Universidad Autónoma de Baja California
Facultad de Ciencias Químicas e Ingeniería
Maestría y Doctorado en Ciencias e Ingeniería

Análisis e Implementación de Algoritmos de Inteligencia Computacional dentro de un Marco de Trabajo para la Toma de Decisiones Médicas Asistidas por Computadora

TESIS

Para Obtener el Grado de

Maestro en Ciencias

Presenta:

José Antonio Martínez Maldonado

Bajo la dirección de:

Dr. Juan Ramón Castro Rodríguez

Co-dirigido por:

Dr. Manuel Castañón Puga

Tijuana, Baja California, México

Marzo del 2017

A las personas que realmente confiaron en mi, esas personas a las que puedo llamar amigos, a mi incansable familia que siempre me apoyo en este paso importante de mi vida y a mi compañera de vida.

Universidad Autónoma de Baja California

FACULTAD DE CIENCIAS QUÍMICAS E INGENIERÍA

COORDINACIÓN DE POSGRADO E INVESTIGACIÓN

FOLIO No. 203

Tijuana, B. C., a 3 de marzo de 2017

C. José Antonio Martínez Maldonado
Pasante de: Maestro en Ciencias
Presente

El tema de trabajo y/o tesis para su examen profesional, en la Opción TESIS

Es propuesto, por los CC. Dres. Juan Ramón Castro Rodríguez y Manuel Castañón Puga

Quienes serán los responsables de la calidad del trabajo que usted presente, referido al tema: Análisis e Implementación de Algoritmos de Inteligencia Computacional dentro de un Marco de Trabajo para la Toma de Decisiones Médicas Asistidas por Computadora.

el cual deberá usted desarrollar, de acuerdo con el siguiente orden:

I.- INTRODUCCIÓN
II.- MARCO TEÓRICO
III.- PROPUESