edad, DE, sexo, cirugía, sitio de infección, FR\_obesidad, FR\_Tratamiento, FR\_otros, enfermedades degenerativas, Neumonía.

Después de aplicar la reducción de atributos en el conjunto VIII quedaron los siguientes: servicio, inmunocompromiso, sitio de infección, cirugía y enfermedades degenerativas. En los conjuntos VIII y IX quedaron los atributos anteriores mas la edad y DE. Posteriormente se revisaron el porcentaje de instancias clasificadas correctamente, el conjunto que tuvo el mejor desempeño fue el conjunto V, esto se puede ver en la *Tabla XXXIII*.

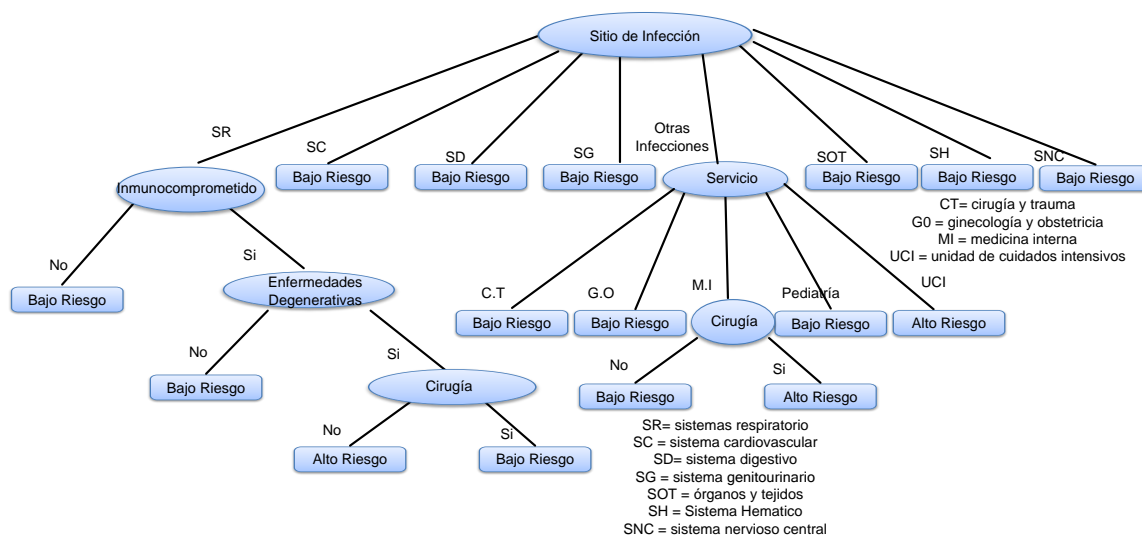
**Tabla XXXIII. Comparación de modelo en revisión del paciente.**

Porcentaje de instancias clasificadas correctamente	
Conjunto V	89.39%
Conjunto VIII	88.25%
Conjunto IX	85.22%

Por esta razón el modelo definitivo se obtuvo a partir del conjunto V. el cual se puede ver



*Figura 32.* En este modelo tiene al atributo cirugía el cual representa los antecedentes quirúrgicos del paciente, pero los antecedentes quirúrgicos del paciente tiene más elementos de información como los son el tipo de cirugía y el grado de contaminación de la misma, sin embargo al agregar estos atributos al conjunto de datos, generaba modelos de poca calidad, por esta razón se decidió crear un conjunto de datos con pacientes que tenían un “si” en el atributo cirugía.



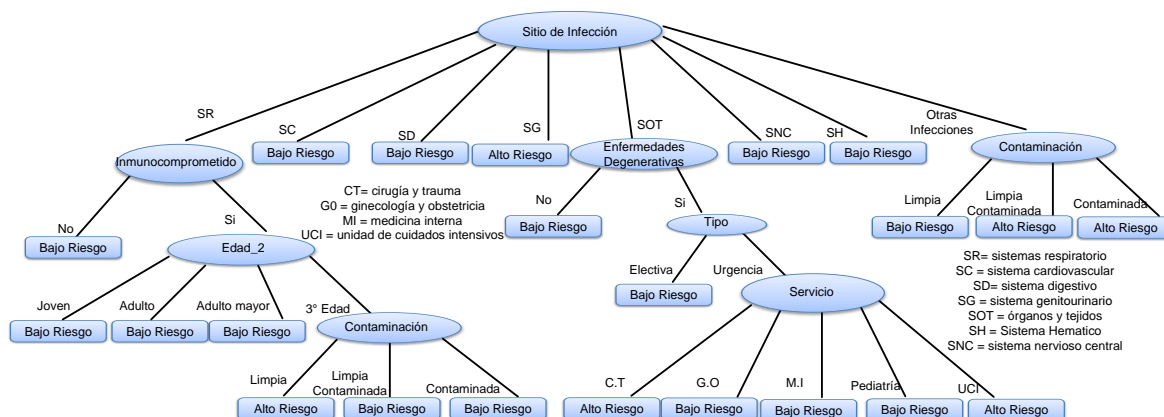
**Figura 32. Modelo de árbol de decisión para la revisión del paciente.**

Para obtener este modelo predictivo definitivo se analizaron los conjuntos de datos VII y XI, entonces después de aplicar la reducción de los atributos sobre el conjunto VII quedaron los siguientes atributos: servicio, inmunocompromiso, sitio de infección, contaminación, tipo y enfermedades degenerativas y para el conjunto X quedaron las anteriores más el atributo edad. Posteriormente se compararon el porcentaje de instancias clasificadas correctamente esto se puede ver en la *Tabla XXXIV*.

**Tabla XXXIV. Comparación de los modelos revisión del paciente con solo cirugía**

Porcentaje de instancias clasificadas correctamente	
Conjunto VII	88.45%
Conjunto X	90.83%

En la *Tabla XXXIV* se puede observar que el que tuvo un porcentaje de instancias clasificadas correctamente fue la del conjunto X, el modelo resultante de esto se puede ver en la *Figura 33*.



**Figura 33. Modelo de árbol de decisión para la revisión del paciente con cirugía**

A continuación se describirá el modelo de detección de IN.

### V.9.2.3. Detección de IN.

El objetivo de los anteriores modelos predictivos es detectar oportunamente una IN, sin embargo en esta fase se desea conocer el riesgo que tiene un paciente después de adquirir una IN.

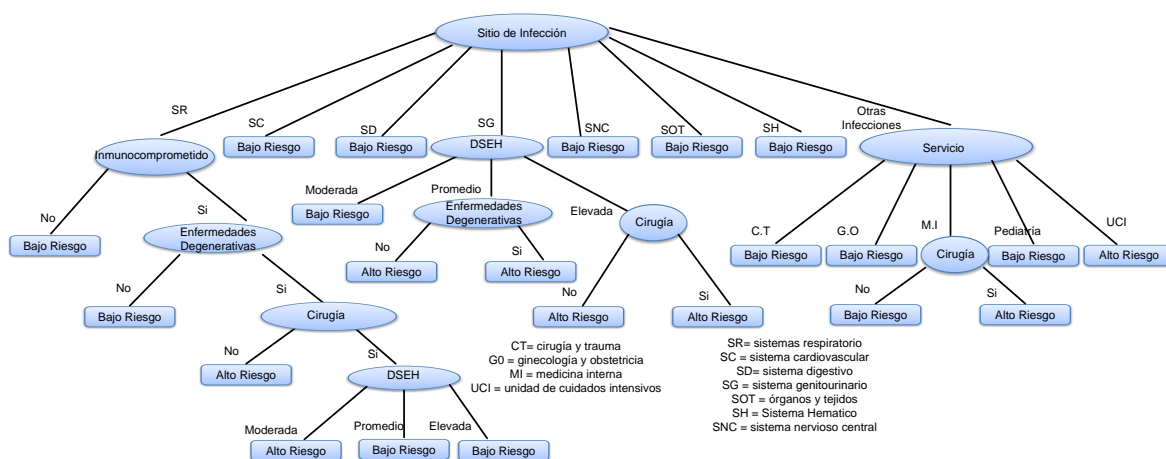
Para obtener un modelo definitivo se utilizaron los conjuntos V y IX, en los cuales se aplicó la metodología para la reducción de atributos. Los atributos que se tienen disponibles son los siguientes: servicio, inmunocompromiso, Edad, DE, sexo, cirugía, sitio de infección, FR\_obesidad, FR\_Tratamiento, FR\_otros, enfermedades degenerativas, Neumonía, microorganismo y DSEH

Después de aplicar la reducción de atributos sobre ambos conjuntos quedaron los siguientes: servicio, inmunocompromiso, sitio de infección, cirugía, enfermedades degenerativas y DSEH. Posteriormente se comparó el porcentaje de instancias clasificadas correctamente, el conjunto que tuvo el mejor desempeño fue el conjunto IX, esto se puede ver en la *Tabla XXXV*.

**Tabla XXXV. Comparación de los modelos en detección de IN**

Porcentaje de instancias clasificadas correctamente	
Conjunto V	88.45%
Conjunto IX	89.39%

En la *Figura 34* se muestra el modelo predictivo resultante y se obtuvo a partir del conjunto IX. A continuación se presenta el modelo basado en casos de solo cirugías.



**Figura 34. Modelo de árbol de decisión para la fase de detección de IN.**

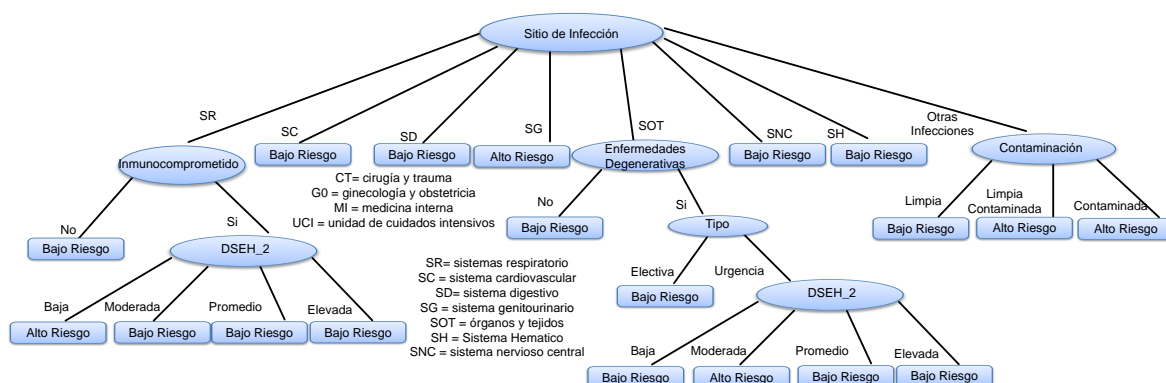
Para obtener este modelo predictivo se analizaron los conjuntos de datos VII y XI, entonces después de aplicar la reducción de los atributos sobre el conjunto VII y XI quedaron los siguientes atributos: servicio, inmunocompromiso, sitio de infección, contaminación, tipo, DSEH y enfermedades degenerativas. Posteriormente se compararon el porcentaje de instancias clasificadas correctamente esto se puede ver en la *Tabla XXXVI*.

**Tabla XXXVI. Comparación de los modelos en detección IN con solo cirugía**

Porcentaje de instancias clasificadas correctamente	
Conjunto VII	89.65%
Conjunto X	90.24%



En la *Tabla XXXVI* se puede observar que el que tuvo un porcentaje de instancias clasificadas correctamente fue la del conjunto X, el modelo resultante se puede ver en la *Figura 35*.



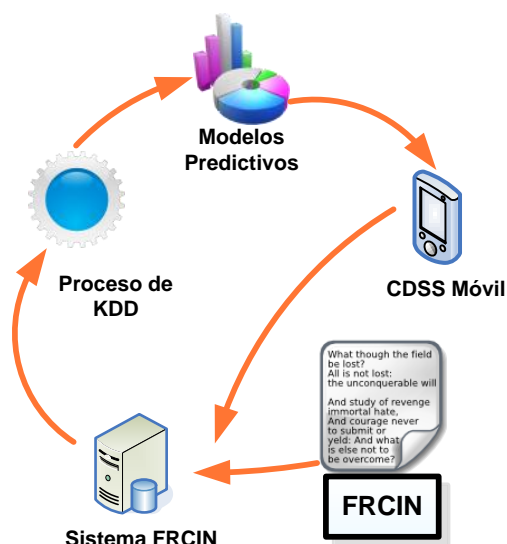
**Figura 35.** Modelo de árbol de decisión para la fase de detección de IN de solo cirugía.

El siguiente paso dentro del proceso de extracción de conocimiento es la evaluación del conocimiento adquirido, el cual se muestra en el capítulo VII.

Anteriormente se ha mencionado que uno de los elementos del CDSS es el conocimiento el cual puede ser adquirido mediante un proceso de extracción de conocimiento y la utilización de técnicas de minería de datos. Entonces el último paso del proceso de extracción de conocimiento, es la consolidación del conocimiento adquirido representado por los modelos predictivos. Es decir en la implementación de los modelos predictivos en el CDSS móvil y en el sistema FRCIN.

En el próximo capítulo se muestra la consolidación del conocimiento en los sistemas CDSS móvil y en el sistema FRCIN. Estos sistemas tienen el objetivo de realizar una valoración de los pacientes para la detección oportuna de una infección nosocomial.

La *Figura 36* muestra el ciclo de extracción de conocimiento donde a partir del FRCIN se creó una base de conocimiento y se aplico el proceso de extracción de conocimiento para obtener modelos predictivos los cuales serán embebidos en el CDSS móvil y en el sistema FRCIN.



*Figura 36. Ciclo de extracción de conocimiento.*

## V.10. Resumen

Teniendo identificados la base de conocimiento y el caso de estudio se procedió con el inicio al proceso de extracción de conocimiento. El primer paso fue la definición del objetivo el cual es el de obtener modelos predictivos para la detección oportuna de infecciones nosocomiales. Con esto en claro, se procedió a realizar una selección inicial de atributos a partir de la FRCIN.

En base a esos atributos se identifico el proceso de generación información en base a eventos los cuales son: ingreso del paciente, revisión del paciente, ingreso del análisis de laboratorio y detección de IN. Cada fase de generación de información tiene su respectivo conjunto de atributos.

Después se desarrollo una aplicación para la creación de una vista minable, posteriormente se realizo un análisis de exploración de los diversos métodos de minería de datos para el tratamiento de información médica en base a este análisis se eligieron los siguientes métodos: *Naive Bayes* y arboles de decisión (C4.5).

En base a esos métodos elegidos se inicio un proceso de trasformación y generalización de datos sobre la vista minable. Con la vista minable trasformada se inicio el proceso de reducción de la dimensionalidad para posteriormente aplicar los métodos minería de datos.

Con este último paso se obtuvieron los modelos predictivos los cuales se almacenaron en un depósito de modelos, posteriormente estos modelos se compararon y en base al porcentaje de instancias clasificadas correctamente, precisión y alcance de las clases se eligieron los mejores modelos predictivos los cuales corresponden a las fases de generación de información de eventos.

En el siguiente capítulo se muestra la consolidación de estos modelos para su uso en un sistema móvil y otra versión de escritorio con el objetivo de ayudar en la toma de decisión en la detección de infecciones nosocomiales.

**Diseño e implementación de CDSS**

---

**VI.1. Introducción**

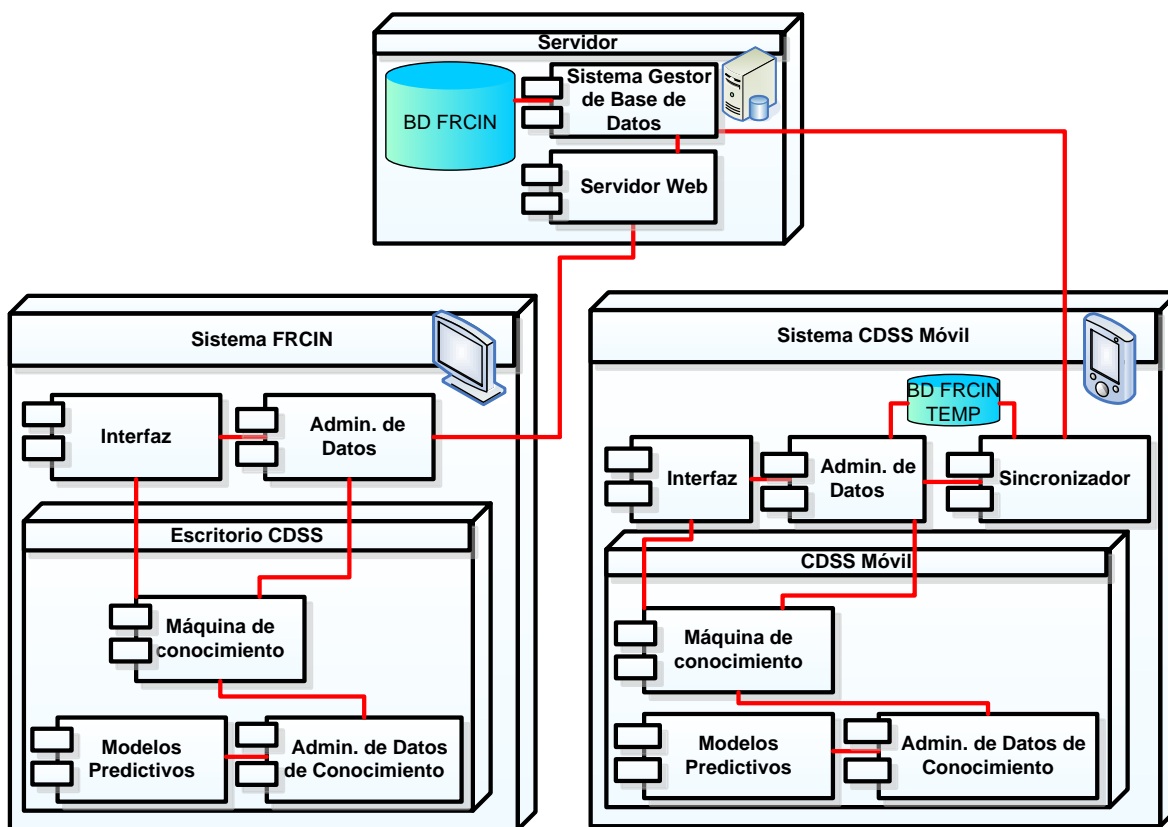
Este capítulo describe el diseño del CDSS que se propone en este trabajo de tesis, el cual fue realizado a partir de los requerimientos obtenidos, que se analizaron en el capítulo IV, y en el diseño de los modelos predictivos, que especificamos en el capítulo V. Para representar y especificar este diseño emplearemos la notación del lenguaje unificado de modelado (UML, por sus siglas en inglés *Unified Modeling Language*) (Fowler, 2004). Este diseño de CDSS, para dispositivos heterogéneos, está compuesto por tres elementos: una aplicación móvil, y una base (fija), además de un servidor.

Primero presentaremos el diseño estructural del sistema mediante un diagrama arquitectónico. Seguido por el diseño de la funcionalidad del sistema que reflejan los requisitos obtenidos, mediante diagramas de casos de uso. Posteriormente explicaremos el comportamiento principal del sistema mediante diagramas de secuencia.

Para finalizar el capítulo se describe el desarrollo de un prototipo basado en los diseños anteriores de CDSS, en el cual se describe la implementación del diseño arquitectónico, y posteriormente se describen las funcionalidad y características del prototipo desarrollado en base a los requerimientos obtenidos en el caso de estudio, para finalmente describir la funcionalidad este prototipo con un escenario de uso.

## VI.2. Diseño arquitectónico

En esta sección describimos el diseño arquitectónico. Jacobson et al. (1997), definen el diseño arquitectónico como la organización estática del software en subsistemas conectados a través de interfaces, y como estos subsistemas interactúan entre sí. En esta sección estableceremos un diseño arquitectónico (Ver *Figura 37*), compuesto por un servidor, un sistema uno móvil y otra base (fijo).



*Figura 37. Diseño arquitectónico del sistema.*

Se diseñó un sistema cliente-servidor, para que de esa forma los usuarios pudieran compartir la información que maneja el sistema. Además, se tiene el requerimiento de emplear tanto dispositivos móviles, como fijos. Entonces, resulta necesaria una arquitectura donde dispositivos heterogéneos tengan interoperabilidad con una misma fuente de información.

Cada uno de los componentes de esta arquitectura se diseño en base a los elementos de básicos de los DSS. Enseguida se describirá cada uno de estos componentes y sus elementos:

- **Servidor.** Tiene la función de proporcionar servicios para la gestión y administración de la información. El cual se compone de los siguientes elementos:
  - Servidor Web. proporciona servicio al sistema FRCIN y brinda una interface para la gestión de la información.
  - Sistema gestor de base de datos. se encarga de proporcionar y almacenar la información contenida en la base de datos.
- **Sistema Móvil CDSS.** Este componente le acerca al usuario, mediante una interfaz montada en un dispositivo móvil, las funcionalidades del sistema. El cual está compuesto por los siguientes elementos:
  - Interfaz. Proporcionaría el dialogo con el tomador de decisión y mediante la interacción con la máquina de conocimiento, o directamente con el administrador de datos, accede a las funcionalidades del sistema.
  - Administrador de datos. Maneja el contenedor la base de datos temporal, para dar soporte a la desconexión. Este elemento interactúa con el componente interfaz, y con la máquina de conocimiento.
  - Sincronizador. Obtiene la información para depositarla en la base la base de datos temporal, en caso de que en la base de datos temporal existan cambios, el sincronizador gestionara estos cambios para que sean actualizados en la base de datos.
  - Máquina de conocimiento. Este elemento interactúa con los elementos administradores de datos y administrador del conocimiento, para proporcionara contexto a los modelos predictivos acoplándolos con los datos necesarios.
  - Administrador de datos de conocimiento. Interactúa directamente con los modelos predictivos y la máquina de conocimiento siendo un mecanismo para la trasformación de datos de acuerdo a la entrada del modelo predictivo.

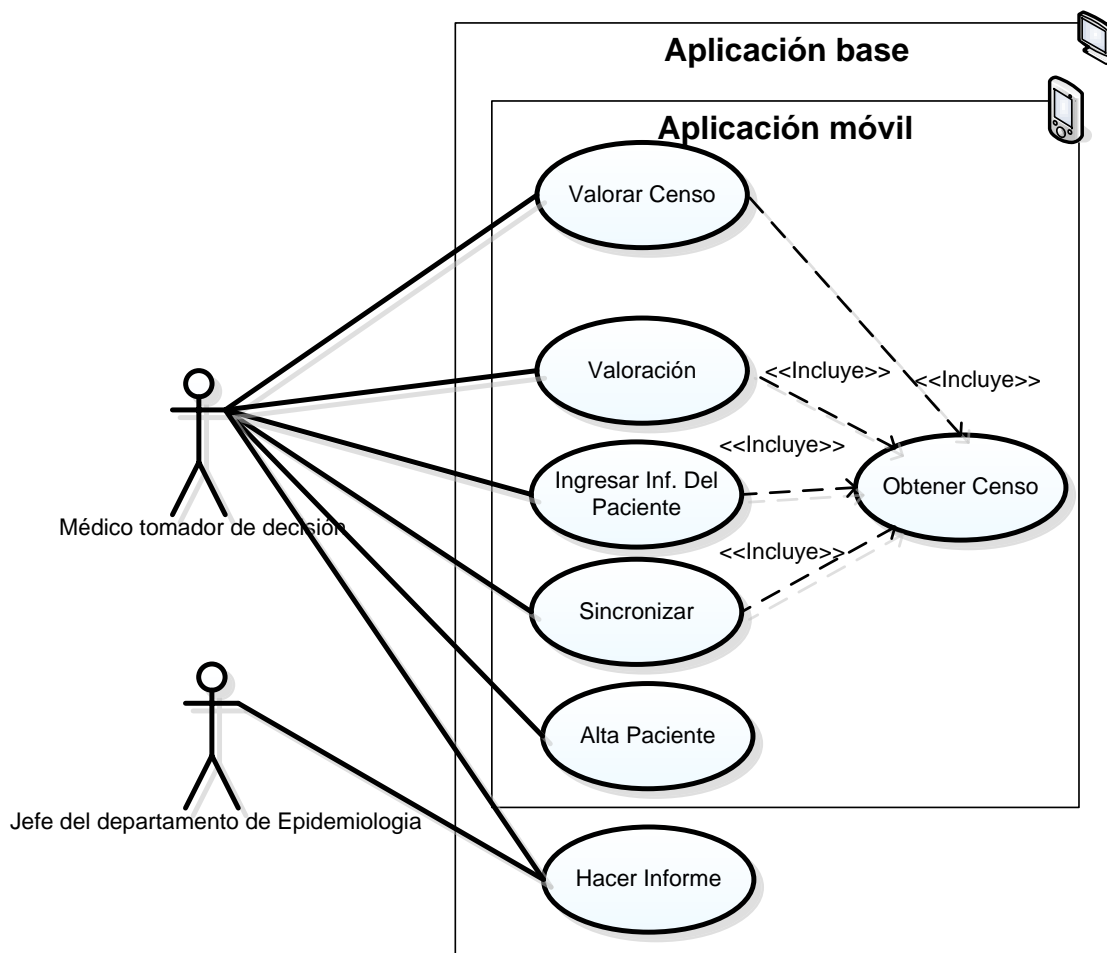
- Modelos predictivos. Interactúa directamente con el administrador de datos de conocimiento, en este modulo se encuentran los modelos predictivos obtenidos por medio del proceso de extracción de conocimiento.
  - BD FRCIN TEMP. Interactúa directamente con el modulo de sincronizador y el modulo de administrador de datos. la base de datos FRICN temporal contiene la información relacionada de los paciente y la almacena temporalmente hasta que se sincronice con el sistema gestor de base de datos mediante el modulo de sonorización.
- **Sistema FRCIN.** Este subsistema comparte las características de diseño del subsistema móvil, pero mediante una interfaz montada en un dispositivo fijo.

Una vez definido el diseño arquitectónico del sistema, se explicara la funcionalidad que proporcionará.

### **VI.3. Funcionalidad (casos de uso)**

En esta sección presentamos los requisitos funcionales del sistema que giran en torno al acceso compartido de información y conocimiento, y el uso de herramientas con modelos predictivos para la toma de decisiones. Para ello en la notación de UML se identifica y presenta los casos de uso, la cual es una representación simple en términos de actores (que interactúan con el sistema) y casos de uso (que documentan la acción de del actor que lo ejecuta) (Bennett, 2004).

En base los requerimientos funcionales y arquitectónicos diseñamos un sistema móvil, y un sistema base (FRCIN). Para simplificar el diseño modelaremos con diagramas de casos de uso, primero las funcionalidades del sistema móvil y base. Seguido se describirá brevemente las funcionalidades principales (ver *Figura 38*).



*Figura 38. Diagrama de casos de uso para el sistema móvil y de base*

La aplicación móvil les brindara toda la funcionalidad para que el tomador de decisión pueda realizar sus actividades y pueda valorar al paciente. De forma breve se explicara cada caso de uso y posteriormente se detallara el caso de uso de valoración.

El médico tomador de decisión **obtiene el censo** de pacientes que se encuentran hospitalizados, posteriormente el sistema CDSS móvil resalta los pacientes con alto riesgo de contraer una infección nosocomial, el médico podrá **ingresar información del paciente**, en base a esta nueva información el médico podrá realizar una nueva **valoración**. En caso de que sea necesario el médico podrá dar de **alta un paciente** en el censo, cuando el médico termine de hacer sus actividades podrá **sincronizar** el dispositivo móvil con el



sistema base, de esta forma podrá actualizar todos los cambios realizados en el sistema móvil a la base de de datos.

Para la construcción de este CDSS se detallaron todos los elementos de análisis y diseño de este. Sin embargo, para fines de este documento de tesis, se describen brevemente solo los casos de uso principales (véase Tabla XXXVII):

**Tabla XXXVII. Casos de uso principales**

<i>Caso de uso</i>	<i>Valoración</i>
Actores:	Médico tomador de decisión
Propósito:	Realizar una valoración sobre el paciente utilizando diferentes modelos predictivos
Descripción:	Para realizar una valoración sobre un paciente, el <i>médico tomador de decisión</i> tendrá que elegir a uno de ellos en el censo de pacientes posteriormente eligiera cual modelo predictivo utilizar ya sea el <i>Naive Bayes (Figura 45)</i> o el <i>C4.5 (Figura 46)</i> . Cada uno de estos tiene diversos elementos de información los cuales están divididos en fases. El <i>Naive Bayes</i> consta de las siguientes tres fases: ingreso del paciente, revisión del paciente, ingreso de análisis de laboratorio, cada uno de estas fases requiere diferentes elementos de información, en el ingreso de información, el sistema desplegara la siguientes atributos: edad, servicio, días estancia, si el paciente es inmunocomprometido, en la fase de revisión del paciente el sistema móvil desplegara los anteriores atributos más el sitio de infección y los factores de riesgo del paciente, en la última fase el sistema presentara lo anterior mas si el paciente tiene análisis de laboratorio. En las tres fases anteriores el sistema tiene un opción de valoración, cuando el <i>médico tomador de decisión</i> utilice el sistema desplegara la probabilidad de contraer una infección nosocomial. Para la opción del <i>C4.5</i> el <i>médico tomador de</i>

---

*decisión* podrá elegir entre las siguientes fases: ingreso del paciente, revisión del paciente y detección de IN, en la fase de ingreso del paciente tiene los mismos atributos que en su fase homologa en clasificador bayesiano, en la fase de revisión del paciente, el sistema presentara si el paciente tiene antecedentes quirúrgicos, sitios de infección y factores de riesgo, en la última fase el sistema desplegara en caso de tenerlos los días de sobre estancia hospitalaria. En las tres fases de este modelo se tendrán la opción de valorar con *C4.5* desplegando el resultado el cual puede ser *Bajo riesgo* ó *Alto riesgo* según sea el caso, así mismo los tres tendrá la opción de razonamiento la cual desplegara la regla y árbol de decisión utilizado para llegar a esa conclusión.

<i>Caso de uso</i>	<b><i>Valorar censo</i></b>
Actores:	Médico tomador de decisión
Propósito:	Consultar información para la toma de decisiones sobre que pacientes que se encuentran hospitalizados
Descripción:	Cuando el <i>médico tomador de decisión</i> , <b>obtiene el censo</b> de pacientes para conseguir la información relacionada con el paciente como es el No. De afiliación, nombre, apellido paterno, servicio, No. de cama, edad, días de estancia y días de sobre-estancia hospitalaria, si el paciente es inmunocomprometido. Antes de desplegar estos datos del paciente el sistema realiza una valoración, en caso de que el paciente presente riesgo de contraer una infección nosocomial, el sistema resaltara ese paciente de tal forma que el <i>médico tomador de decisión</i> le preste atención ( <i>Figura 43a</i> ).
Comentario:	n/a
Prerrequisito:	n/a

---

En el caso de uso de ingreso de información del paciente corresponde cuando el médico ingreso y consulta de información sobre el paciente como el sitios de infección (*Figura*

43c), diagnóstico (*Figura 43d*), factores de riesgo (*Figura 44a*), antecedentes quirúrgicos (*Figura 44d*), análisis de laboratorio (*Figura 44c*), desenlace de pacientes (*Figura 44d*) y las observaciones del paciente.

El caso de uso sincronizar se encarga de obtener toda la información de la base de datos del sistema FRCIN sobre los pacientes que se encuentran en el hospital, esta opción se puede ver en la *Figura 43a*, asimismo se inicia esta opción cuando se desea almacenar la información capturada del sistema móvil al sistema FRCIN.

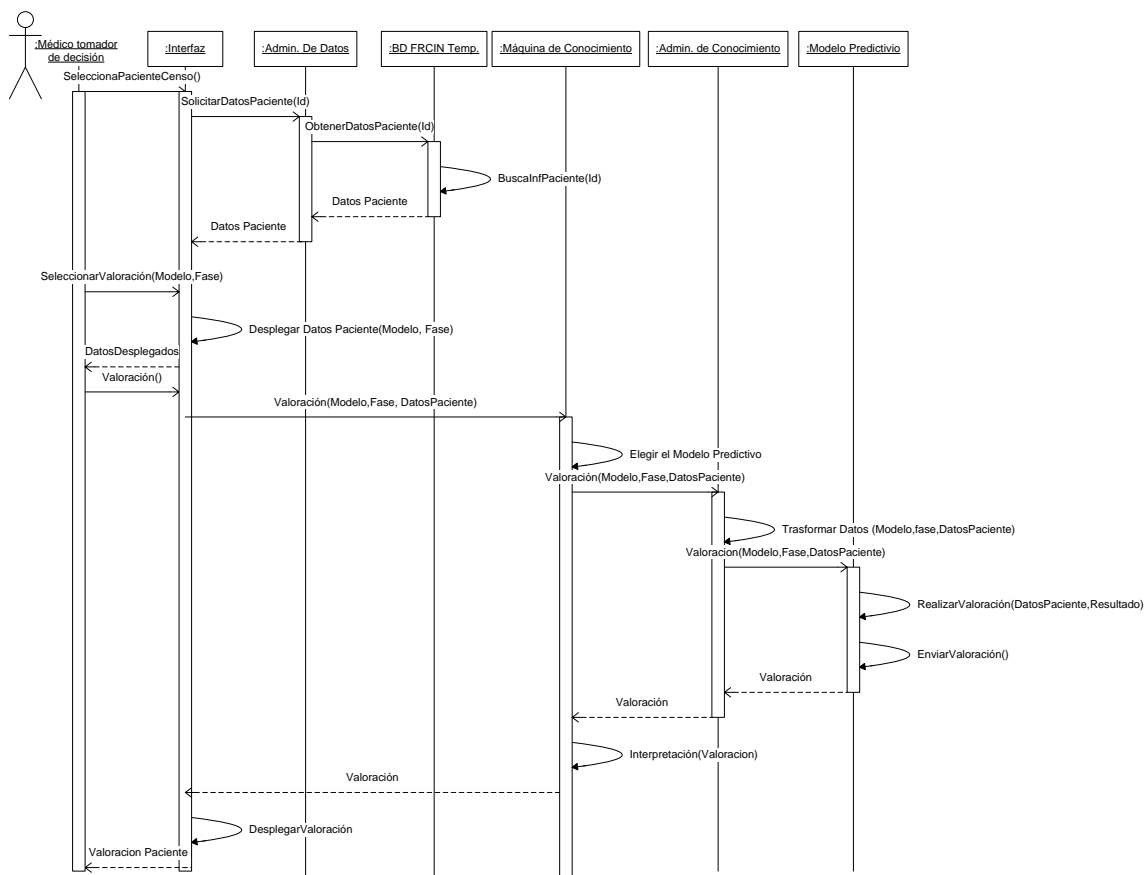
El caso de uso de alta paciente inicia cuando el personal de la UVEH desea dar de alta a un paciente que no se encuentra en el censo de pacientes. El caso de uso hacer informe inicia cuando el jefe de departamento de epidemiología desea hacer el informe mensual de infecciones nosocomiales.

Una vez establecida la funcionalidad del sistema se mostrará el comportamiento del sistema mediante los diagramas de secuencia.

#### **VI.4. Comportamiento (diagrama de secuencia)**

Describiremos el comportamiento por medio de diagramas de secuencia los cuales muestran las interacciones de los objetos relacionados en una secuencia de tiempo. Enseguida se muestran los diagramas de valoración y valoración censo.

En la *Figura 39* se analizará el caso de valoración, donde el paciente hace utilizar diferentes modelos predictivos para la valoración de pacientes con el objetivo de calcular los riesgos que el paciente tiene de contraer una infección nosocomial.



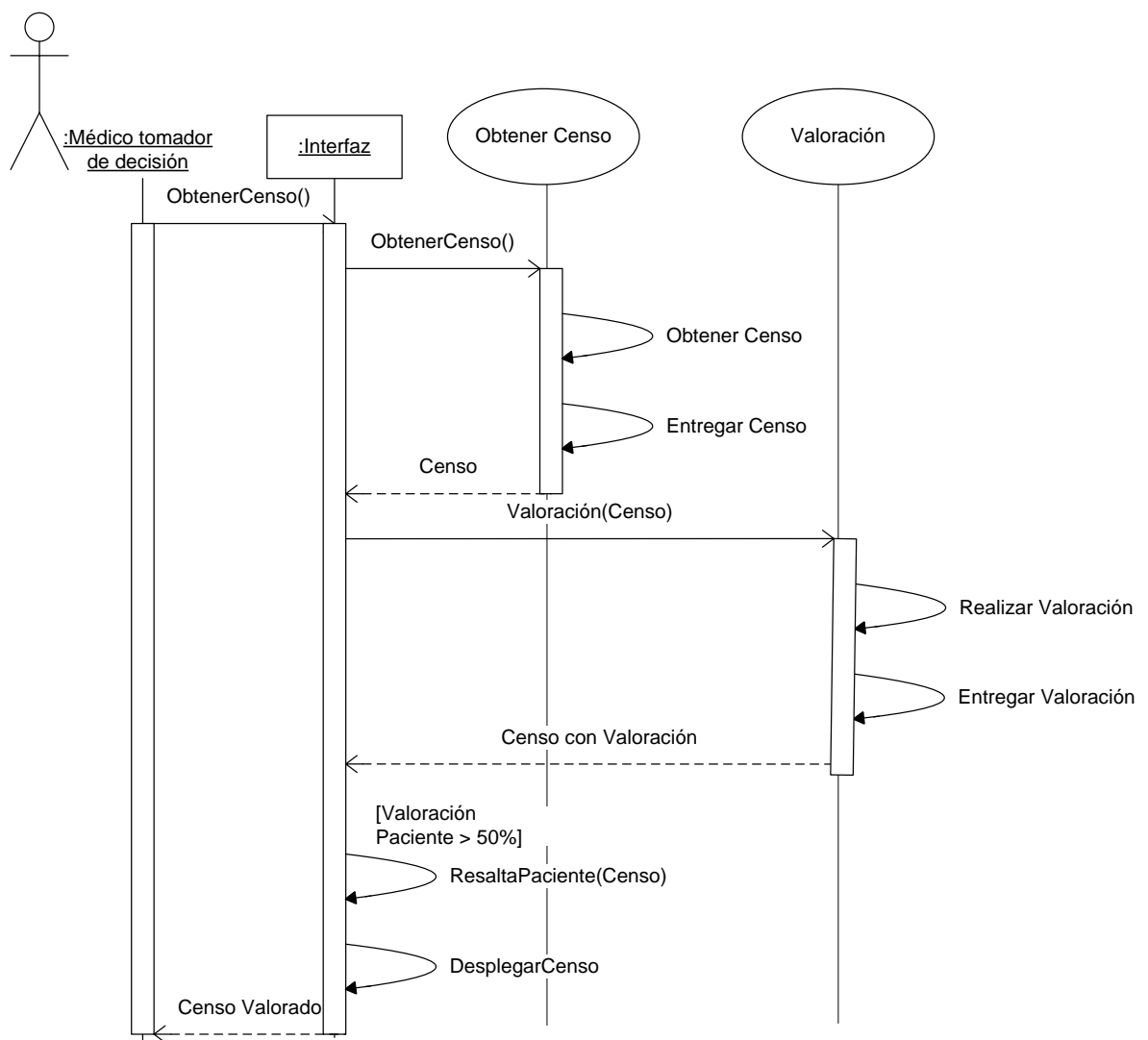
**Figura 39. Diagrama de secuencia del caso de uso “Valoración”**

1. El Actor interactúa con la interfaz, al momento de seleccionar a un paciente en el censo.
2. La interfaz interactúa con el administrador de datos, para solicitar los todos los datos disponibles de ese paciente.
3. El administrador de datos buscar los datos en la base de datos temporal del FRCIN.
4. La base de datos temporal de la FRCIN busca los datos del paciente y se le entrega al administrador de datos.
5. El administrador de datos le entrega los datos del paciente a la interfaz.
6. El actor elige el modelo y la fase en la cual se desea evaluar y se la envía a la interfaz.
7. La interfaz despliega la información correspondiente a ese modelo y la fase que el actor desea evaluar.

8. El actor visualiza la información y posteriormente interactúa con la interfaz para solicitar la valoración del paciente.
9. La interfaz interactúa con la máquina de conocimiento enviándole el modelo, la fase elegida por el usuario y la información del paciente.
10. La máquina de conocimiento elige el modelo de acuerdo a la información del paciente y se la envía al administrador de conocimiento
11. El administrador de conocimiento realiza la transformación de datos de acuerdo al modelo, fase e información proporcionados por la máquina de conocimiento, y se los envía al modelo predictivo.
12. El modelo predictivo realiza la valoración y posteriormente envía la valoración al administrador de conocimiento.
13. El administrador de conocimiento envía el resultado a la máquina de conocimiento.
14. La máquina de conocimiento interpreta la valoración y se lo entrega a la interfaz
15. La interfaz despliega la valoración.

En el paso 6 se pueden crear diversos sub-flujos, debido a que el actor puede elegir diversos modelos predictivos y algunos de ellos tienen diferentes eventos de información, esto se mostrará más a detalle en la implementación

En la *Figura 40* se describe el caso de uso de valorar censo.



**Figura 40. Diagrama de secuencia del caso de uso “Valorar censo”**

1. El Actor interactúa con la interfaz solicitando el censo de pacientes.
2. La Interfaz interactúa con el caso de uso obtener censo solicitando el censo de pacientes.
3. El caso de uso obtener censo entrega el censo a interfaz
4. La interfaz interactúa con el caso de uso valoración enviándole el censo de pacientes para su valoración individual.
5. El caso de uso valoración realiza la valoración y posteriormente la entrega a la interfaz.

6. La interfaz verifica la valoración de la valoración y en caso de que el paciente tenga más del 50% de probabilidad de contraer una infección nosocomial lo desplegara en un color rojo.

Una vez que se conocer los elementos principales de diseño del prototipo, es importante dar a conocer los detalles más importantes de su implementación.

## **VI.5. Implementación**

En esta sección se describe el desarrollo del prototipo realizado basándonos en el diseño de CDSS que describimos en la sección anterior. El cual describe un diseño arquitectónico compuesto por el sistema FRCIN, sistema móvil CDSS y un servidor. A continuación se mostrara la implementación del diseño arquitectónico.

### **VI.5.1. Implementación del diseño arquitectónico propuesto**

Los elementos de software para la implementación de este prototipo está representada en la *Figura 41*, es un esquema que muestran la relación de los elementos del CDSS. Describiremos cada uno de los elementos del esquema arquitectónico.

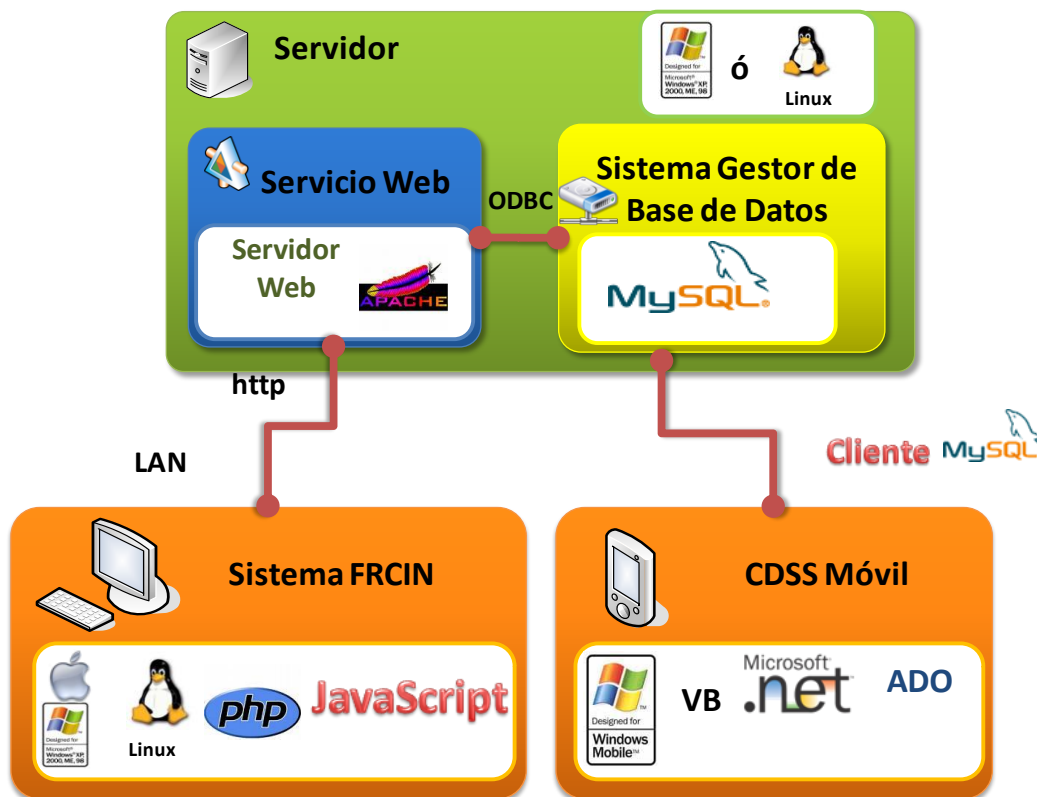


Figura 41. Diseño de la implementación del sistema

- **Servidor.** Este fue implementado a partir de dos elementos un servicio web y una base de datos. Adicionalmente en esta implementación se tomaron en cuenta la consideración de que este subsistema funcionara con un sistema operativo Windows o Linux. Enseguida describiremos los elementos de este subsistema:
  - **Servicio web.** proporciona servicio para el sistema FRCIN, de tal forma que adquiera todas las ventajas del ambiente web, para este modulo se utilizo un servidor *Apache* el cual es un software libre en código abierto
  - **Sistema gestor de base de datos.** Este nos proporciona el registro histórico del sistema. Este fue implementado utilizando *MySQL* como servidor de base de datos. El servicio web se conecta con esta base de datos a través de ODBC, y se utilizan consultas SQL para obtener los datos.
- **Subsistema FRCIN.** Este fue implementado para correr en computadoras de escritorio o portátiles con un sistema operativo *Microsoft Windows XP* o 2000. Este fue



desarrollado sobre *PHP* y se utilizo *JavaScript* para las validaciones e interacciones del usuario con el sistema.

- **Subsistema móvil.** Este fue implementado para correr en una computadora de bolsillo o PDA, tipo PocketPC con un sistema operativo *Microsoft Windows Mobile (PocketPC 2003)*. Fue desarrollado en *Visual Basic* pero utilizando el *framework* punto *NET* de *Microsoft* pero para computadoras de bolsillo. Utiliza un cliente de *MySQL* para comunicarse con la base de datos en caso de no tener acceso a ella utiliza repositorios temporales del *ADO.Net*.

Ahora que hemos descrito como es que implementamos el diseño arquitectónico de este sistema procederemos a describir cómo es que implementamos el diseño de la funcionalidad del sistema.

### **VI.5.2. Implementación del diseño de la funcionalidad propuesto**

Este prototipo de CDSS, que está diseñado para proporcionar una funcionalidad enfocada principalmente al apoyo a la toma de decisiones mediante modelos predictivos obtenidos por medio del proceso de extracción del conocimiento, el cual esta embebido a su proceso organizacional y su flujo de trabajo.

Para este sistema desarrollamos dos tipos diferentes de interfaz una para computadoras de escritorio y otra para computadoras de bolsillo, como se muestra a continuación en la Figura 42.

## Formato de Registro

En Caso de Infección Nosocomial

Nuevo Registro

Informe

Num. de Afiliación:  Nombre/Apellidos:  Descenlace: Todos

Núm. de Afiliación	Nombres	Apellidos	Edad	Servicio	D.E.	D.S.E.H.	Valoración
				Medicina Interna	3	0	59%
				Cirugía y Trauma	7	1	1%
				Cirugía y Trauma	9	1	1%
				Medicina Interna	13860	3	59%

Formato de Registro  
En Caso de Infección Nosocomial

29/09/2008 | Electiva | Limpia | TAI MPD | No es causa (x)

Formato de Registro  
En Caso de Infección Nosocomial

Información Básica Factores de Riesgo Sitios de Infección Antecedentes Quirúrgicos Información Microbiológica Tratamiento Observaciones Valoración

Redes Bayesianas

Ingreso del Paciente  
Prob. de Bajo Riesgo de  41%  
IN

Prob. de Alto Riesgo de  59%  
IN

Revisión Del Paciente  
Prob. de Bajo Riesgo de  43%  
IN

Prob. de Alto Riesgo de  57%  
IN

Ingreso de Análisis de Microbiológico  
Prob. de Bajo Riesgo de  50%  
IN

Prob. de Alto Riesgo de  50%  
IN

C4.5

Guardar y Salir Cancelar

Obtener C

CDSS Movil 12:56

CDSS Movil 8:54

Folio Afi Javier Hernandez  
218 12: Antecedentes Quirúrgicos:  
207 12: Fecha: 10/15/08  
219 12:  
209 12: Tipo:  
210 12: Grado  
213 12: Contaminaci  
220 12: Nombre

Alta Paci  
Censo Pacier

Nuevo AQ. Valorar Redes Bayesianas  
Probabilidad de Bajo Riesgo  
Antecedentes C  30%  
Censo Pacier Probabilidad Alto Riesgo  
 70%

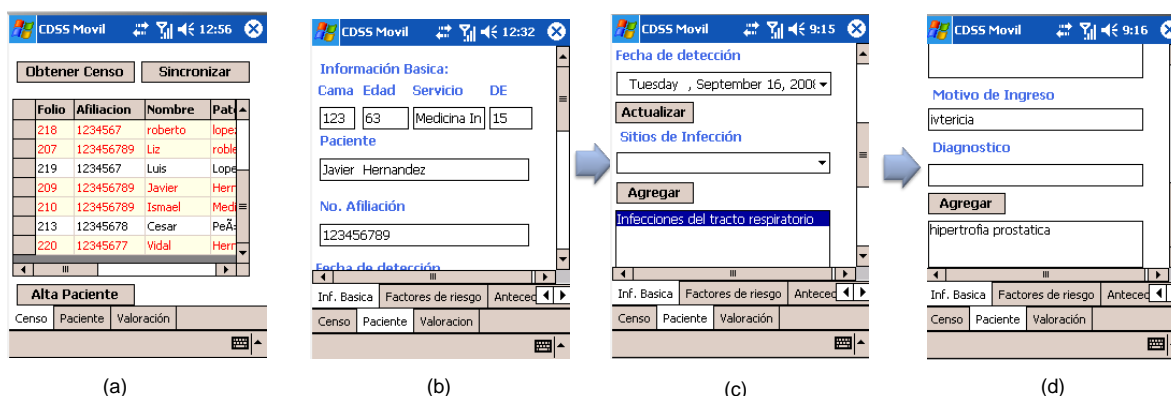
Revisión del paciente Ingreso de Análisis  
Redes Bayesianas C4.5  
Censo Paciente Valoración

(a) (b)

Figura 42. Desarrollo de la interfaz de usuario para (a) computadora de escritorio y (b) dispositivo móvil.

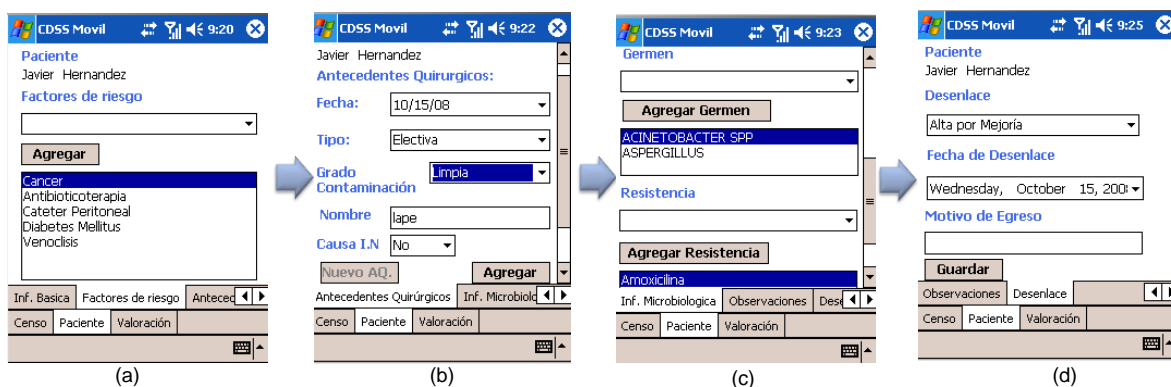
De acuerdo con los requerimientos obtenidos a través del caso de estudio, es importante que el sistema al apoyo a la decisión este embebido dentro de su flujo de trabajo y que se tenga una meta en común que es la detección de infecciones nosocomiales.

Entonces cómo una de las posibles fuentes de detección de infección nosocomial (IN), puede ser a partir del censo, en la propuesta se decidió darle soporte mediante el uso de dispositivos móviles y la utilización de modelos predictivos para la valoración de pacientes.



**Figura 43.** Flujo de información a través del censo, censo valorado (a), información básica del paciente (b), fecha de detección de IN y sitio de infección que tiene el paciente (c), motivo de ingreso y diagnóstico del paciente (d).

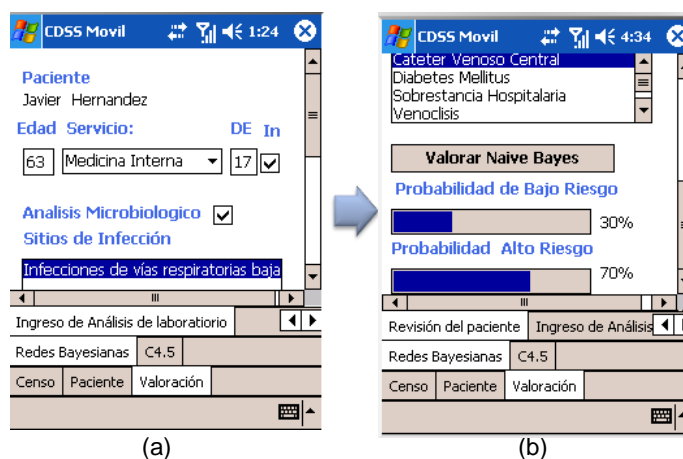
Para realizar esta valoración se necesita tener información sobre el paciente, la cual se va generando conforme a su estadía en el hospital. Entonces cuando el médico tomador de decisión realiza el recorrido a partir del censo de pacientes, esto corresponde al caso de uso ingresar información del paciente, este puede ingresar elementos como son fecha de detección de IN, los sitios de infección que tiene el paciente, motivo de ingreso y diagnóstico, factores de riesgo, antecedentes quirúrgicos y información microbiológica, esto se observa en la *Figura 43* y *Figura 44*.



**Figura 44.** Flujo de información, factores de riesgo del paciente (a), antecedentes quirúrgicos, información microbiológica (b) y el desenlace del paciente (d).

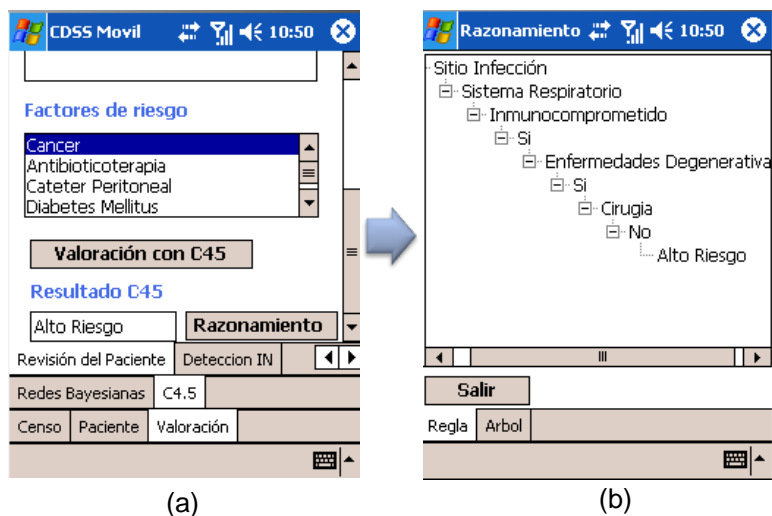
De acuerdo con la información que se tenga disponible se puede entrar a la opción de valoración del paciente mediante los diferentes modelos predictivos implementado en el dispositivo móvil, los cuales se obtuvieron por el proceso de extracción de conocimiento.

En valoración se tienen disponibles dos modelos predictivos, que son el clasificador *Naive Bayes* y el *C4.5*. Para realizar una valoración solo se tiene que seleccionar a un paciente del censo y posteriormente entrar a la opción de valoración, en *Figura 45* se muestra el acceso a los modelos.



**Figura 45.** Opciones de valoración que tiene el médico tomador de decisión, se encuentra en la fase de ingreso de análisis de laboratorio, con el modelo de Naive Bayes(a), valoración con Naive Bayes (b)

En la *Figura 46* se muestran los un ejemplo de los modelos basados en reglas (a), asimismo se presenta la regla que utilizo para llegar a esa conclusiones.



**Figura 46.** Opción de valoración que tiene el médico tomador de decisión con árboles de decisión C4.5 (a), regla utilizada para llegar a ese resultado (b).

A continuación se presentara un escenario de uso del CDSS desarrollado.

### VI.5.3. Escenario de uso del prototipo del CDSS desarrollado

El escenario comienza cuando el personal de la UVEH inicia el recorrido dentro del hospital, el médico o pasante deberá de sincronizar el CDSS móvil con el sistema FRCIN para obtener los datos del Censo previamente capturados en el sistema FRCIN, antes de que el sistema móvil despliegue el censo, este realizaría una valoración con la información disponible y en caso de que detecte un paciente con alto riesgo de contraer un IN este lo marcara en rojo, como se muestra en la *Figura 43a*. Cuando el caso personal de la UVEH entre a la opción de valoración del CDSS móvil, la cual está dividida por diferentes modelos predictivos y por las diferentes etapas en las cuales se generara la información del paciente. En caso de que el CDSS móvil mediante el censo le indique al usuario que el paciente tiene un alto riesgo de contraer una IN, se dirigirá a donde se encuentra el paciente y leerá su expediente médico para obtener información e ingresarla al CDSS (*Figura 43c*, *Figura 43d*, *Figura 44a*, *Figura 44b*, *Figura 44c* y *Figura 44d*) y de esta forma tener más elementos de información los cuales corresponderían a la fase de revisión del paciente, y

así realizar nuevamente una valoración (*Figura 45a, Figura 45c*) con los diferentes modelos predictivos implementados en el CDSS. Cuando el personal de la UVEH solicite análisis de laboratorio lo cual correspondería fase de ingreso de laboratorio, los resultados pueden ser ingresados al CDSS (*Figura 44c*) y usados para una valoración del paciente (*Figura 45a, Figura 45b*). En caso de que el usuario utilice modelos predictivos *C4.5* que es el árbol de decisión (*Figura 46a*) el CDSS puede brindar una explicación mostrando la regla utilizada para llegar a esa decisión (*Figura 46b*). Cuando el personal de la UVEH termina el recorrido del hospital este regresará y sincronizará con el sistema FRCIN para que toda la información capturada con el dispositivo sea actualizada en la base de datos del sistema FRCIN. Con los datos capturados se realiza el última paso del sub-proceso de control el cual es la creación de información mensual de IN.

## **VI.6. Resumen**

En este capítulo se estableció la base del diseño del CDSS, en términos arquitectónicos y funcionales que se han definido en este trabajo de tesis.

Para ello se especifico un diseño arquitectónico compuesto por tres-sub-sistemas: uno de tipo móvil, uno de tipo base y un servidor. La funcionalidad principal de este CDSS es el apoyo a la toma de decisiones en base a modelos predictivos obtenidos a través de un proceso de extracción de conocimiento.

Después se describió la implementación del prototipo tanto como arquitectónico como funcional tanto como los elementos del *Software* en los dispositivos de *Hardware* después se muestra brevemente una descripción de las características de funcionalidad del sistema. Finalmente se describe un escenario de utilización del sistema que muestra la implementación de la funcionalidad del CDSS.

**Evaluación de los modelos predictivos y del CDSS**

---

**VII.1. Introducción**

En este capítulo se presentan la evaluación realizada sobre los modelos predictivos a partir de un conjunto de prueba, así como una evaluación previa del CDSS móvil y el sistema FRCIN con usuarios.

En la evaluación de los modelos predictivos se dividió en dos partes, la primera se obtuvo mediante una validación cruzada, que utiliza un sub-conjunto de datos de entrenamiento para realizar la evaluación (III.3.3.1), la segunda evaluación se obtuvo mediante un conjunto independiente de datos que no se usó para el entrenamiento de los modelos.

En cuanto a la evaluación con los usuarios, el objetivo es conocer su perspectiva respecto a la utilidad y soporte a la decisión del CDSS desarrollado, mediante un conjunto de entrevistas y escenarios de uso para aplicar los modelos predictivos. A continuación se muestra la evaluación de los modelos predictivos.

**VII.2. Evaluación de los modelos predictivos**

En esta sección se describen las pruebas realizadas sobre los diferentes modelos obtenidos por medio del proceso de extracción de conocimiento, mostrados en el capítulo V. Primero se presenta la evaluación obtenida mediante validación cruzada, posteriormente

se presenta la evaluación obtenida por un conjunto de datos que no se uso para la creación de los modelos predictivos.

Las medidas que se tomaron en consideración son: porcentaje de instancias clasificadas correctamente, precisión y alcance de las clases (III.3.3). La precisión, como se había establecido anteriormente, mide la probabilidad de que si un sistema clasifica una tupla en una cierta clase, la tupla realmente pertenezca a esa clase, por otro lado, el alcance mide la probabilidad de que si una tupla pertenece a cierta clase, el sistema lo asigne a la clase.

A continuación se presenta la evaluación mediante validación cruzada.

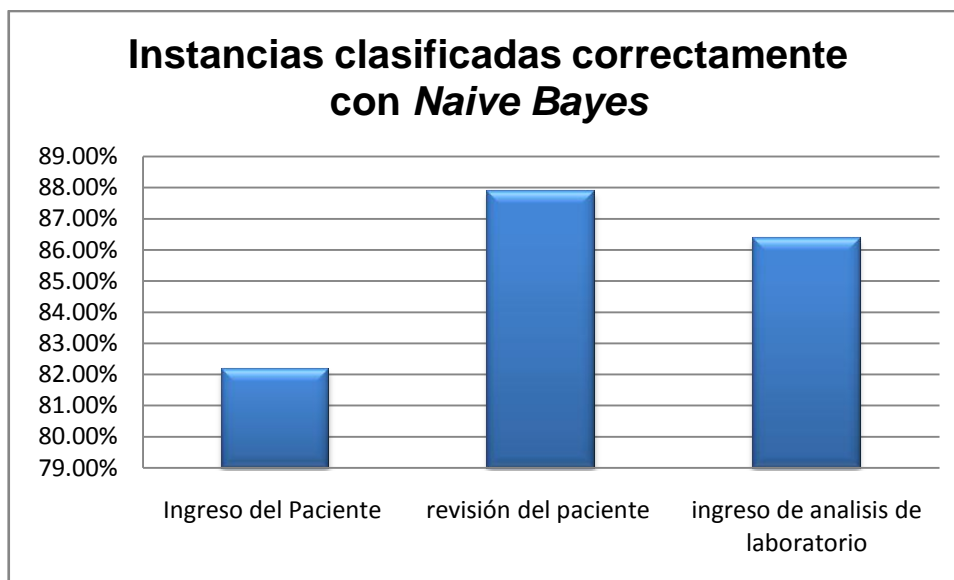
### **VII.2.1. Evaluación de los modelos mediante validación cruzada**

En esta sección se muestra las medidas de desempeño de cada uno de los modelos predictivos obtenidos para cada fase de generación de información, los datos de evaluación se obtuvieron mediante una validación cruzada de 10 particiones. A continuación se muestran los resultados.

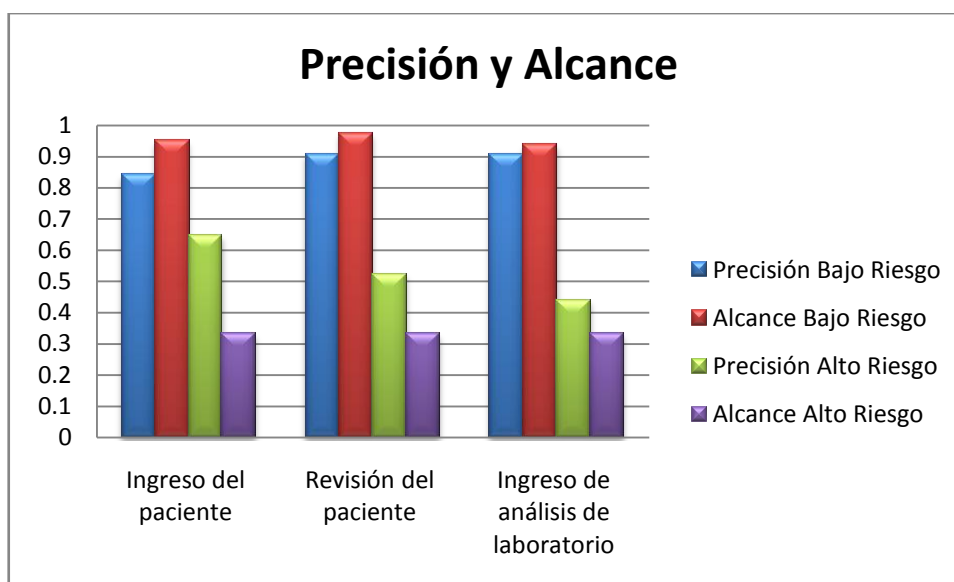
La *sFigura 47* muestra el porcentaje de instancias clasificadas correctamente, en ella se puede ver que la fase de ingreso del paciente tiene un 82% de instancias clasificadas correctamente por otro lado revisión de paciente tiene mayor porcentaje (87.88%) lo que nos da un aumento de 7%. Enseguida se muestran las medidas de precisión y alcance de las clases de alto riesgo y de bajo riesgo.

La *Figura 48* muestra una comparación de las medidas de precisión y alcance en ella se puede observar que la fase de revisión del paciente tiene la mayor alcance en la clase de bajo riesgo, sin embargo la que tuvo mejor precisión en la clase alto riesgo fue la de ingreso del paciente con una diferencia del 44% con respecto a la precisión de la clase alto riesgo en el evento de ingreso del análisis de laboratorio.





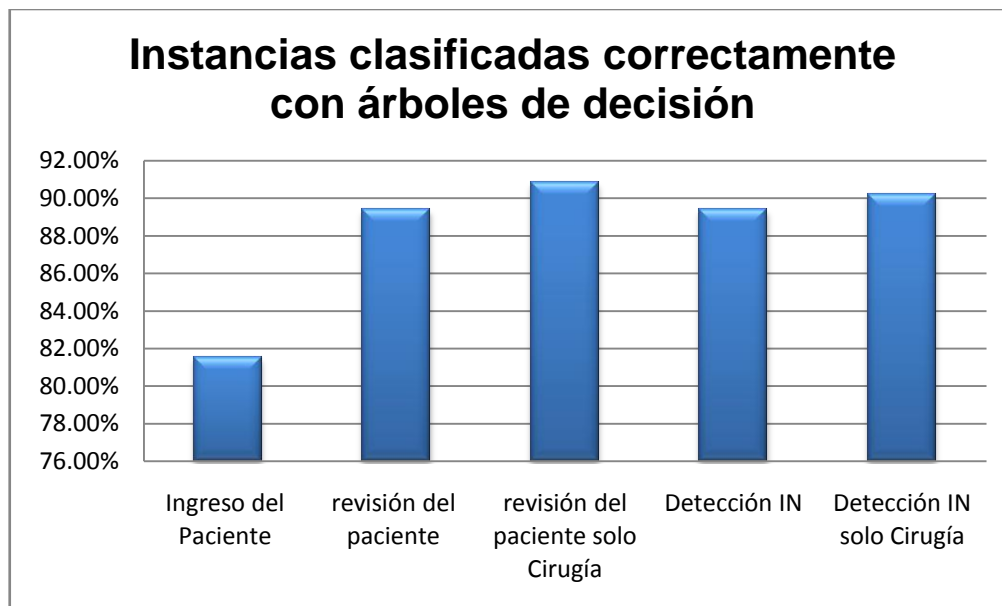
*sFigura 47. Instancias clasificadas correctamente con Naive Bayes.*



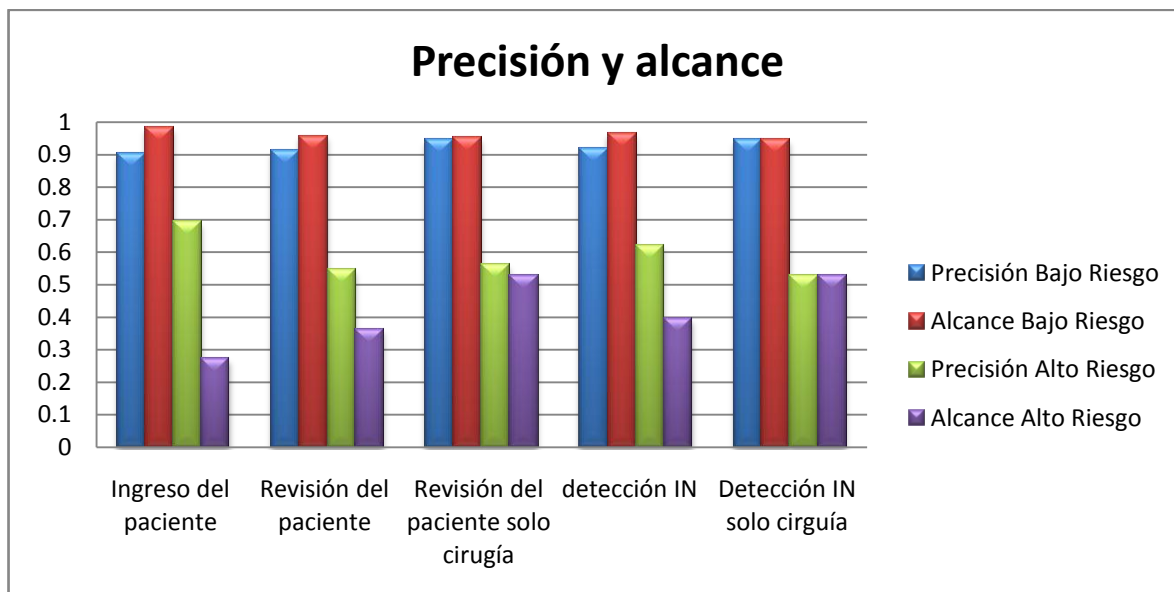
*Figura 48. Comparación de las medidas de precisión y alcance de acuerdo a los eventos de generación de información con Naive Bayes.*

En la *Figura 49* se muestra el porcentaje de instancias clasificadas correctamente con los arboles de decisión en la cual se puede observar la que tuvo el más alto porcentaje fue en la fase de revisión del paciente con solo cirugía mostrando una diferencia del 11% con respecto a la fase de ingreso del paciente.

La *Figura 50* muestra la comparación de los valores de precisión y alcance para las clases de bajo riesgo y bajo alto riesgo, las fases que tienen mejor precisión en la clase alto riesgo es la fase de ingreso del paciente y detección de IN con una diferencia del 31% con respecto a la fase de detección IN de solo cirugías.



*Figura 49. Instancias clasificadas correctamente con árboles de decisión.*



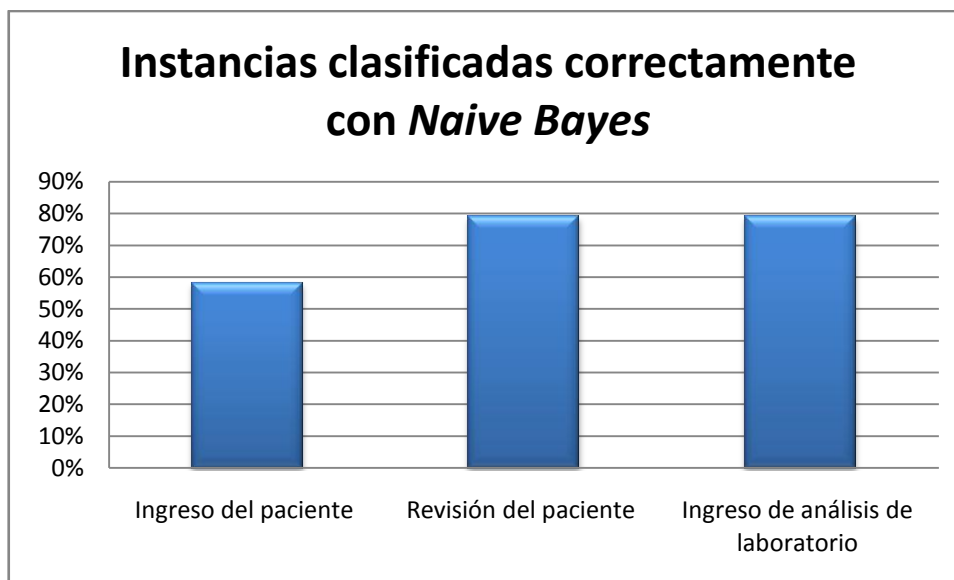
*Figura 50. Comparación de las medidas de precisión y alcance de acuerdo a los eventos de generación de información con árboles de decisión.*

## VII.2.2. Evaluación de los modelos predictivos con el conjunto de prueba.

Para realizar la prueba de estos modelos predictivos obtenidos en el capítulo V se utilizó un conjunto de prueba que consta de 24 casos, la distribución de las clases del atributo desenlace es la siguiente: 18 casos pertenecen a la clase bajo riesgo y 6 pertenecen a la clase alto riesgo.

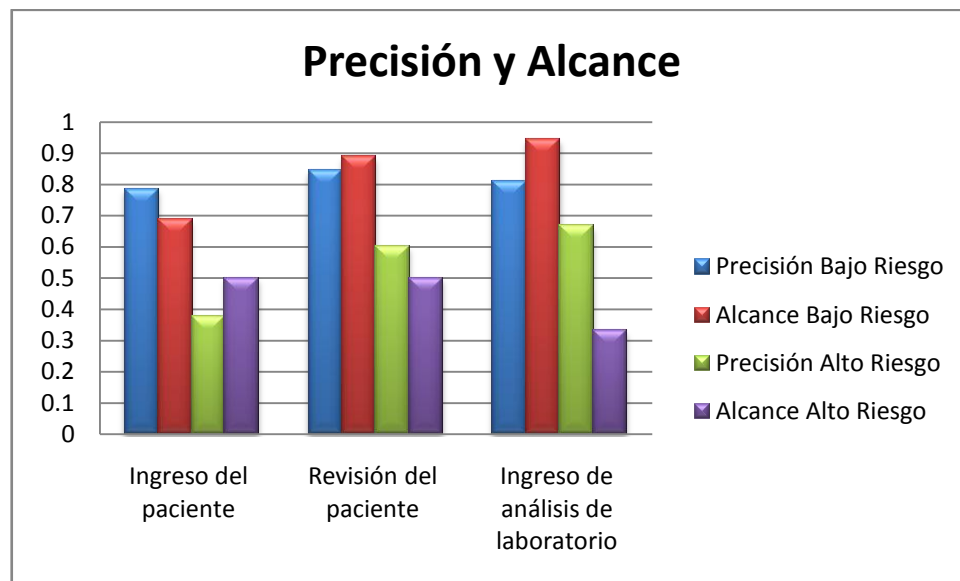
Se utilizó este conjunto para realizar la evaluación de los modelos predictivos se comenzó mostrando la información sobre el clasificador *Naive Bayes* con sus respectivos eventos de generación de información posteriormente se mostraron los resultados de la evaluación con árboles de decisión.

La *Figura 51* muestra el porcentaje de instancias clasificadas correctamente por el método de *Naive Bayes*, se puede observar en el evento de revisión del paciente y ingreso de análisis de laboratorio tienen los más altos porcentajes de instancias clasificadas con un 80%.



*Figura 51. Porcentaje de instancias clasificadas s clasificadas correctamente.*

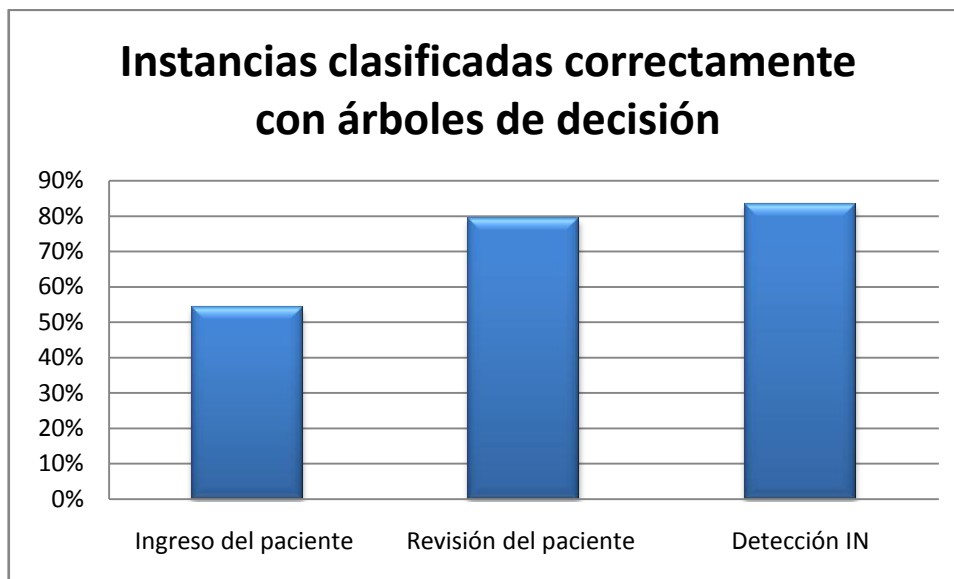
En la *Figura 52* se puede ver las medidas de evaluación como lo son la precisión y el alcance de las clases bajo riesgo y alto riesgo, se puede ver el evento que tiene mayor alcance de la clase bajo riesgo es el evento de ingreso de análisis de laboratorio con una diferencia del 37% comparada el alcance de bajo riesgo en ingreso del paciente. Asimismo el evento que tiene mejor precisión en la clase de alto es el de ingreso de análisis de laboratorio con una diferencia del 78% comparado con el evento de ingreso del paciente.



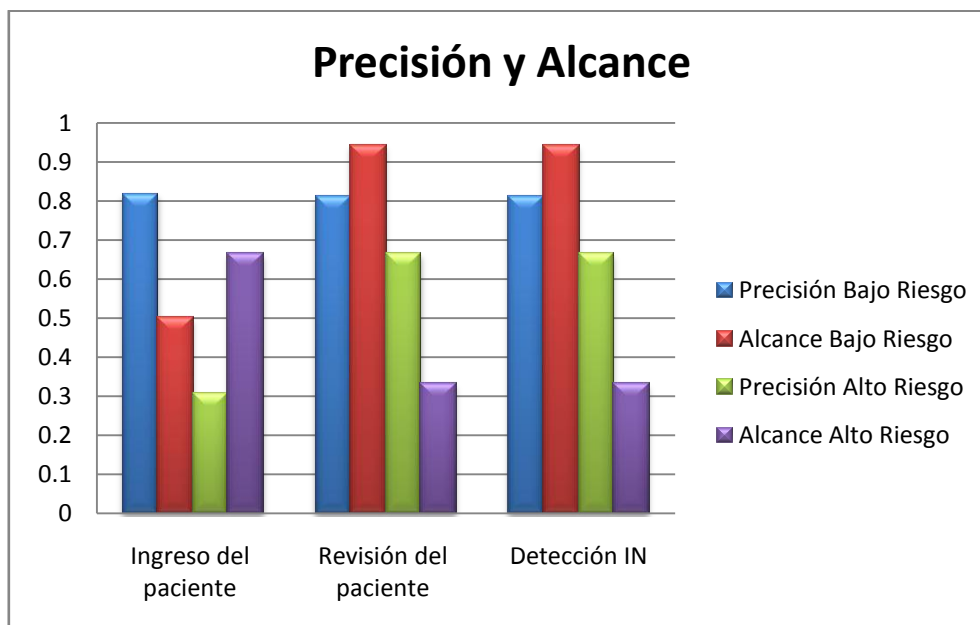
*Figura 52. Comparativa de las medidas de precisión y alcance del clasificador Naive Bayes.*

En la *Figura 53* se muestra el porcentaje de instancias clasificadas correctamente por el método por arboles de decisión, en la misma se puede ver, en los eventos de generación de información de revisión del paciente y ingreso de análisis de laboratorio tienen los más altos porcentajes de instancias clasificadas con un 80% y un 83% respectivamente.

En la *Figura 54* se puede ver las medidas de evaluación como lo son la precisión y el alcance, se puede observar que en los tres eventos tienen la misma precisión en la clase bajo riesgo, por otro lado en los eventos de revisión del paciente y detección de IN tienen el mismo alcance en la clase bajo riesgo con una diferencia del 89% con respecto a la fase de ingreso del paciente. Asimismo en la precisión de la clase de alto riesgo en los eventos de revisión del paciente y detección de IN tienen una diferencia de 98% con respecto a ingreso del paciente.



*Figura 53. Porcentaje de instancias clasificadas s clasificadas correctamente*



*Figura 54. Comparativa de las medidas de precisión y alcance de en arboles de decisión*

Con esto finaliza la evaluación realizada al conjunto de prueba, a continuación se presenta la evaluación realizada al sistema con usuarios.

### **VII.3. Pruebas con usuarios al CDSS**

Esta evaluación tiene como objetivo conocer la opinión de los usuarios acerca de la utilidad que tiene el soporte a decisión del CDSS desarrollado para ayudar en la detección oportuna de infecciones nosocomiales. A continuación se presenta la definición del problema de evaluación.

#### **VII.3.1. Definición del problema de evaluación.**

En este trabajo se ha diseñado y desarrollado un prototipo de CDSS que tiene embebido modelos predictivos obtenidos mediante el proceso de extracción de conocimiento. Entonces evaluamos la percepción de utilidad, facilidad de uso del CDSS, asimismo se evaluó la percepción que tienen los médicos sobre las diferentes técnicas de minería de datos implementadas en el CDSS.

Debido a que el personal de la UVEH son trabajadores de conocimiento especializado en el área de epidemiología, solo se tuvo a disposición de dos médicos el jefe de departamento de epidemiología y la epidemióloga, por lo tanto los resultados de la evaluación son cualitativos. Enseguida describiremos el diseño de la evaluación.

#### **VII.3.2. Diseño de la evaluación**

Las preguntas de investigación para esta evaluación son las que se plantean a continuación,

- ¿Qué percepción de utilidad tendrán los médicos respecto a la utilidad del CDSS que esté basado en modelos predictivo obtenidos mediante un proceso de extracción de conocimiento?
- ¿Qué percepción tendrán los médicos respecto a la facilidad de uso del CDSS móvil y sistema de escritorio?

- ¿Qué percepción tendrán los médicos respecto a los diferentes tipos de modelos predictivos implementados en el CDSS?

A continuación se presentan las actividades que se llevaron a cabo en la realización de la evaluación con la finalidad de determinar los aspectos de utilidad, facilidad de uso, manejo de información y perspectivas de sobre los diferentes modelos implementados en el CDSS.

### **VII.3.3. Actividades realizadas**

Las tareas que se realizaron son las siguientes introducción al proyecto, capacitación sobre la funcionalidad del prototipo, realización de las actividades y recolección de los resultados.

En la introducción del proyecto se presentó y explicó a los médicos la finalidad de la evaluación; además, de los diferentes modelos predictivos implementados en el CDSS y su origen el cual es el FRCIN.

En la capacitación sobre la funcionalidad se les proporcionó una breve capacitación sobre el uso de las aplicaciones (CDSS móvil y el sistema FRCIN). En cuanto al funcionamiento del prototipo se presentó a la UVEH una serie de actividades (apéndice D) para interactuar con el prototipo, para que de esta forma pudieran probar la funcionalidad del mismo.

Para realizar las actividades se proporcionó al jefe de departamento y a la epidemióloga una computadora y una PDA. Las actividades de los médicos estuvieron de acuerdo con sus actividades del PDVIN.

- *“Ingreso de información del paciente en el sistema FRCIN”*
- *“Ingreso de información del paciente en el CDSS móvil”*



- *“Ingreso de factores de riesgo del paciente”*
- *“Valoración de riesgo de contraer un IN con Naive Bayes en los diversos eventos de generación de información”*
- *“Valoración de riesgo de contraer un IN con árboles de decisión en los diversos eventos de generación de información ”*
- *“Censo valorado de pacientes”*
- *“Sincronización con el sistema FRCIN”*
- *“Creación del informe FRCIN”.*

Al final de las actividades se realizó una entrevista para conocer sus percepciones, además se realizaron discusiones respecto a las mismas. Para llevar a cabo esta evaluación, se presentó el sistema a los principales roles involucrados del PDVIN, es decir al jefe de departamento de epidemiología y a la epidemióloga. La evaluación se llevó a cabo de forma separada, en una única sesión de aproximadamente 30 minutos por cada usuario, haciendo uso de una grabadora. A continuación se muestran los resultados de esta evaluación.

#### **VII.3.4. Análisis cualitativo de los resultados encontrados en la evaluación**

Se presentaron los comentarios obtenidos para cada una de las percepciones de esta evaluación, las preguntas que se plantearon se pueden encontrar en el apéndice D. primeramente se presentan la percepción de utilidad del CDSS.

##### **VII.3.4.1. Utilidad percibida**

Para obtener esta percepción de utilidad de los médicos se realizaron diversas actividades con el objetivo de mostrar la funcionalidad de los sistemas tanto en su versión móvil como la de escritorio.

Las funcionalidades que se mostraron en ambos sistemas son las siguientes: ingreso de información del paciente, censo valorado de pacientes, valoración del paciente en los

diversos eventos de generación de información y en el sistema FRCIN se mostro la funcionalidad de creación del informe mensual. Después de realizar esas actividades se realizo una serie de preguntas, una de ellas es la siguiente: ¿considera que el CDSS es útil para su trabajo?, de lo cual comentaron lo siguiente:

*“Me facilita la toma de decisión ya que el CDSS está basado en datos estadísticos que tienen en consideración los factores de riesgo para trata de evitar eventos futuros de caso de infección nosocomial”*

*“El sistema considera todos los elementos de información de la forma FRCIN los cuales son los necesarios para el proceso de detección y vigilancia epidemiologia del hospital”*

#### **Jefe de departamento de epidemiologia**

*“Si en ambos sistemas contiene toda la información relacionada con el paciente y está acorde a nuestro flujo de trabajo de la detección y vigilancia y control de infecciones nosocomiales”*

*“El sistema administra y maneja toda la información necesaria para llevar a cabo mis actividades para la detección de infecciones nosocomiales”*

#### **Epidemióloga**

#### **VII.3.4.2. Facilidad de uso percibida**

Para obtener la percepción de la facilidad de uso de los sistemas móvil tanto en su versión de escritorio, se realizaron una serie de actividades correspondientes al ingreso de información del paciente y su posterior consulta de información estas actividades se realizaron en ambos sistemas. Al finalizar la sesión de actividades se realizo una serie de

preguntas la primera de ella fue: ¿considera fácil el uso de los sistemas CDSS móvil y el sistema FRCIN? De lo cual comentaron lo siguiente:

*“me adaptaría fácilmente a utilizarlo, debido a que contiene tiene todos los elementos del proceso de detección y vigilancia de infecciones nosocomiales”*

*“El sistema móvil es fácil de usar y tiene la ventaja de que ya no tengo que cargar todos los papeles sobre los pacientes ya que el sistema tiene toda la información relacionada de los mismos”*

**Jefe de departamento epidemiología**

*“Considero al sistema fácil de usar, pero creo que necesitaría un poco mas de práctica”*

**Epidemióloga**

### **VII.3.4.3. Tipos de modelos**

Para la evaluación de utilidad de los diferentes modelos predictivos implementados en las diversas fases de generación de información se procedió de la siguiente manera: se tomaron varios ejemplos del conjunto de datos que no se uso para el entrenamiento del los modelos.

Posteriormente se mostraron ante los médicos para que valoraran la situación del caso de forma cualitativa y determinaran si es un caso con alto riesgo y bajo riesgo de infección nosocomial. Posteriormente se realizo lo mismo con el CDSS para que vieran la valoración con los diversos modelos predictivos implementados. Adicionalmente a los diversos casos le agregaron elementos de información correspondientes a los eventos de generación de información. Al finalizar la sesión de actividades se realizo una serie de preguntas la primera de ella fue: ¿Cuál de los diferentes modelos predictivos considera más útil? De lo cual comentaron lo siguiente:

*“Ambos son útiles, en algunos casos el método basado en probabilidad se acercaba con más certeza al caso clínico del paciente sin embargo el método del árbol nos decía paso a paso el razonamiento que uso para llegar a ese resultado”*

*“las reglas mostradas del árbol de decisión se parece al razonamiento clínico que usamos para determinar de un paciente*

#### **Jefe de departamento de epidemiología**

*“Considero que el modelo basado en probabilidad son más útiles ya que muestran la valoración del paciente en términos de probabilidad”*

#### **Epidemióloga**

A continuación se muestra una discusión de los resultados.

Una vez realizada la sesión de evaluación se analizó la información obtenida. Esta información se presenta de manera condensada en la Tabla XXXVIII.

**Tabla XXXVIII. Condensado de los resultados obtenidos en la evaluación preliminar del CDSS.**

<b>Tópico</b>	<b>Jefe de departamento de epidemiología</b>	<b>Epidemióloga</b>
<b>Utilidad</b>	Lo considero útil	Lo considero útil
<b>Manejo de información</b>	Considera los elementos de información del FRCIN	El manejo de información es suficiente
<b>Facilidad de uso</b>	Contiene los elementos con los cuales trabajo	Considera fácil el uso del sistema
<b>Tipos de modelos</b>	Ambos tipos de modelos son útiles	Considera mejor el modelos basado en probabilidad

### **VII.3.5. Discusión de los resultados**

De acuerdo a los resultados obtenidos, específicamente en cuanto en la utilidad del CDSS, el jefe de departamento y la epidemióloga expresaron que los diferentes modelos predictivos son adecuados para la valoración de riesgo de infección nosocomial que tiene los pacientes del hospital. Adicionalmente el jefe de departamento de epidemióloga expreso que el CDSS está embebido dentro del todo el PDVIN. Lo cual demuestra que se realizo un una correcta captura del proceso.

En base a la captura del proceso se realizo una obtención de requerimientos del sistema donde se identificaron los elementos de información utilizados durante el PDVIN, esto se ve reflejado en el CDSS debido a que por el jefe de departamento de epidemiologia y la epidemióloga sobre el manejo de información el cual consideraron adecuado a sus necesidades.

Por otro lado en el aspecto de la facilidad de uso de los sistemas, la epidemióloga considera que el sistema móvil es fácil de usar, sin embargo expreso que necesita de más práctica, debido a la ausencia de experiencia en el manejo de este tipo de dispositivos

En cuanto la percepción que tienen el jefe de departamento de epidemiologia y la epidemióloga sobre los modelos predictivos implementados en los sistemas CDSS móvil y el sistema FRCIN, ambos coincidieron en que los modelos basados en probabilidad reflejan con mas certeza la el caso clínico del paciente, sin embargo expresaron que las reglas mostradas por los arboles son muy útiles ya que representan paso a paso el razonamiento que uso el modelo para llegar a ese resultado.

Finalmente, tanto el feje de departamento de epidemiologia y la epidemióloga, expresaron aceptación por el CDSS, haciendo énfasis en los beneficios que se tendrían con el soporte brindado al PDVIN, a si que el CDSS refleja la forma en cómo ellos trabajan.

#### **VII.4. Resumen.**

En este capítulo se mostro la evaluación de los modelos predictivos implementados en el CDSS móvil y en el sistema FRCIN utilizando un conjunto de datos que no se utilizo para la creación de los mismos. Con estas pruebas se demostró que los modelos obtenidos por medio del proceso de extracción de conocimiento son lo suficiente robustos en cuanto al porcentaje de instancias clasificadas correctamente y los valores de precisión y alcance.

También en este capítulo se describió la evaluación preliminar del CDSS realizada con el jefe de departamento y el jefe de cirugía, el objetivo de esta evaluación y las actividades que se llevaron a cabo para su realización. Los resultados principales que se obtuvieron reflejan una buena aceptación del CDSS, ya que tanto el jefe de departamento de epidemiología como la epidemióloga consideran que el soporte proporcionado por el CDSS es útil para la detección de infecciones nosocomiales, además que su forma de trabajo se ve reflejada, al considerar los aspectos técnicos, sociales y de información existentes en el PDVIN.

## Capítulo VIII

---

### Conclusiones

---

#### VIII.1. Conclusiones

Las tendencias actuales sugieren que el desarrollo de CDSS's involucren una perspectiva socio-técnica y movilidad para dar un mejor apoyo al flujo de trabajo en el contexto de decisiones y movilidad. En este trabajo de tesis se presenta un CDSS diseñado a partir de este enfoque se obtienen los requerimientos del siguiente caso de estudio: proceso de detección y vigilancia de infecciones nosocomiales (PDVIN).

De acuerdo al estudio se identificaron cuatro fases de generación de información en el llenado del formato de registro de caso de infección nosocomial (FRCIN). El FRCIN es la base para el proceso extracción de conocimiento y la aplicación de los métodos de minería de datos como los arboles de decisión, el clasificador *Naive Bayes* para la obtención de modelos predictivos son embebidos en un CDSS móvil con el objetivo de detectar una infección nosocomial (IN) y usadas en las distintas fases correspondientes al proceso del PDVIN. Para finalizar se evaluó un prototipo de CDSS para conocer la percepción de utilidad que tienen los médicos sobre un CDSS que estén basados en técnicas de minería de datos.

Bajo este contexto el primer paso fue análisis de un escenario clínico real con técnicas cualitativas y de investigación de campo para capturar el proceso del PDVIN, consecuencia de este estudio se encontró que, el diseño del CDSS debe estar embebido en el contexto del proceso organizacional, para brindarle soporte con tecnología de ayuda a la toma de decisión al PDVIN, utilizando dos modelos predictivos: arboles de decisión y *Naive Bayes*.

Adicionalmente se encontró que durante el PDVIN existe una gran necesidad de movilidad, debido a que el personal de la unidad de vigilancia epidemiología del hospital (UVEH) tiene que realizar un recorrido para visitar a cada uno de los pacientes.

Se identificó que la información del paciente se genera de forma dinámica a causa que los datos se generan conforme a su estadía en el hospital se decidió dividir en eventos de generación de información que corresponden a los modelos predictivos por las técnicas de minería de datos.

Por un lado, otro producto interesante del caso de estudio es conocer la importancia de aprovechar la memoria organizacional con la que cuentan los hospitales, tal como lo es el FRCIN debido a que representa la base de conocimiento para la extracción de modelos predictivos, que es la base del CDSS móvil desarrollada.

Entonces en base a los eventos de generación de información se inició el proceso de extracción de conocimiento con el objetivo de obtener modelos predictivos para la detección oportuna de IN.

En el proceso de extracción de conocimiento se identificaron que los métodos de árboles de decisión y el clasificador *Naive Bayes* son los más robustos para nuestro conjunto particular de datos.

Consideramos que para llevar a cabo el proceso de transformación de datos y generalización de conceptos es necesaria la participación del experto del proceso en este caso el jefe de departamento de epidemiología.

Consideramos que en todo proceso de extracción de conocimiento debe de contar un proceso de reducción de la dimensionalidad de la vista minable como el que se realizó en este proceso para aumentar la calidad de los modelos predictivos.



Como producto final del proceso de extracción de conocimiento se obtuvo los diferentes modelos predictivos (*Naive Bayes* y en arboles de decisión) en sus respectivos eventos de generación de información para la detección de infección nosocomial, el clasificador bayesiano en la fase de revisión del paciente tiene un 88% de instancias clasificadas correctamente por otro lado el modelo de árbol de decisión tiene un 90% en la fase de detección de IN solo cirugía.

Con los requerimientos con una perspectiva socio-técnica obtenida mediante el caso de estudio y con el conjunto de modelos predictivos se diseñó una arquitectura compuesta por tres sub-sistemas: uno de tipo móvil, uno de tipo base y un servidor. La funcionalidad principal de este CDSS es el apoyo a la toma de decisiones en base a modelos predictivos obtenidos a través de un proceso de extracción de conocimiento donde se utilizaron técnicas de minería de datos.

El diseño del CDSS móvil da soporte al PDVIN y a sus necesidades de movilidad al personal de la UVEH. Entonces con el diseño se procedió con la implementación de los sistemas móviles y base.

Finalmente se desarrolló un prototipo de CDSS con el fin de obtener una evaluación para obtener la percepción de utilidad analizándose de forma cualitativa.

Los resultados de esta evaluación arrojan que el jefe de departamento de epidemiología y la epidemióloga reflejan aceptación del CDSS que está basado en técnicas de minería de datos debido a que consideran que el soporte proporcionado por el CDSS es útil para la detección de infecciones nosocomiales, además que su forma de trabajo se ve reflejada, al considerar los aspectos técnicos, sociales y de información existentes en el PDVIN.

Otro aspecto interesante arrojado por esta evaluación, fue la percepción de utilidad que tienen los medios sobre los diversos modelos predictivos implementados en el CDSS, en el cual desde un inicio se consideró que los modelos predictivos producto de los árboles de

decisión tendrían mayor aceptación debido a que muestran el razonamiento que uso para llegar a esa conclusión sin embargo los que tuvieron mayor percepción de utilidad fueron los modelos predictivos bayesianos.

Finalmente en este trabajo de tesis consideramos que este es el enfoque adecuado para que en el futuro cualquier aplicación que se haga en esa dirección tenga este ciclo de desarrollo.

## **VIII.2. Aportaciones**

Las contribuciones principales de este trabajo de tesis giran alrededor del objetivo de este trabajo: *“explorar cómo las técnicas de minería de datos pueden apoyar a un proceso de toma de decisiones en un contexto clínico, a través del desarrollo y evaluación de un prototipo que utilice cómputo móvil”* enseguida se describen brevemente estas contribuciones:

- Realizar un diseño de un CDSS construido desde una perspectiva socio-técnica para darle soporte a un proceso organizacional como el PDVIN. Donde se embeben modelos predictivos obtenidos a partir de un proceso de extracción de conocimiento.
- Un conjunto de modelos predictivos obtenidos mediante el uso de clasificador bayesiano simple y arboles de decisión, que tienen como objetivo la detección oportuna de una infección nosocomial.
- El desarrollo de un prototipo funcional de un CDSS móvil con el que evaluamos cualitativamente la percepción de los médicos con respecto a la utilidad del CDSS de la herramienta y la generación del informe mensual de infecciones nosocomiales.
- Exploramos la percepción de utilidad que tienen los médicos sobre los diferentes modelos predictivos implementados en el CDSS.

- Finalmente la realización de este trabajo genero una publicación: “Minería de Datos y Cómputo Móvil en la Toma de Decisiones Médicas”, presentado en el Symposium de Computación Clínica e Informática Médica celebrado en Mexicali, B.C., México en octubre del 2008.

### **VIII.3. Trabajo a futuro**

Como se ha mencionado, el estudio del proceso de toma de decisiones de trabajadores del cuidado de la salud es muy importante, debido a que sus actividades están encaminadas a este fin. Se encontró interesante proponer el siguiente trabajo futuro:

- Realizar el proceso de extracción de conocimiento con un mayor conjunto de datos y de esta forma poder crear modelos predictivos para valorar el esquema de tratamiento microbiológico
- Explorar el uso de máquinas de soporte vectorial para la obtención de modelos predictivos más robustos.
- Para hacer posible que la implementación práctica de un CDSS hace falta la integración de bases de información como el expediente electrónico de pacientes esto con el fin de poder acercar sin redundancia de datos la información que los trabajadores de conocimiento para tomar decisiones.

## Referencias

- Bardram, J.E. y Bossen, C.** 2005, Mobility Work: The spatial dimension of collaboration at a hospital. *Journal of computer supported Cooperative Work*. 14: 131-169 p.
- Bellazzi, R. Zupan B., 2008.** Predictive data mining in clinical medicine: Current issues and guidelines. *Journal of medical Informatics* 77:81-97 p
- Berlin, A. Sorani, M. Sim, I. 2006.** A taxonomic description of computer-based clinical decision support systems. *Journal of Biomedical Informatics*. 39: 655-660 p.
- Bramer M. 2007.** Principles of Data Minig, Undergraduate Topics in Computer Science. First edition, Springer-verlag Berlin-Heidelberg 70 pp.
- Buntine, W. L. 1994.** Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159-225 p.
- Checkland y J Scholes. 1999** “Soft System Methodology in Action”. JohnWiley and Sons, 418 pp.
- Chickering, M ., D .Geiger, D Heckerman, 1996,** learning Bayesian networks in NP complete, in learning from data: Artificial and Statistics. Springer Verlag. 121-130 p.
- Choo, W.C. 1999.** La organización inteligente: El empleo de la información para dar significado, crear conocimiento y tomar decisiones. Editorial Oxford University Press. México D.F. 346 pp.
- Clark P y Boswell R. 1991,** Rule induction with CN2: Some recent improvements. En : Proceedings of the Fifth European Working Session on Learning. Berlin: Springer, 151–163 p.
- Coiera, E. 2003.** Guide to Health Informatics. Holder Arnold Publication. Second edition. London. 334 pp.
- Collen MF. 1999.** The evolution of computer communications. *MD Comput.* 1999; 16 (4): 72–76 pp.

**Cooper, G. y Herskovits, E. 1992.** A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9: 309-347p.

**Cuenca J. Fernández G., López de Mántaras R., Verdejo F. 1986.** Inteligencia Artificial: Sistemas Expertos. Alianza Editorial 131 pp

**Dahlström, Ö. Thyberg, I. Hass, U. SKogh, T. Timpka, T. 2006.** Designing a Decision Support System for Existing Clinical Organizational Structures: Considerations From a Rheumatology Clinic. *Springer*. 30: 325-330 p.

**DeSanctis, G. y Gallupe, R.B., 1987.** A Foundation for the study of Group Decision Support System. *Management Science*. 33(5): 589-609 p.

**Dillon, S.M. 1998.** Descriptive decision making: comparing theory with practice. *Operational research society of new Zealand annual conference*. 33-61 p.

**Englehardt y Nelson. 2001.** Health Care Informatics: An Interdisciplinary Approach. Mosby, first edition , 549 pp

**Fayyad, U. Shapiro, G. y Smyh, P. 1996.** From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*. 37-42 p.

**Fisher, S. Steward, T. MEHTA, S. WAX, R. LAPINSKY, S., 2003.** Handheld Computing in Medicine. *Journal of the American Medical Informatics Association*. 10(2): 139-140 p.

**Fowler, M. 2004.** UML distilled. Tercera edición. Addison Wesley, Boston, 175 pp.

**Fox, J. 2001.** Designing Safety into Medical Decisions and Clinical Processes. *SPRINGER SAFECOMP 2001, Budapest, Hungary, September 26-28, 2001, LNCS 2187*, 1-2 pp.

**French, S. 1998.** Notas de curso de maestría: Analysis and Decision Support System. University of Manchester. No publicado.

**González, V. M., Tentori, M.E., Morán, E.B., Favela, J., y Martínez, A.I. 2005.** Under Standing mobile work in a distributed information space: implications for design of ubicomp technology. *Proc. of CLIHC 2005, ACM Press. Cuernavaca México*. 52-63pp.

**Gorry G.A. y Scott-Morton M.S. 1971.** A framework for management information system. *Sloan Manage.* 13(1): 55-70 p.

**Han J. y Kamber M.. 2006.** *Data Mining: Concepts And Techniques.* Second Edition, Elsevier. 618 pp.

**Hernández, O. Ramírez, Q. Ferri, R. 2004** *Introducción a la Minería de Datos.* Pearson Prentice Hall. Primera edición, 544 pp.

**Holroyd Brian R, MD, Michael J. Bullard, MD, Timothy A. D. Graham, MD, MSc, Brian H. Rowe, MD, MSc, 2007** *Decision Support Technology in Knowledge Translation, Academic Emergency medicine,* 14:942-948 p.

**Hunt E. B. Marin J. y Stone P.T 1966.** *Experiments in induction.* Academic press 14p.

**Imhoff, M, Webb A y Goldschmidt A. 2000.** *Health informatics.* Elsevier, *Intensive Care Medicine.* 27:1-5 p.

**Jacobson, I., Griss, M. y Jonsson, P. 1997.** *Software Reuse: Architecture, Process and Organization for Business Success.* Addison Wesley. New York. 497pp.

**Kaplan, B. 2001.** *Evaluating informatics applications –clinical decision support system-literature review.* ELSEVIER *International Journal of Medical Informatics.* 64: 15-17 p.

**Laplante, P.A. Nelly, C.J., y Sangwan, R. 2006.** *Healthcare Professionals' Perceptions of Medical Software and What to do About it.* *Computer IEEE.* 39(4): 29-29 p.

**lavrač, N. 1999.** *DMachine Learning for data Mining in Medicine.* J. Stefan Institute, Jamova 39: 47-48 p.

**lavrač, N. 2001.** *Data Mining in Medicine: Selected Techniques and Applications.* J. Stefan Institute, Jamova. 39: 1-2 p.

**López de Castilla Vásquez C. 2005.** *Clasificadores por redes bayesianas.* Tesis de Maestría. Universidad de Puerto Rico, Mayagüez, Puerto rico. 13,14 17-23 p.

**Mangalampalli, A. Mangalampalli, M. Chakravarthy, R. Jain, K. 2006.** *A neural network based clinical decision –support system for efficient diagnosis and fuzzy- based*

prescription diseases using homoeopathic medicinal system. *ELSEVIER Expert System With Applications*. 30: 109- 112 p.

**Marakas, G.M. 2003** Decision Suport Systems in 21 Century. Prentice Hall. Second edition. New Jersey. 611 pp.

**Mark, D:B. 2002.** Tome de decisiones en medicina clínica. En: Braunwald, E., E., Fauci, A.S., Kasper, D.L., Hauser S.L. Longo, D.L y Jameson J.L. (eds). 15ª. Edición Harrison's: principios de medicina interna. McGraw-Hill. New York. 1521 pp.

**Michalowski, W., Rubin, S., Slowinski, R. y Wilk, S. 2003.** Movable Clinical support system for pediatric emergencies, *ELSEVIER Decision Support System* 36(2): 161-176 p.

**Mitchell T.M. 1997.** Machine Learning. First Edition. Computer Science Series, McGraw-Hill International Editions, Singapore. 543 pp.

**Monk, A. y Howard, S. 1998.** The Rich Picture: A Tool for Reasoning About Work Context. *Interactions, ACM*. 5(2):21-30 p.

**NOM-045-SSA2,-2005,** Norma Oficial Mexicana Para la vigilancia epidemiológica, prevención y control de las infecciones nosocomiales, unidad de calidad y normatividad validado y registrado 28-nov-2005., folio 015.

**Ofstad, H. 1961.** An inquiry into the freedom of decision. Norwegian Universities Press. Oslo. 15 pp.

**Osheroff, J. A. Pifer E.A., Siting, D.F., Jenders, R.A. y Teich J.M., 2004.** Clinical Decision Suport Implementers, *WorkBook. Healthcare Information and Management System Society*. 68 pp.

**Pacheco-Soto, M.J. 2004.** Soporte a la Toma de Decisiones con un Enfoque a la Ingeniería de Procesos. Tesis de Maestría. CICESE. Ensenada, México. 324 pp.

**Patel V. L., Kaufman D. R y Arocha, J. 2002** Emerging paradigms of cognition in medical decision-making. *Journal Biomedical Informatics*. 35: 52-75 p

**Perreault L. y Metzger. 1999.** A pragmatic framework for understanding clinical decision support. *Journal of Healthcare Information Management*. 5-22 p.

**PROY-NOM-045-SSA2, 2007** Procedimiento para realizar la vigilancia epidemiológica de infecciones nosocomiales en las unidades médicas hospitalarias del IMSS, unidad de organización y calidad validado y registrado 24-agosto-2007, folio 064.

**Quinlan J.R, 1986**, “Induction of decision trees”. *Machine Learning*, Vol. 1(1), 81–106p

**Quinlan, 1993**. C4.5: “Programs for Machine Learning”. Morgan, Kaufmann, San Mateo, CA, USA.

**San Pedro, J., Burstein, F., Cao, P., Churilov L., Zaslavsky A., y Wassertheil. J. 2004.** Mobile Decision Support for Triage in Emergency Departments. *Proc. of IFIP Int. Conf. 2004, on Decision Support System: 372-379 p.*

**Schwarz, G. 1978.** Estimating the dimension of a model. *Annals of Statistics*, 6: 461-464p.

**Shannon C., Weaver W. 1948.** The mathematical theory of communication. University of Illinois Press, Urbana. 5-7p,

**Shim, J. Warkentin, M. Courtney, F. Power, D. Sharda, R. Carlsson, C. 2002.** Past, present, and future of decision support technology. *Decision Support System*. 33: 111-125 p.

**Simon, H.A. 1960.** The new science of management decision. Harper & Row, publishers. New York. 50 p.

**Sumathi S y Sivanandam S.N. 2006.** Introduction to Data Mining and its Applications. *Studies in Computational intelligence, Volume 29.* Springer-verlag Berlin-Heidelberg. 653 pp.

**Tsumoto, S. 2004.** Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. *Information Sciences* 162: 65–80 p.

**Witten, I. y Frank, E. 2005.** *Data Mining Practical Machine Learning Tool and Techniques.* Second edition. Morgan Kaufmann Publishers. 543 pp.



## **Apéndice A Modelado del proceso PDVIN con Rads y DFD**

### **Diagramas rol actividad**

Los Diagramas Rol Actividad o RAD (por sus siglas en ingles Rol Activity Diagrams) son un herramienta poderosa para modelar procesos, en la cual podemos representar actividades, roles, interacciones y decisiones.

La idea central del RAD es poder representar el proceso como un conjunto de actividades e interacciones de los roles. Con esta técnica, pueden evidenciarse algunas situaciones que impiden el correcto funcionamiento del proceso como cuellos de botella, excesivas actividades para un rol, etc.

Mostramos a continuación, los RAD para el proceso de detección, vigilancia y control de infecciones nosocomiales del hospital

Las abreviaturas utilizadas en los RAD son:

UVEH. = Unidad de vigilancia epidemiológica del hospital

FRCIN = formato de registro de caso infección nosocomial

El rad inicial se llama detectando, vigilando y controlando la infección nosocomial

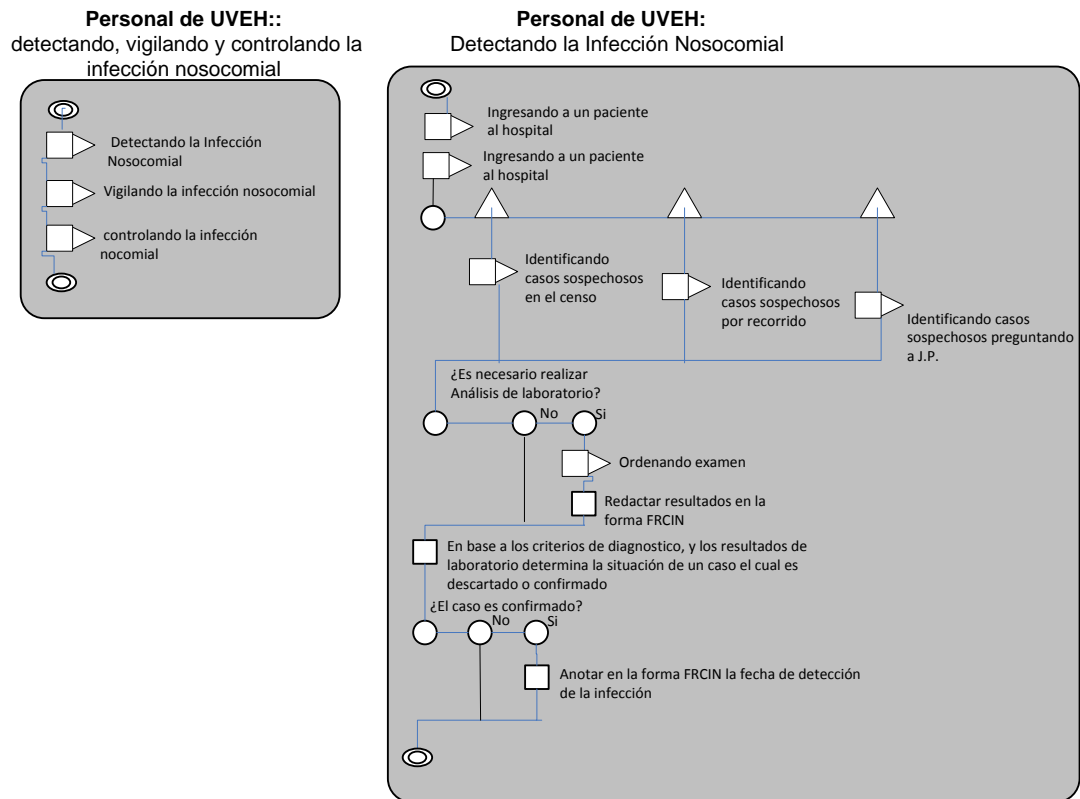
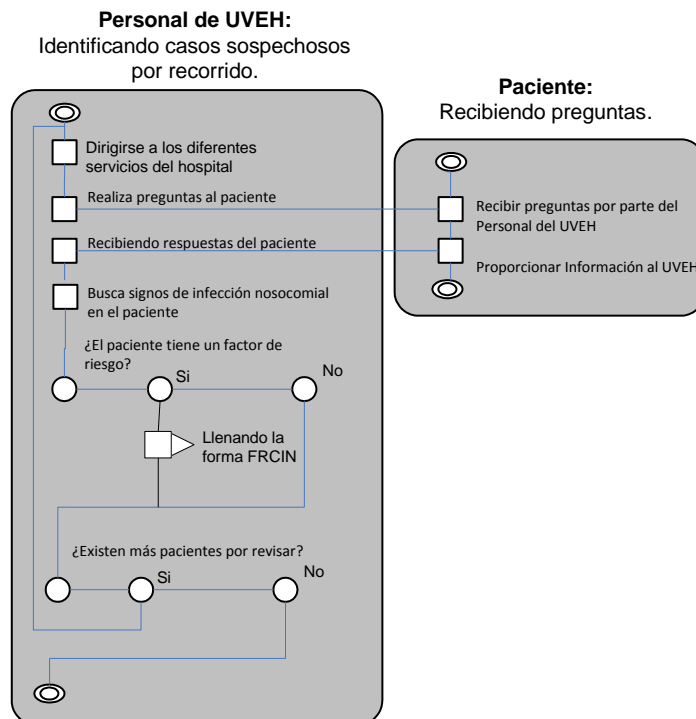
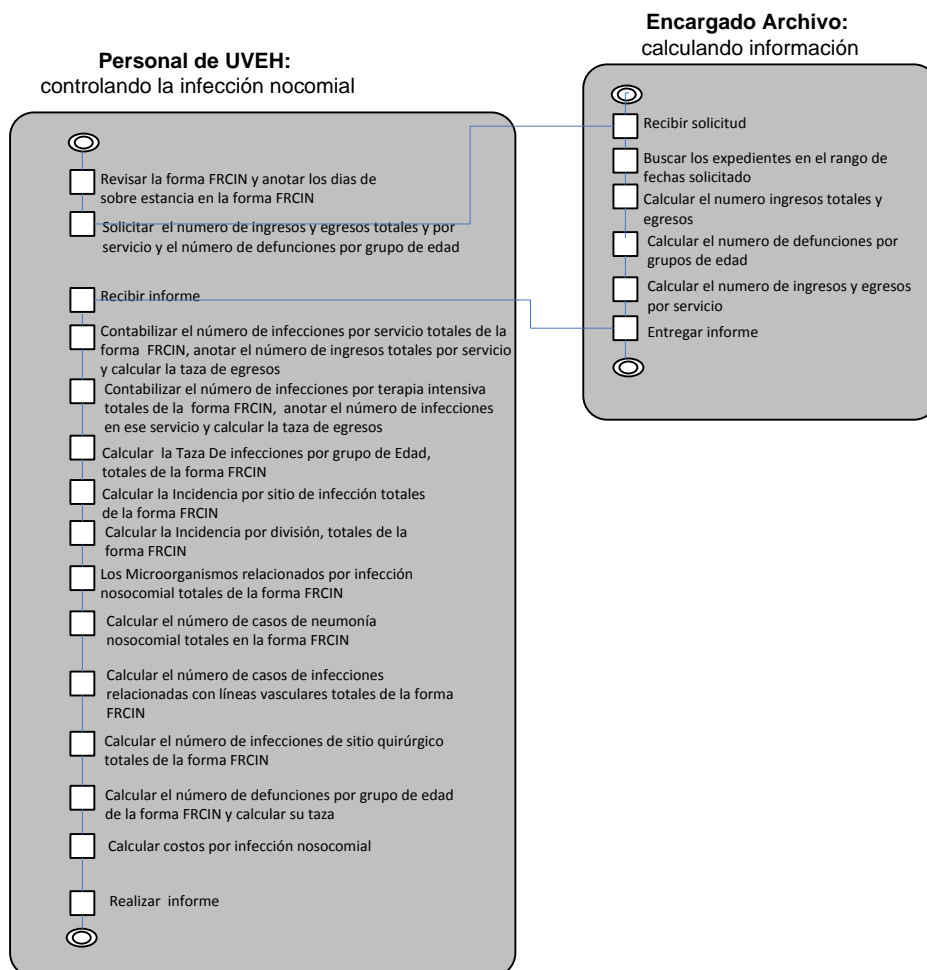


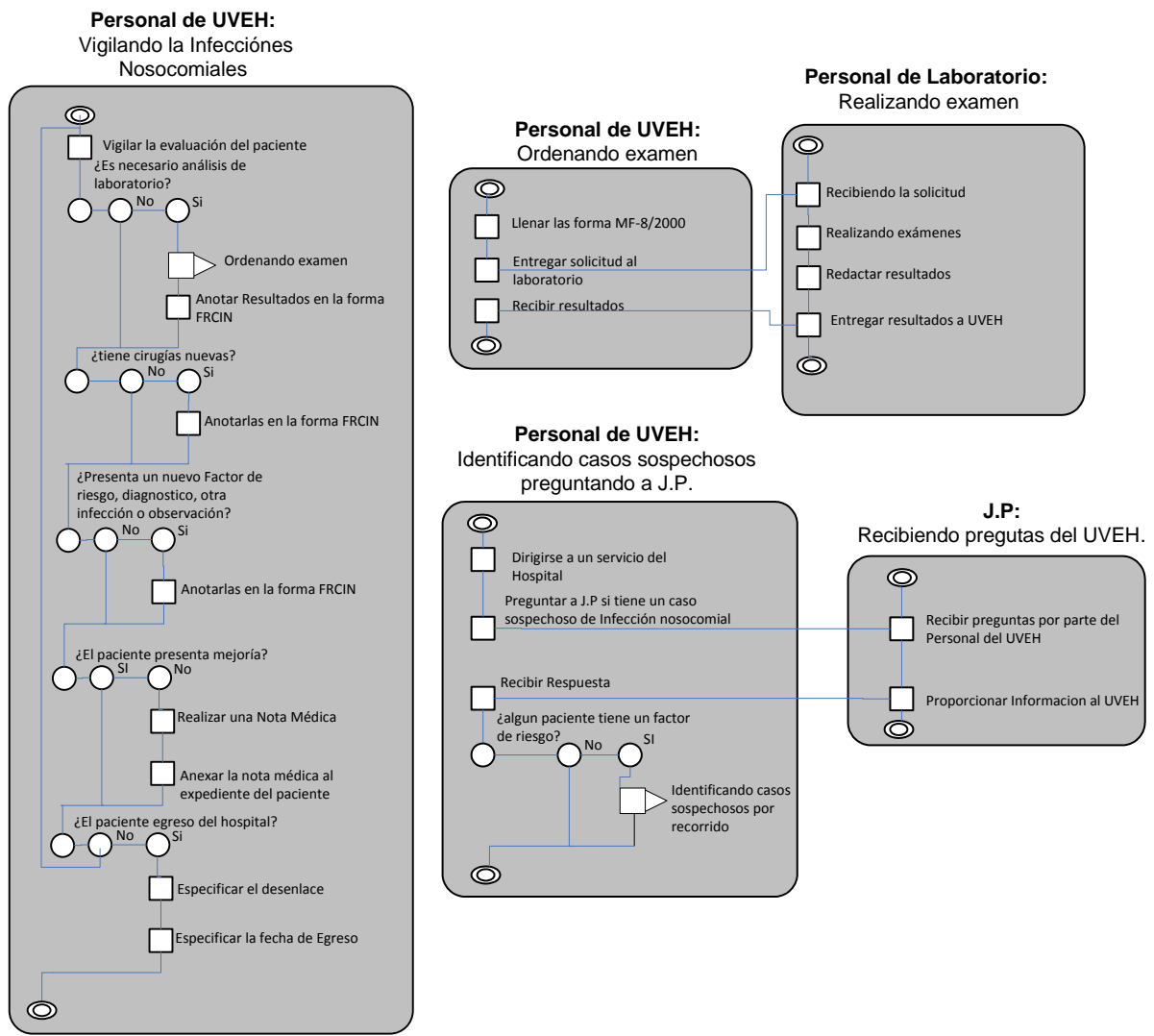
Figura 55. Rad detectando vigilando la infección nosocomial



**Figura 56. Rad identificando casos sospechosos por recorrido**



**Figura 57. Rad sobre controlando la infección nosocomial**



**Figura 58. Vigilando infecciones nosocomiales.**

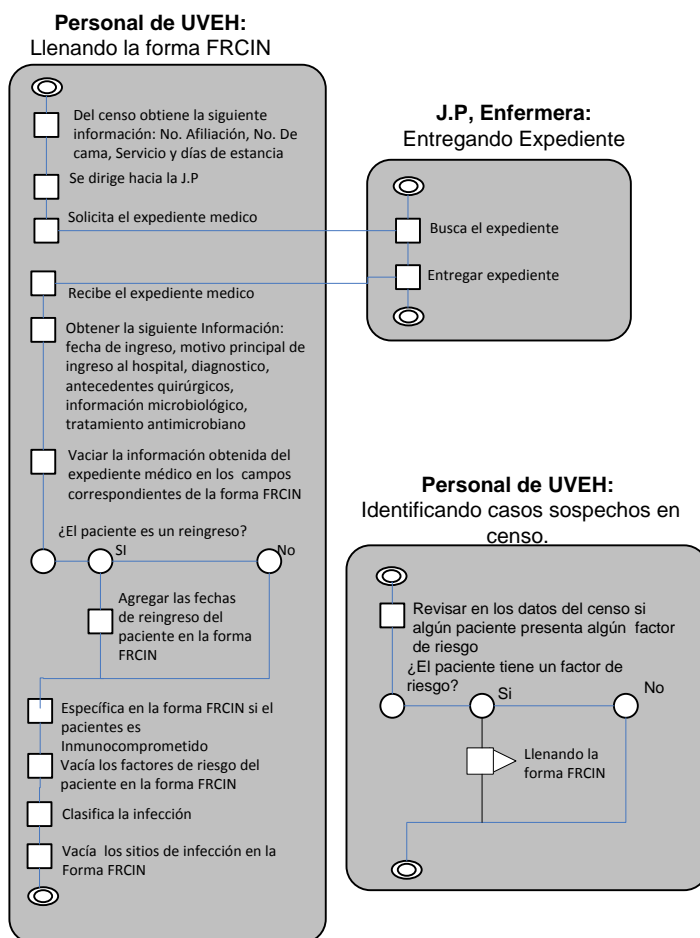


Figura 59. Rad Llenado de la forma FRCIN

## Diagrama de flujo de datos

El Diagrama de flujo de Datos, como su nombre lo dice, nos provee de la visión general acerca de cómo fluyen los datos en nuestro proceso, si sufren cambios, etc.

“El diagrama de flujo de datos (DFD), es una herramienta que permite visualizar un sistema como una red de procesos funcionales, conectados entre sí por ‘conductos’ y ‘tanques de almacenamiento’ de datos. Siendo éste, una de las herramientas más comúnmente usadas.

- Los componentes de un diagrama típico de flujo de datos:

- Proceso.
- Flujo.
- Almacén.

A continuación presentamos el diagrama de flujo de datos del proceso de detección, vigilancia y control de infecciones nosocomiales

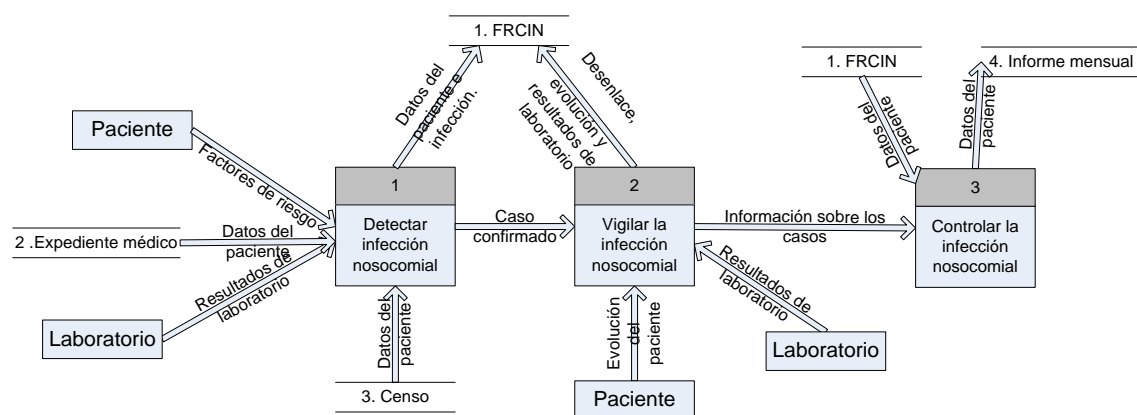


Figura 60. Vista global del diagrama del flujo de datos

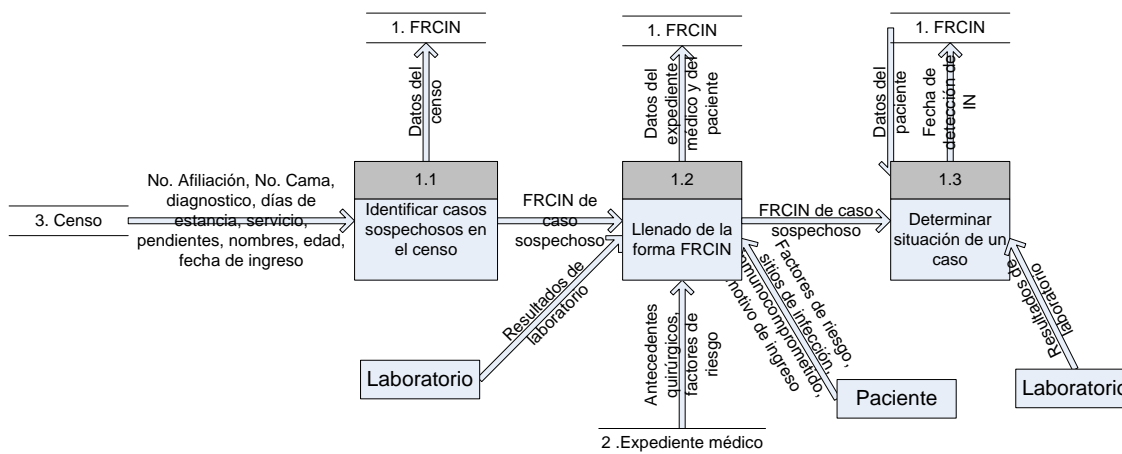


Figura 61. DFD sobre en la fase de detección de infección nosocomial

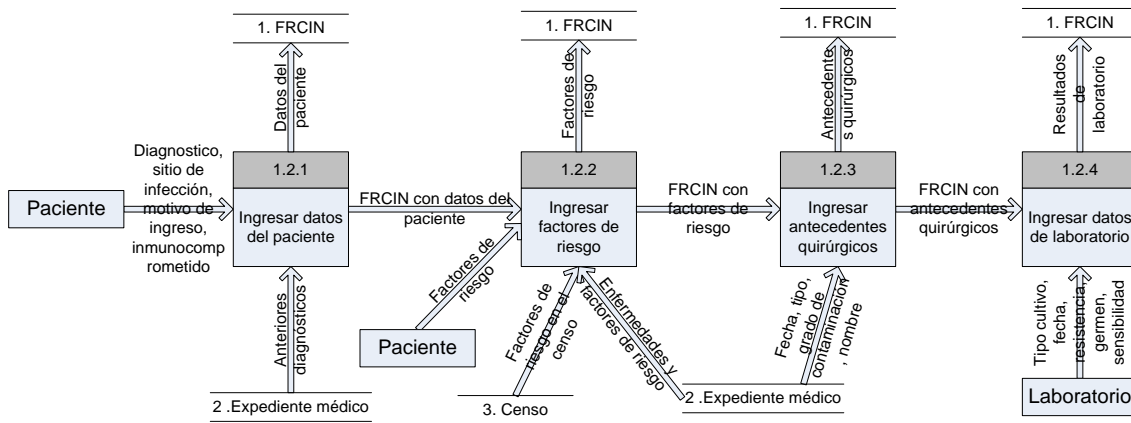


Figura 62. DFD de llenado de la forma FRCIN

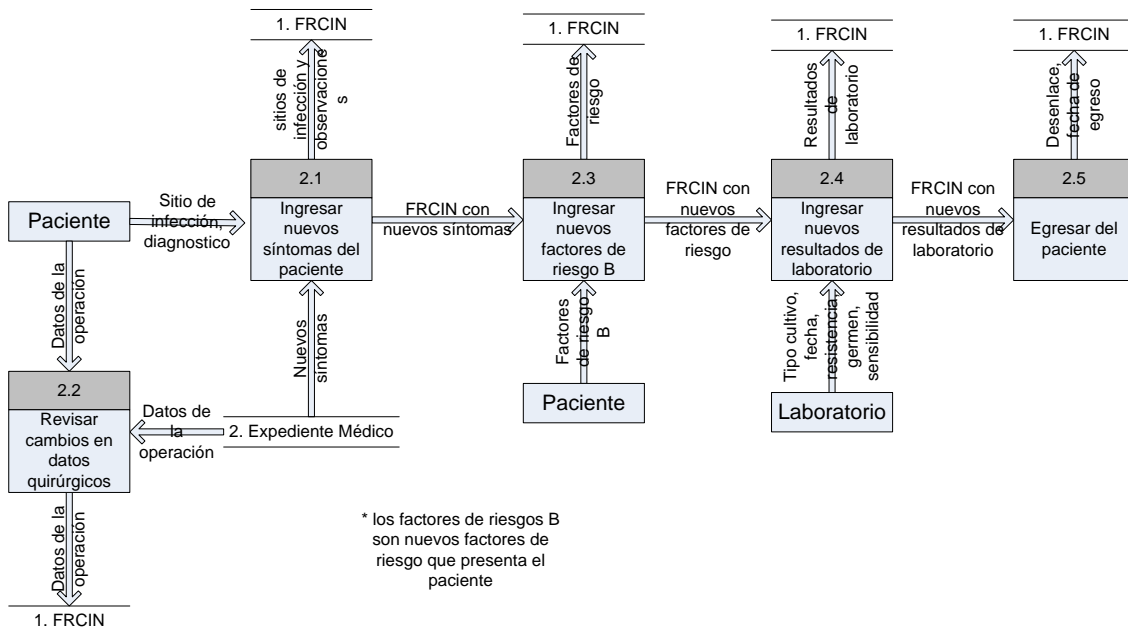


Figura 63. DFD sobre la fase de vigilancia

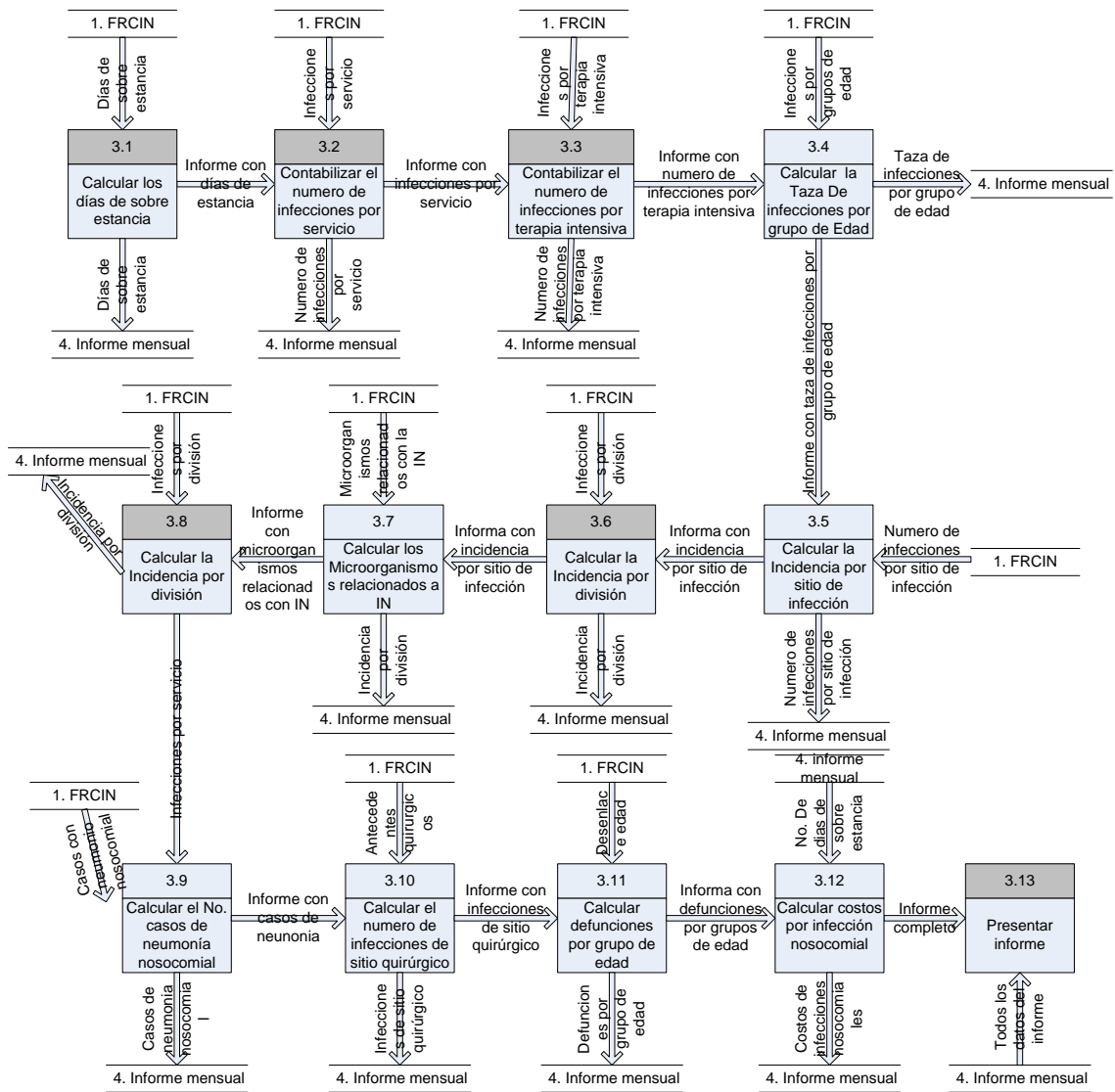


Figura 64. DFD sobre el proceso de control de infección nosocomial



## Apéndice B. Valores del atributo antibiótico y germen

En este apéndice se encuentran los valores posibles que puede tomar el atributo sensibilidad y germen. Esto se puede ver en la *Tabla XXXIX*.

*Tabla XXXIX. Valores posibles del atributo sensibilidad y germen*

Antibióticos
<i>Ampicilina, Amikacina, Cefalotina, Cefotaxima, Ceftazidima, Ceftriaxona, Ciprofloxacino, Clindamicina, Gatifloxacina, Gentamicina, Imipemen, Metronidazol, Meropenem, Eritromicina, Cloranfenicol, Acido Nalidixico, Acido Nalidixico, Amoxicilina</i>
Germen
<i>ACINETOBACTER SPP, AEREOMONAS SPP, CAMPYLOBACTER SPP, ASPERGILLUS, BURKHOLDERIA CEPACIA, CANDIDA PARAPSILOSIS, CANDIDA KRUSEI, CANDIDA GUILLERMONDII, CANDIDA GLABRATA, CANDIDA DUBLINIENSIS, CANDIDA ALBICANS, CANDIDA PSEUDOTROPICALIS, CANDIDA SPP, CANDIDA TROPICALIS, CITOMEGALOVIRUS, CITROBACTER FREUNDII, CITROBACTER SPP, CLAMYDIA TRACHOMATIS, CLOSTRIDIUM SPP, CLOSTRIDIUM DEFFICILE, CLOSTRIDIUM PERFRINGES, CORINEBACTERIUM SPP, CRYPTOCOCCUS SPP, CRYPTOSPOIDIUM, ENTAMOEBIA HISTOLYTICA, ENTEROBACTER AEROGENES, ENTEROBACTER AGGLOMERANS, ENTEROBACTER AMNIGENUS, ENTEROBACTER ASBURIAE, ENTEROBACTER CANCEROGENUS, ENTEROBACTER CLOACAE, ENTEROBACTER FREUNDII, ENTEROBACTER SP, ENTEROCOCCUS AVIUM, ENTEROCOCCUS CASSELI FLAVUS, ENTEROCOCCUS DURANS, ENTEROCOCCUS FAECALIS, ENTEROCOCCUS FAECIUM, ENTEROCOCCUS GALLINARUM, ENTEROCOCCUS MUNDTI, ENTEROCOCCUS RAFFINOSUS, ENTEROCOCO BOVIS, ENTEROCOCO SPP, EPSTEIN BAR, ESCHERICHIA COLI, FLAVOBACTERIUM, HAEMOPHILUS INFLUENZAE, HEPATITIS A, HEPATITIS B, HEPATITIS C, HEPATITIS OTRAS, HERPES VIRUS, KLEBSIELLA SPP, KLEBSIELLA OXYTOCA, KLEBSIELLA PNEUMONIAE, KLEBSIELLA PNEUMONIAE SPP, KLEBSIELLA PNEUMONIAE SPP. OZAENAE, KLEBSIELLA RHINOSCLEROMATIS, KLEBSIELLA TERRIGENA, LEGIONELLA, LEUCONOSTIOC, LISTERIA SPP, MICROSCODIDIUM, MORGANELLA MORGANI, MORGANELLA SPP, MYCOBACTERIUM SPP, MYCOBACTERIUM TUBERCULOSIS, PROTEUS MIRABILIS, PROTEUS SPP, PROTEUS VULGARIS, PROVIDENCIA RETIGERII, PSEUDOMONAS AERUGINOSA, PSEUDOMONAS CEPACIA,</i>

*PSEUDOMONAS FLUORESCENS, PSEUDOMONAS SPP, PSEUDOMONAS STUTZERI, RALSTONIA PICKETTII, ROTAVIRUS, SALMONELLA GRUPO A, SALMONELLA SPP, SARAMPION, SERRATIA MARCESCENS, SERRATIA SPP, STAPHYLOCOCCUS EPIDERMIDIS, STAPHYLOCOCCUS CAPRAE, STAPHYLOCOCCUS CAPITIS, STAPHYLOCOCCUS AUREUS, SHIGELLA, SPHINGOMONAS PAUCIMOBILIS, STREPTOCOCCUS NEUMONIAE (NEUMOCOCO), STREPTOCOCCUS DEL GRUPO C Y GRUPO G, STENOTROFOMONAS SPP, STENOTROFOMONAS MALTOPHILA, STAPHYLOCOCCUS WARNERI, STAPHYLOCOCCUS SIMULANS, STAPHYLOCOCCUS SCHLEIFERI, STAPHYLOCOCCUS LUGDUNENSIS, STAPHYLOCOCCUS HOMINIS, STAPHYLOCOCCUS HAEMOLYTICUS, STREPTOCOCCUS PYOGENES, STREPTOCOCCUS SPP, STREPTOCOCCUS VIRIDANS (S. MILLERI), TORULOPSIS GLABRATA, VIH, VIRUS SINCISAL RESPIRATORIO, MICROCOCCUS SPP, ALCALIGENES XYLOSOXIDANS*

*OTROS GERMENES, NO SE AISLO, NO SE CULTIVO*

## Apéndice C Modelos predictivos del obtenidos del análisis de exploración

En el análisis de exploración se obtuvieron diversos modelos, nos limitaremos a mostrar un modelo por cada método de minería de datos utilizado, todos ellos corresponden al evento de revisión del paciente. Para finalizar con este apéndice se muestra los modelos para la detección de neumonía.

=== información del modelo ===

Instancias: 264

Atributos: 8

servicio  
 inmunocompromiso  
 sitio\_infeccion1  
 Cirugia  
 E\_Degenerativas  
 Cat\_Edad  
 Cat\_DE  
 Desenlace

Tipo de validación: cruzada de 10 particiones

### Método de árbol ID3

```
sitio_infeccion1 = Sistema_respiratorio
| E_Degenerativas = 0
| | servicio = Cirugia_y_Trauma: Alta_Mejoria
| | servicio = Ginecologia_y_obstetricia: Alta_Mejoria
| | servicio = Medicina_Interna
| | | inmunocompromiso = 0: Alta_Mejoria
| | | inmunocompromiso = 1
| | | | Cirugia = 0
| | | | | Cat_Edad = ninos: null
| | | | | Cat_Edad = adolescentes: null
| | | | | Cat_Edad = adulto: Alta_Mejoria
| | | | | Cat_Edad = adulto_mayor: Alta_Mejoria
| | | | Cirugia = 1
```

```

| | | | Cat_Edad = ninos: null
| | | | Cat_Edad = adolescentes: null
| | | | Cat_Edad = adulto: Alta_Mejoria
| | | | Cat_Edad = adulto_mayor: Alta_Mejoria
| | servicio = Pediatria: Alta_Mejoria
| | servicio = Unidad_de_cuidados_intensivos: null
| E_Degenerativas = 1
| Cirugia = 0
| | Cat_DE = Baja_estancia: null
| | Cat_DE = Estancia_media: Alta_Mejoria
| | Cat_DE = Sobreestancia: Finado
| | Cirugia = 1
| | | Cat_Edad = ninos: null
| | | Cat_Edad = adolescentes: null
| | | Cat_Edad = adulto: Alta_Mejoria
| | | Cat_Edad = adulto_mayor
| | | | Cat_DE = Baja_estancia: null
| | | | Cat_DE = Estancia_media: Alta_Mejoria
| | | | Cat_DE = Sobreestancia
| | | | servicio = Cirugia_y_Trauma: Alta_Mejoria
| | | | servicio = Ginecologia_y_obstetricia: null
| | | | servicio = Medicina_Interna: Alta_Mejoria
| | | | servicio = Pediatria: null
| | | | servicio = Unidad_de_cuidados_intensivos: null
sitio_infeccion1 = Sistema_cardiovascular: Alta_Mejoria
sitio_infeccion1 = Sistema_digestivo: null
sitio_infeccion1 = Sistema_genitourinario
| | Cat_DE = Baja_estancia: Alta_Mejoria
| | Cat_DE = Estancia_media: Alta_Mejoria
| | Cat_DE = Sobreestancia
| | servicio = Cirugia_y_Trauma
| | | Cirugia = 0
| | | | inmunocompromiso = 0: Alta_Mejoria
| | | | inmunocompromiso = 1
| | | | E_Degenerativas = 0: Finado
| | | | E_Degenerativas = 1: Alta_Mejoria
| | | Cirugia = 1: Alta_Mejoria
| | servicio = Ginecologia_y_obstetricia: null
| | servicio = Medicina_Interna
| | | E_Degenerativas = 0
| | | | inmunocompromiso = 0: Finado
| | | | inmunocompromiso = 1
| | | | Cirugia = 0
| | | | | Cat_Edad = ninos: null
| | | | | Cat_Edad = adolescentes: null

```

```

| | | | | Cat_Edad = adulto: Alta_Mejoria
| | | | | Cat_Edad = adulto_mayor: Alta_Mejoria
| | | | | Cirugia = 1: Finado
| | | | E_Degenerativas = 1
| | | | inmunocompromiso = 0
| | | | Cirugia = 0: Finado
| | | | Cirugia = 1: Alta_Mejoria
| | | | inmunocompromiso = 1: Alta_Mejoria
| | servicio = Pediatria: null
| | servicio = Unidad_de_cuidados_intensivos: null
sitio_infeccion1 = Sistema_nervioso_central: null
sitio_infeccion1 = Organos_y_tejidos
| servicio = Cirugia_y_Trauma
| | E_Degenerativas = 0: Alta_Mejoria
| | E_Degenerativas = 1
| | | Cat_DE = Baja_estancia: Alta_Mejoria
| | | Cat_DE = Estancia_media
| | | Cirugia = 0: Alta_Mejoria
| | | Cirugia = 1: Alta_Mejoria
| | | Cat_DE = Sobreestancia
| | | Cat_Edad = ninos: Alta_Mejoria
| | | Cat_Edad = adolescentes: null
| | | Cat_Edad = adulto: Alta_Mejoria
| | | Cat_Edad = adulto_mayor
| | | | Cirugia = 0: Alta_Mejoria
| | | | Cirugia = 1: Alta_Mejoria
| | servicio = Ginecologia_y_obstetricia: Alta_Mejoria
| | servicio = Medicina_Interna
| | | Cat_DE = Baja_estancia: Alta_Mejoria
| | | Cat_DE = Estancia_media
| | | | inmunocompromiso = 0
| | | | E_Degenerativas = 0: Alta_Mejoria
| | | | E_Degenerativas = 1: Finado
| | | | inmunocompromiso = 1
| | | | E_Degenerativas = 0
| | | | | Cat_Edad = ninos: null
| | | | | Cat_Edad = adolescentes: Alta_Mejoria
| | | | | Cat_Edad = adulto: Alta_Mejoria
| | | | | Cat_Edad = adulto_mayor: Alta_Mejoria
| | | | | E_Degenerativas = 1: Alta_Mejoria
| | | | Cat_DE = Sobreestancia
| | | | | inmunocompromiso = 0: Alta_Mejoria
| | | | | inmunocompromiso = 1
| | | | | E_Degenerativas = 0
| | | | | Cat_Edad = ninos: null

```

```

| | | | | Cat_Edad = adolescentes: null
| | | | | Cat_Edad = adulto
| | | | | Cirugia = 0: Finado
| | | | | Cirugia = 1: Alta_Mejoria
| | | | | Cat_Edad = adulto_mayor
| | | | | Cirugia = 0: Alta_Mejoria
| | | | | Cirugia = 1: Alta_Mejoria
| | | | E_Degenerativas = 1
| | | | Cirugia = 0
| | | | | Cat_Edad = ninos: null
| | | | | Cat_Edad = adolescentes: null
| | | | | Cat_Edad = adulto: Alta_Mejoria
| | | | | Cat_Edad = adulto_mayor: Alta_Mejoria
| | | | | Cirugia = 1
| | | | | Cat_Edad = ninos: null
| | | | | Cat_Edad = adolescentes: null
| | | | | Cat_Edad = adulto: Alta_Mejoria
| | | | | Cat_Edad = adulto_mayor: Alta_Mejoria
| servicio = Pediatria: Alta_Mejoria
| servicio = Unidad_de_cuidados_intensivos: Alta_Mejoria
sitio_infeccion1 = Hematico: Alta_Mejoria
sitio_infeccion1 = Otras_infecciones
| servicio = Cirugia_y_Trauma: Alta_Mejoria
| servicio = Ginecologia_y_obstetricia: null
| servicio = Medicina_Interna
| | Cirugia = 0
| | | E_Degenerativas = 0
| | | | inmunocompromiso = 0: Alta_Mejoria
| | | | inmunocompromiso = 1: Finado
| | | E_Degenerativas = 1: Alta_Mejoria
| | Cirugia = 1
| | | E_Degenerativas = 0: Alta_Mejoria
| | | E_Degenerativas = 1: Finado
| servicio = Pediatria: null
| servicio = Unidad_de_cuidados_intensivos: Finado

```

=== **Resumen de evaluación** ===

Instancias clasificadas correctamente	220	83.6818 %
Instancias clasificadas incorrectamente	35	13.2576 %

==== Detalle de la exactitud por clase ====

TP Tasa FP tasa precisión alcance F medida Clase

```
0.946  0.719  0.902  0.946  0.923  Alta_Mejoria
0.281  0.054  0.429  0.281  0.34  Finado
```

==== matriz de confusión ====

```
a  b  <-- classified as
211 12 | a = Alta_Mejoria
23  9 | b = Finado
```

## Árbol de regresión

REPTree

```
sitio_infeccion1 = Sistema_respiratorio
| E_Degenerativas = 0 : Alta_Mejoria (18/2) [6/1]
| E_Degenerativas = 1
| | Cirugia = 0 : Finado (4/1) [2/0]
| | Cirugia = 1 : Alta_Mejoria (10/3) [2/1]
sitio_infeccion1 = Sistema_cardiovascular : Alta_Mejoria (1/0) [0/0]
sitio_infeccion1 = Sistema_digestivo : Alta_Mejoria (0/0) [0/0]
sitio_infeccion1 = Sistema_geniturinario : Alta_Mejoria (23/2) [11/3]
sitio_infeccion1 = Sistema_nervioso_central : Alta_Mejoria (0/0) [0/0]
sitio_infeccion1 = Organos_y_tejidos : Alta_Mejoria (110/7) [62/3]
sitio_infeccion1 = Hematico : Alta_Mejoria (0/0) [1/0]
sitio_infeccion1 = Otras_infecciones : Finado (10/5) [4/3]
```

==== Resumen de evaluación ====

Instancias clasificadas correctamente	227	85.9848 %
Instancias clasificadas incorrectamente	37	14.0152 %

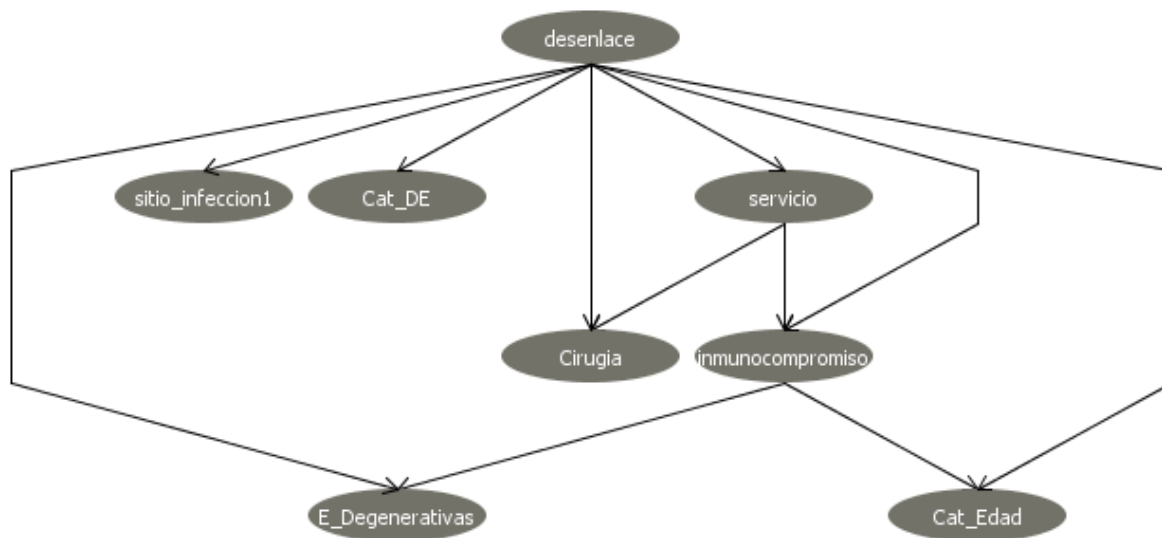
==== Detalle de la exactitud por clase ====

TP	Tasa	FP	tasa	precisión	alcance	F	medida	Clase
0.978	0.97	0.876	0.978	0.924	Alta_Mejoria			
0.03	0.022	0.167	0.03	0.051	Finado			

==== matriz de confusión ====

```
a  b  <-- classified as
226  5 | a = Alta_Mejoria
32  1 | b = Finado
```

## Red Bayesiana



*Figura 65. Red Bayesiana*

### === Resumen de evaluación ===

Instancias clasificadas correctamente	223	84.4697 %
Instancias clasificadas incorrectamente	41	15.5303 %

### === Detalle de la exactitud por clase ===

TP	Tasa	FP	tasa	precisión	alcance	F	medida	Clase
0.944	0.848	0.886	0.944	0.914	Alta_Mejoria			
0.152	0.056	0.278	0.152	0.196	Finado			

### === matriz de confusión ===

```

a b <-- classified as
218 13 | a = Alta_Mejoria
28 5 | b = Finado
  
```

## Red neuronal

### === Resumen de evaluación ===



Instancias clasificadas correctamente	227	85.9848 %
Instancias clasificadas incorrectamente	37	14.0152 %

=== Detalle de la exactitud por clase ===

TP Tasa	FP tasa	precisión	alcance	F medida	Clase
0.948	0.758	0.898	0.948	0.922	Alta_Mejoria
0.242	0.052	0.4	0.242	0.302	Finado

=== matriz de confusión ===

```

a b <-- classified as
219 12 | a = Alta_Mejoria
25 8 | b = Finado

```

## Maquina de soporte vectorial

v=== Resumen de evaluación ===

Instancias clasificadas correctamente	230	87.1212 %
Instancias clasificadas incorrectamente	34	12.8788 %

=== Detalle de la exactitud por clase ===

TP Tasa	FP tasa	precisión	alcance	F medida	Clase
0.996	1	0.875	0.996	0.931	Alta_Mejoria
0	0.004	0	0	0	Finado

=== matriz de confusión ===

```

a b <-- classified as
230 1 | a = Alta_Mejoria
33 0 | b = Finado

```

## Reglas de asociación

Tertius

1. /\* 0.524453 0.109848 \*/ inmunocompromiso = 1 ==> E\_Degenerativas = 1 or Cat\_Edad = adulto\_mayor or desenlace = Finado

2. /\* 0.519864 0.060606 \*/ inmunocompromiso = 1 ==> servicio = Medicina\_Interna or E\_Degenerativas = 1 or Cat\_Edad = adulto\_mayor
3. /\* 0.503948 0.128788 \*/ inmunocompromiso = 1 ==> E\_Degenerativas = 1 or Cat\_Edad = adulto\_mayor
4. /\* 0.498335 0.113636 \*/ inmunocompromiso = 1 ==> sitio\_infeccion1 = Sistema\_geniturinario or E\_Degenerativas = 1 or Cat\_Edad = adulto\_mayor
5. /\* 0.494042 0.064394 \*/ inmunocompromiso = 1 ==> Cirugia = 0 or E\_Degenerativas = 1 or Cat\_Edad = adulto\_mayor
6. /\* 0.481467 0.090909 \*/ inmunocompromiso = 1 ==> servicio = Medicina\_Interna or Cat\_Edad = adulto\_mayor or desenlace = Finado
7. /\* 0.476905 0.094697 \*/ inmunocompromiso = 1 ==> servicio = Medicina\_Interna or sitio\_infeccion1 = Hematico or Cat\_Edad = adulto\_mayor
8. /\* 0.475430 0.136364 \*/ Cirugia = 1 and desenlace = Alta\_Mejoria ==> servicio = Cirugia\_y\_Trauma or inmunocompromiso = 0
9. /\* 0.474290 0.106061 \*/ inmunocompromiso = 1 ==> sitio\_infeccion1 = Sistema\_respiratorio or E\_Degenerativas = 1 or Cat\_Edad = adulto\_mayor
10. /\* 0.472389 0.098485 \*/ inmunocompromiso = 1 ==> servicio = Medicina\_Interna or Cat\_Edad = adulto\_mayor

Número de hipótesis consideradas 37200

Número de hipótesis exploradas 18785

## Método de los k con generalización

### ==== Resumen de evaluación ====

Instancias clasificadas correctamente	213	80.6818 %
Instancias clasificadas incorrectamente	51	19.3182 %

### ==== Detalle de la exactitud por clase ====

TP Tasa	FP tasa	precisión	alcance	F medida	Clase
0.87	0.636	0.905	0.87	0.887	Alta_Mejoria
0.364	0.13	0.286	0.364	0.32	Finado

### ==== matriz de confusión ====

```

a  b  <-- classified as
201 30 | a = Alta_Mejoria
21  12 | b = Finado

```

## Árbol de decisión para la detección de neumonía

J48 unpruned tree

```

-----
sitio_infeccion1 = Sistema_respiratorio
| servicio = Cirugia_y_Trauma: 1 (15.0)
| servicio = Ginecologia_y_obstetricia: 0 (1.0)
| servicio = Medicina_Interna: 1 (25.0/2.0)
| servicio = Pediatria: 0 (1.0)
| servicio = Unidad_de_cuidados_intensivos: 1 (0.0)
sitio_infeccion1 = Sistema_cardiovascular: 0 (1.0)
sitio_infeccion1 = Sistema_digestivo: 0 (0.0)
sitio_infeccion1 = Sistema_genitourinario
| Cat_Edad2 = Joven: 0 (0.0)
| Cat_Edad2 = Adulto: 0 (6.0)
| Cat_Edad2 = adulto_mayor
| | Cat_DE = Baja_estancia: 1 (1.0)
| | Cat_DE = Estancia_media: 0 (2.0)
| | Cat_DE = Sobreestancia
| | | E_Degenerativas = 0: 0 (3.0)
| | | E_Degenerativas = 1: 1 (2.0)
| | Cat_Edad2 = 3_edad
| | inmunocompromiso = 0: 0 (4.0)
| | inmunocompromiso = 1
| | | servicio = Cirugia_y_Trauma: 0 (8.0/1.0)
| | | servicio = Ginecologia_y_obstetricia: 0 (0.0)
| | | servicio = Medicina_Interna
| | | | Cat_DE = Baja_estancia: 0 (0.0)
| | | | Cat_DE = Estancia_media: 0 (5.0/1.0)
| | | | Cat_DE = Sobreestancia: 1 (3.0/1.0)
| | | servicio = Pediatria: 0 (0.0)
| | | servicio = Unidad_de_cuidados_intensivos: 0 (0.0)
sitio_infeccion1 = Sistema_nervioso_central: 0 (0.0)
sitio_infeccion1 = Organos_y_tejidos: 0 (172.0/4.0)
sitio_infeccion1 = Hematico: 0 (1.0)
sitio_infeccion1 = Otras_infecciones: 0 (14.0)

```

Número de hojas: 25

Tamaño del árbol: 33

Instancias clasificadas correctamente	245	92.803 %
Instancias clasificadas incorrectamente	19	7.197 %

==== Detalle de la exactitud por clase ====

TP Tasa FP tasa precisión alcance F medida Clase  
 0.963 0.224 0.95 0.963 0.956 0  
 0.776 0.037 0.826 0.776 0.8 1

==== matriz de confusión ====

a b <-- classified as  
 207 8 | a = 0  
 11 38 | b = 1

### Clasificador Naive Bayes para la detección de neumonía

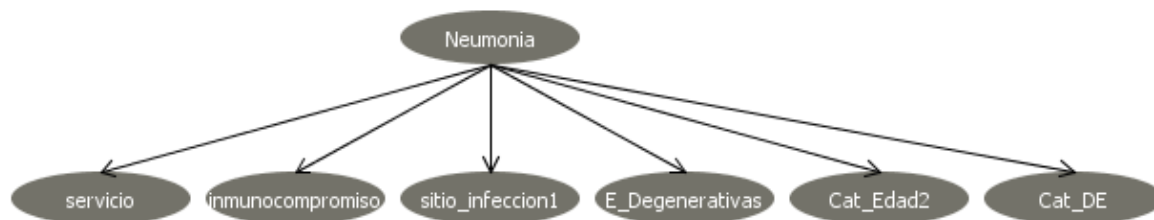


Figura 66. Clasificador naive bayes para la detección de Neumonía

Tabla XL. Distribución de probabilidad condicional del atributo neumonía

Neumonía		
	No	SI
	0.8132075471698114	0.18679245283018867

Tabla XLI. Distribución de probabilidad condicional del atributo enfermedades degenerativas.

Enfermedades Degenerativas		
	No	Si
No	0.59027778	0.40972222
SI	0.53	0.47

Tabla XLII. Distribución de probabilidad condicional del atributo inmunocompromiso.

Inmunocompromiso	
------------------	--

	No	Si
No	0.29525862	0.70474138
SI	0.17	0.83

**Tabla XLIII. Distribución de probabilidad condicional del atributo servicio.**

Servicio					
	C.T	G.O	M.I	Pediatría	UCI
No	0.44367816	0.03908046	0.46206897	0.04367816	0.01149425
SI	0.39805825	0.00970874	0.57281553	0.00970874	0.00970874

**Tabla XLIV. Distribución de probabilidad condicional del atributo sitio de infección.**

Sitio de infección							
	SR	SC	SD	SG	SNC	SG	SOT
No	0.0205479	0.0068493	0.002283	0.1255707	0.0022831	0.7694063	0.0068493
SI	0.72641509	0.0094339	0.009433	0.1415094	0.0094339	0.0849056	0.0094339

**Tabla XLV. Distribución de probabilidad condicional del atributo DE\_2.**

DE			
	Estancia baja	Estancia media	Sobre estancia
No	0.13625866050	0.3071593533487	0.5565819861
SI	0.04950495049	0.1683168316831	0.7821782178

**Tabla XLVI. Distribución de probabilidad condicional del atributo edad.**

Edad_2				
	Joven	Adulto	Adulto Mayor	
No	0.14516129032	0.2096774193548	0.3801843317	0.264976958525
SI	0.00980392156	0.1862745098039	0.3431372549	0.460784313725

### === Resumen de evaluación ===

Instancias clasificadas correctamente	248	93.9394 %
Instancias clasificadas incorrectamente	16	6.0606 %

### === Detalle de la exactitud por clase ===

TP	Tasa	FP	tasa	precisión	alcance	F	medida	Clase
0.981	0.245	0.946	0.981	0.963	No	=0		

0.755 0.019 0.902 0.755 0.822 SI=1

==== matriz de confusión ====

```
a b <-- clase
211 4 | a = No=0
12 37 | b = SI=1
```

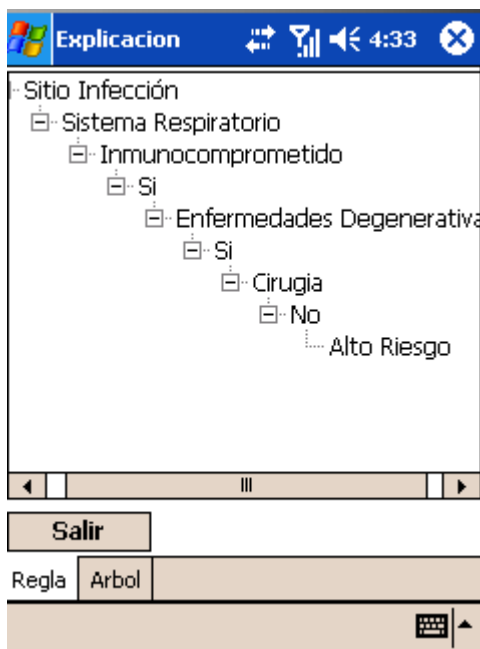
## Apéndice D. Actividades de la evaluación

### Introducción

Uno de los objetivos de este trabajo es obtener modelos predictivos a partir de técnicas de estadística avanzada, conocidas como minería de datos, estas técnicas tienen la capacidad de aprender en base a una memoria organizacional la cual contiene toda experiencia y conocimiento de las decisiones realizadas en el pasado, que en nuestro caso es la forma de registro caso de infección nosocomial.

Entonces se utilizó la forma de registro de caso de infección nosocomial para aplicar el proceso de extracción de conocimiento con el objetivo de obtener modelos predictivos. Podemos clasificar los modelos predictivos en dos: simbólicos y sub-simbólicos, los modelos predictivos simbólicos son aquellos que pueden brindar información de cómo la decisión fue tomada en forma de reglas o en un árbol de decisión.

En la *Figura 67* se muestra un ejemplo de una regla de decisión, la cual se interpreta de la siguiente forma: si el paciente tiene una infección en el sistema respiratorio, es inmunocomprometido, presenta alguna de las enfermedades degenerativas (cáncer, diabetes mellitus, hipertensión arterial sistemática), y no tiene cirugía entonces el paciente tiene un alto riesgo de contraer una infección nosocomial.



*Figura 67. Ejemplo de una regla en la PDA*

Los modelos predictivos sub-simbólicos son aquellos que son una “caja negra” los cuales no pueden brindar un razonamiento un ejemplo de esto son los modelos basado en probabilidad como lo son los clasificadores naive bayes.

La secuencia de actividades que serán llevadas a cabo durante el experimento son las siguientes:

**Escenario 1 ingresar información del paciente.** Este escenario es cuando: “se ingresa la información del censo en el sistema FRCIN para posteriormente sincronizar con el dispositivo móvil y adquirir todo los datos de los pacientes que se encuentren dentro del hospital previamente ingresados en el sistema FRCIN”

Paso 1. En la pantalla inicial del *sistema FRCIN* presione el botón *nuevo registro* como se muestra en la *Figura 68*.



## Formato de Registro

En Caso de Infección Nosocomial

Nuevo Registro

Informe

Num. de Afiliación:  Nombre/Apellidos:  Descenlace:

Núm. de Afiliación	Nombres	Apellidos	Edad	Servicio	D.E.	D.S.E.H.	Valoración	Opciones
0179530443	Ruth	Alvarado Cendijas	11 años	Cirugía y Trauma	5	2	1%	<input type="button" value="Abrir"/>
12344568	Alejandra	Zedillo Lopez	33 años	Medicina Interna	63	0	33%	<input type="button" value="Abrir"/>
123456	Javier	Hernandez Hernandez	63 años	Medicina Interna	54	0	28%	<input type="button" value="Abrir"/>
12345678	Javier	Hernandez Anguiano	63 años	Medicina Interna	24	0	56%	<input type="button" value="Abrir"/>
12345678	Javier	Hernandez Anguiano	63 años	Medicina Interna	24	0	56%	<input type="button" value="Abrir"/>
25917844264f50or	Alejandra	Gomez Dueñas	56 años	Medicina Interna	15	11	56%	<input type="button" value="Abrir"/>

Nuevo Registro

Figura 68. Pantalla inicial del sistema FRCIN

En la Figura 69 se muestra la pantalla en la cual se introducirán los datos del paciente.

## Formato de Registro

En Caso de Infección Nosocomial

Información Básica   Factores de Riesgo   Sitios de Infección   Antecedentes Quirúrgicos   Información Microbiológica   Tratamiento   Observaciones

Núm de afiliación:

Nombre(s):

Apellido paterno:

Apellido materno:

Sexo:

Edad:  año

Con inmunocompromiso:

Servicio:

Motivo de ingreso:

Cama:

Fecha de ingreso:  ...

Fecha de detección:  ...

Fecha de resolución:  ...

Fecha de reingreso:  ...

Guardar y Salir   Cancelar

Figura 69. Para dar de alta un paciente con sus respectivos datos.

Paso 2. Ingresar los datos de la Tabla XLVII

**Tabla XLVII. Información del paciente**

Campo:	Valor:
Núm. De afiliación	12345678
Nombre	Javier
Apellido paterno	Hernandez
Apellido materno	Anguiano
Sexo	Masculino
Edad	63 años
Inmunocomprometido	Si
Servicio	Medicina interna
Motivo de ingreso	
Cama	123
Fecha de ingreso	12/09/2008
Fecha de detección	
Fecha de resolución	
Fecha de reingreso	

Paso 3. Presionar el botón de “*Guardar y Salir*” d

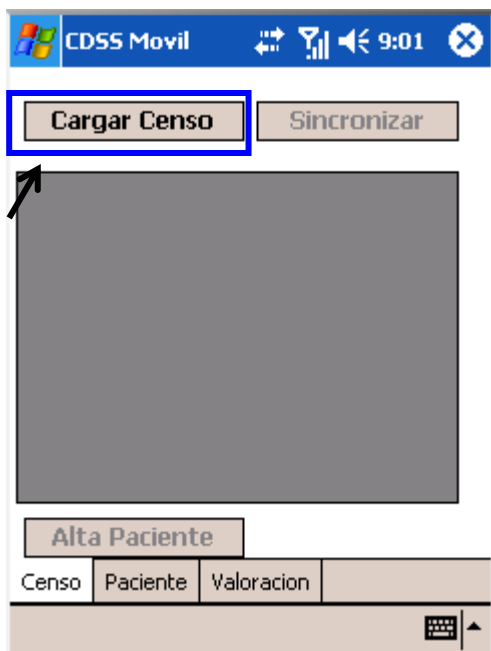
De acuerdo con la fecha ingreso, el paciente tiene 17 días de estancia, con los datos de la Tabla XLVII, usted consideraría que este paciente tiene un riesgo de contraer una Infección nosocomial.

---



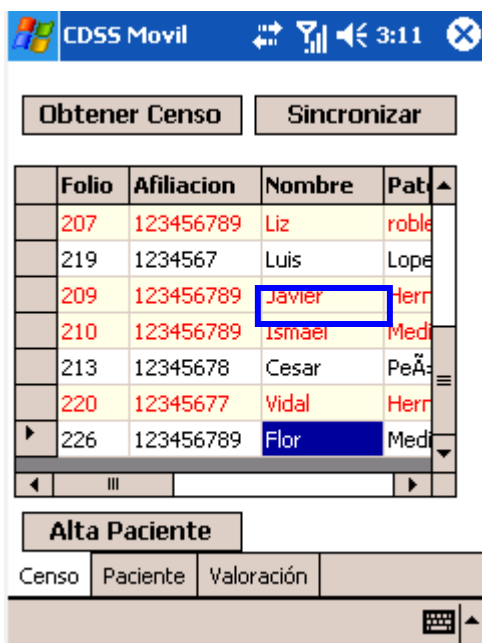
---

Paso 4. Con la pluma de la PDA y el cable conectado, presione el botón de *cargar censo*



*Figura 70. Pantalla inicial del CDSS Móvil*

El cual desplegará el censo de pacientes, como se ve en la siguiente *Figura 70*, cuando este el censo desplegado como se ve en la figura, puede desconectar el cable.



*Figura 71. Censo de pacientes*

En la *Figura 71* Se muestra los pacientes que se encuentran dentro del hospital, los pacientes en rojo, son pacientes que tienen un alto riesgo de contraer una infección nosocomial, entonces el paciente Javier, se considera un paciente con alto riesgo, si se desea los demás elementos de información del censo, solo se tiene que mover la barra horizontal con la pluma.

**Paso 4.** Buscar el paciente Javier y seleccionarlo con la pluma

**Paso 5.** Para poder ingresar o visualizar la información de forma individual del paciente se tiene que tiene que presionar la pestaña de *Paciente*, el cual mostrara la información la como se ve en la *Figura 72*.

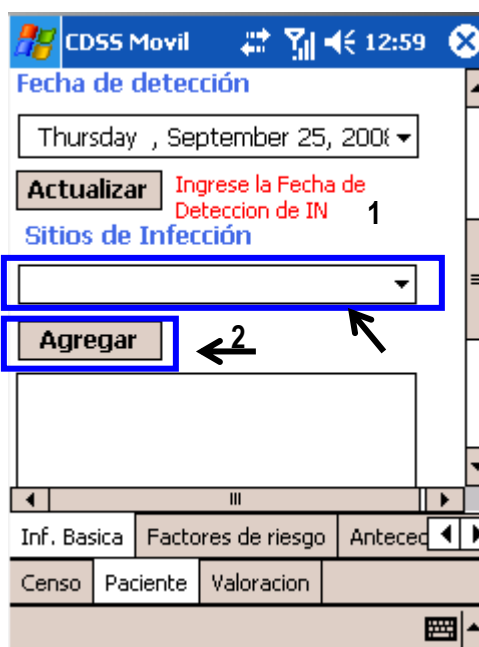
The screenshot shows a mobile application interface for 'CDSS Movil'. At the top, there is a status bar with the time 6:46 and various icons. Below the status bar, the text 'Información Basica:' is displayed in blue. Underneath, there are four input fields: 'Cama' with the value '123', 'Edad' with '63', 'Servicio' with 'Medicina Interna', and 'DE' with '17'. Below these fields, the text 'Paciente' is displayed in blue, followed by a text input field containing 'Javier Hernandez'. Underneath that, 'No. Afiliación' is displayed in blue, followed by a text input field containing '123456'. Below the input fields, there is a section titled 'Fecha de detección' in blue, which is currently obscured by a horizontal scrollbar. At the bottom of the screen, there is a navigation bar with several tabs: 'Inf. Basica', 'Factores de riesgo', 'Anteced', 'Censo', 'Paciente', and 'Valoración'. The 'Paciente' tab is currently selected.

*Figura 72. Información básica del paciente*

Posteriormente para poder visualizar la demás información básica del paciente se tiene que deslizar la barra verticalmente para poder ver la fecha de detección y los sitios de infección que tiene el paciente, en caso de que el paciente no tenga una fecha de detección definida aparece el mensaje en rojo “*ingrese la Fecha de detección de IN*”.

**Escenario 2** ingreso de información del paciente en el dispositivo móvil, este escenario ocurre cuando “el médico realiza el recorrido por los pasillos revisando los síntomas del paciente y el expediente médico”.

Paso 1. Ingresar el sitio de infección del paciente, presionando el cuadro de control el cual desplegara una lista y seleccione la opción de “*Infección en el tracto respiratorio*” y después presione el botón de *Agregar*.



*Figura 73. Sitio de infección del paciente*

Paso 2. Utilice la barra vertical de desplazamiento para recorrer la pantalla, hasta que llegue a ver como se muestra en la *Figura 73*. Presione el botón de “teclado” el cual se encuentra ubicado en la parte inferior derecha de la pantalla. En el cuadro de texto agregue el diagnóstico el siguiente diagnóstico: “*hipertrofia prostática*” y con la pluma presione el botón de “*Agregar*”

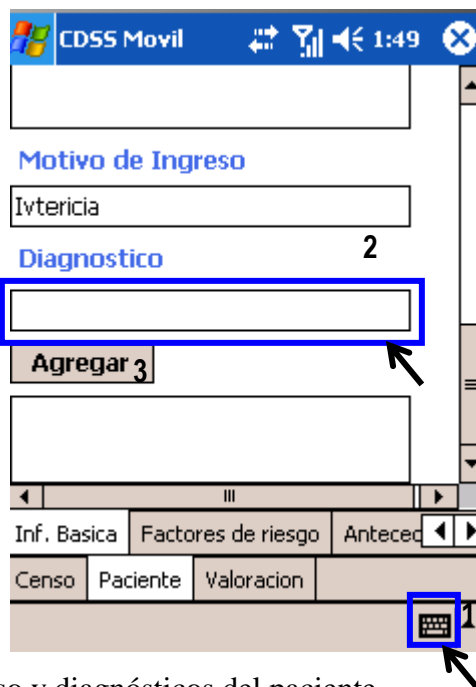


Figura 7 motivo de ingreso y diagnósticos del paciente

Paso 4. Vuela a presionar el botón de *teclado*

**Escenario 3.** Este escenario consiste en el ingreso de los factores de riesgo del paciente los cuales se determinaron con las condiciones actuales del paciente. Los factores de riesgo asignados fueron los siguientes: “*Sobre estancia hospitalaria*”, “*Venoclisis*”, “*Antibioticoterapia*”, “*Edad*”, “*Hipertensión arterial sistemática*”.

Paso 1. Para agregar estos factores de riesgo al sistema, presione con la pluma en el tabulador donde dice “*Factores riesgo*”.

Paso 2. Posteriormente presione sobre el cuadro para que se despliegue la lista de los factores de riesgo disponibles, seleccione los factores de riesgo que tiene el paciente.

Paso 3 Presione el botón de “*Agregar*” y repita los pasos hasta agregar todos los factores de riesgo.

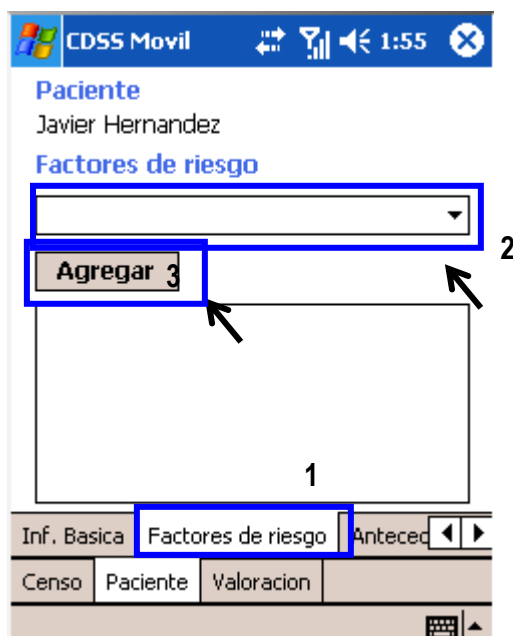


Figura 74. Factores de riesgo del paciente

Entonces se tiene la siguiente información del paciente Javier, esto se puede ver en la Tabla XLVIII.

Tabla XLVIII. Datos clínicos del paciente Javier

Campo:	Valor:
Nombre	Javier
Apellido paterno	Hernandez
Apellido materno	Anguiano
Sexo	Masculino
Edad	63 años
Inmunocomprometido	Si
Servicio	Medicina interna
Días estancia	17
Sitio de infección	<i>Infección en el tracto respiratorio</i>
Diagnostico	hipertrofia prostática
Factores de riesgo	<i>“Sobre estancia hospitalaria”, “Venoclisis”, “Antibioticoterapia”, “Edad”, “Hipertensión arterial sistemática”.</i>

De acuerdo con el estado actual de paciente, usted considera paciente tiene un alto riesgo de contraer una infección nosocomial o tiene un bajo riesgo de contraer una infección?

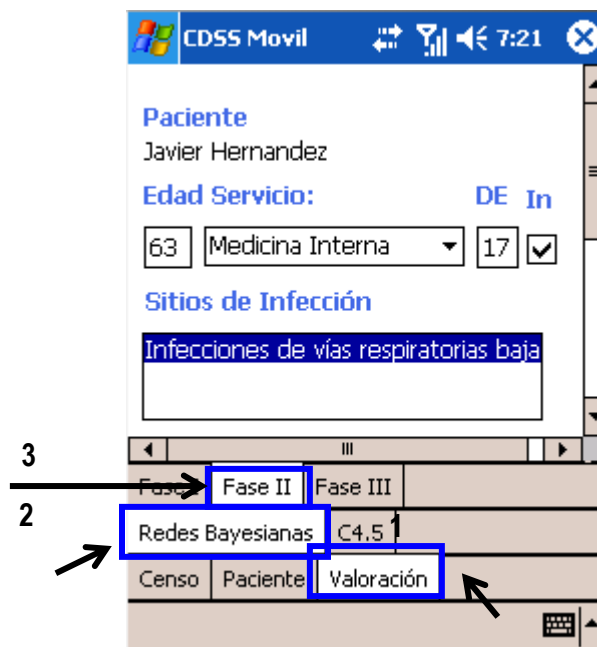
---



---

**Escenario 4.** En este escenario es cuando valoramos si un paciente tiene algún riesgo de contraer una infección nosocomial.

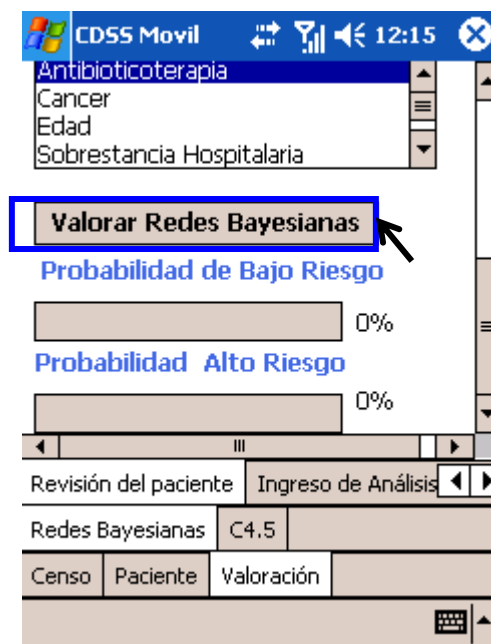
Paso 1. Con la pluma presione la opción de “*Valoración*” y posteriormente presione en la *Fase II*, lo cual desplegará una pantalla como en *Figura 75*. En esta pantalla se desplegará la información necesaria para realizar la valoración del paciente.



*Figura 75. valoración con Naive Bayes*

**Paso 2.** Con la pluma deslice la barra vertical, hasta que se vea como en la *Figura 76*



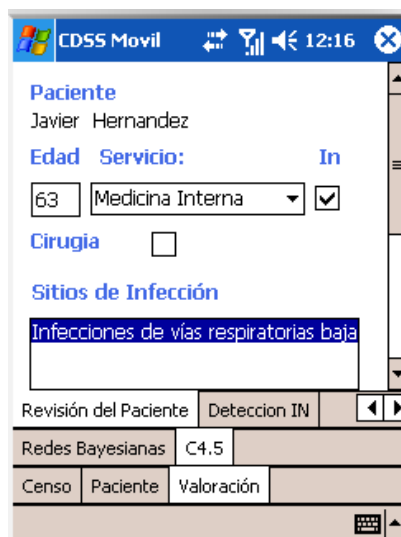


*Figura 76. Valoración con redes bayesianas*

**Paso 3.** Presione con la pluma el botón de *Valorar Redes Bayesianas* y observe el resultado.

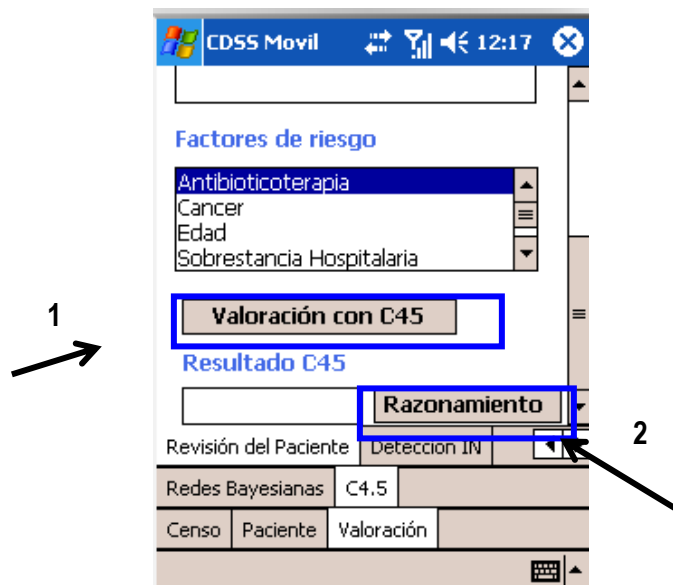
**Escenario 5.** Esta valoración consiste en la evaluación del mismo paciente, pero con diferente modelo predictivo. Para acceder se tiene que hacer lo siguiente

**Paso 1.** Presionar sobre el tabulador que dice *C 4.5* y después sobre la “*revisión del paciente*”. El cual desplegará la imagen como se muestra en la *Figura 77*.



*Figura 77. Valoración con árboles de decisión*

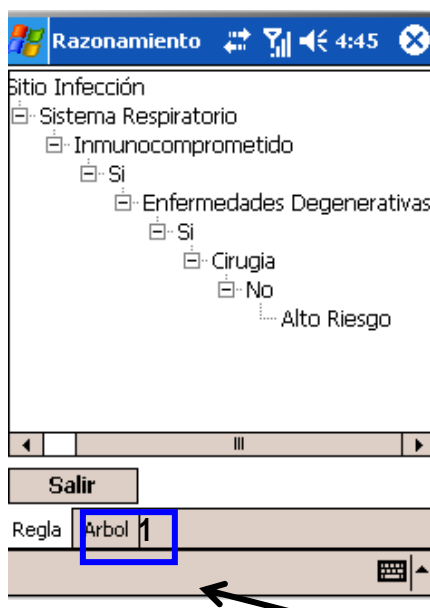
**Paso 2.** Deslizar la barra vertical con la pluma hasta llegar a visualizar el botón de *Valoración C45*. Esto se puede ver en la figura 12.



*Figura 78. Valoración con árboles de decisión*

**Paso 3.** Presione el botón de *valoración C45* y observe el resultado.

**Paso 4.** Presione el botón de razonamiento para poder visualizar la regla y árbol utilizado para llegar a tomar la decisión, en la figura 13 muestra la regla utilizada.



*Figura 79. Regla utilizada para la decisión*

**Paso 5.** Presione en el botón árbol para que muestre el árbol de decisión.

**Escenario 6.** En este escenario suponemos que se realiza el recorrido por el hospital y llegamos hasta el paciente "Paco", el cual se puede ver en el censo de pacientes y en el expediente médico observamos que el paciente tiene una cirugía no registrada.

**Paso 1.** Presione con la pluma en donde dice "Censo" y posteriormente presione sobre "Paco" y por ultimo presione donde dice "Paciente". Como se muestra en la *Figura 80*



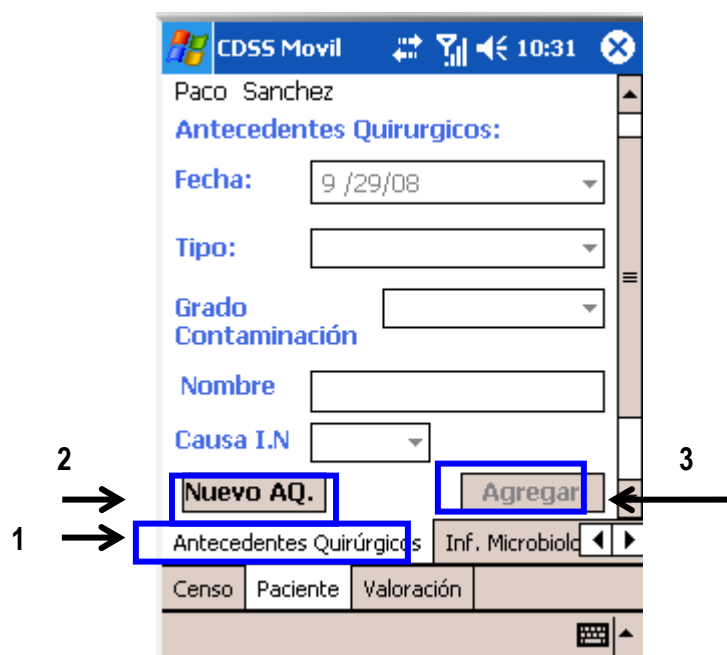
*Figura 80. Censo de pacientes*

Paso 2. Con la pluma presionamos en "Antecedentes Quirúrgicos", posteriormente deslizamos la barra vertical hasta que pueda observar el botón de "Nuevo AQ."

Paso 3 presione con la pluma el botón de "Nuevo AQ."

*Tabla XLIX. Antecedentes quirúrgicos*

Campo:	Valor:
Fecha	9/09/08
Tipo	Electiva
Contaminación	Limpia
Nombre	TAI MPD
Causa I.N	No



*Figura 81. Antecedentes Quirúrgicos*

**Paso 4.** Ingrese la información de la *Tabla XLIX*, en los respectivos campos de la pantalla 15. y presione el botón de “Agregar”

**Paso 5** presione “Valoración” posteriormente presione “C45” y por ultimo presione “Revisión del paciente”. En donde visualizara la *Figura 82*.

CDSS Movil 12:26

**Paciente**  
Paco Sanchez

**Edad** **Servicio:** **In**

68 Cirugía y Trauma

**Cirugia**

**Sitios de Infección**

Infección de piel y tejidos blandos

Revisión del Paciente Deteccion IN

Redes Bayesianas C4.5

Censo Paciente Valoración

Figura 82. Valoración con árboles de decisión

**Paso 6.** Con la pluma deslice la barra vertical de desplazamiento hasta que llegue a visualizar

CDSS Movil 12:30

**Factores de riesgo**

Cateter Venoso Central  
Sobrestancia Hospitalaria  
Antibioticoterapia  
Venoclisis

**Valoración con C45**

**Resultado C45**

Razonamiento

Revisión del Paciente Deteccion IN

Redes Bayesianas C4.5

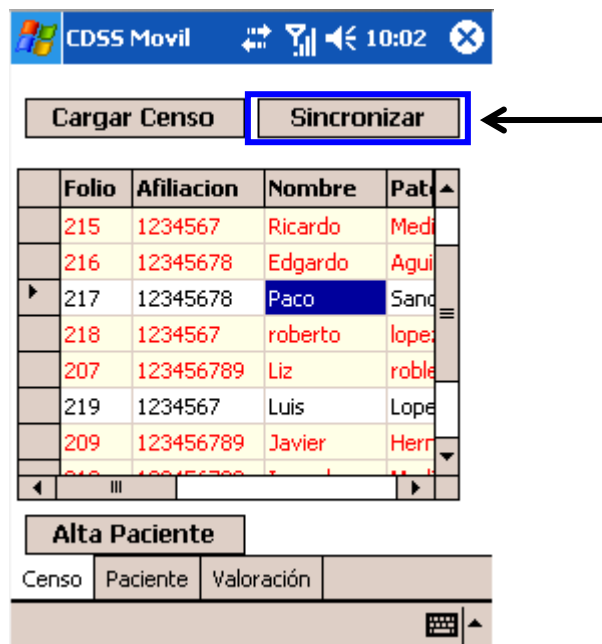
Censo Paciente Valoración

Figura 83. Valoración con árboles de decisión

**Paso 7** presione el botón de “valoración C45” y posteriormente el de razonamiento.

**Escenario 7.**este escenario ocurre cuando se desea actualizar el sistema todos cambios del paciente en el Sistema FRCIN.

**Paso 1.** Con la dispositivo móvil con el cable conectado, presione el botón de sincronizar, como se ver en la *Figura 84*.



*Figura 84. Actualizar los cambios en el sistema FRCIN*

**Paso 2.** En la pantalla inicial de sistema FRCIN, en el campo de Nombre/apellido escriba el nombre de “Paco” y posteriormente presione el botón de “Filtrar”

**Paso 3.** En la columna de opciones presione el botón de “Abrir”

## Formato de Registro

En Caso de Infección Nosocomial

1

2

3

Núm. de Afiliación	Nombres	Apellidos	Edad	Sexo	D.E.	D.S.E.H.	Opciones
12345678	Paco	Sanchez Marron	68 año	M	28	0	Abrir

*Figura 85. Resultado de la búsqueda de “Paco”*

**Paso 4.** Presione la pestaña de “*Antecedentes Quirúrgicos*” y visualice la información capturada previamente en el sistema móvil.

## Formato de Registro

En Caso de Infección Nosocomial

Información Básica | Factores de Riesgo | Sitios de Infección | **Antecedentes Quirúrgicos** | Información Microbiológica | Tratamiento | Observaciones

29/09/2008 | Electiva | Limpia | TAI MPD | No es causa (x)

Fecha:  ...

Tipo:

Grado de contaminación:

Nombre:

¿Causa de una IN?:

Agregar Cirugía

Guardar y Salir | Cancelar

*Figura 86. Antecedentes quirúrgicos*

## Cuestionario

1. La información presentada en esta demostración es suficiente para permitirme entender cómo funcionaría de esta herramienta, en mi trabajo cotidiano dentro del hospital.
2. Asumiendo que tuviera acceso a la tecnología de este CDSS, yo la utilizaría para:

3. ¿apoyarme con el manejo del censo de pacientes?
4. ¿tener acceso a información del paciente?
5. ¿Considera que el CDSS Móvil es sencillo de usarse?
6. ¿Considera que el CDSS Móvil es fácil consultar el censo de pacientes?
7. ¿Considera que el CDSS Móvil podría servir como entrenamiento de la Unidad de vigilancia epidemiológica?
8. ¿Considera que el Sistema FRCIN es sencillo de usarse?
9. ¿Considera que el Sistema FRCIN es fácil consultar el censo de pacientes?
10. ¿Considera que el Sistema FRCIN podría servir como entrenamiento de la Unidad de vigilancia epidemiológica?
11. ¿Cuál de las etapas del CDSS móvil considera que es más útil?
12. ¿Cuál de los modelos considera más útil que tienen más utilidad, los basados en probabilidad (Redes bayesianas) o los basados en arboles (C 45)?
13. ¿Por qué?
14. ¿Las reglas mostradas por parte de del árbol C45, se parecen al razonamiento clínico que usted utiliza?
15. ¿Considera que el CDSS móvil podría ser útil para la detección y vigilancia epidemiológica?, ¿por qué?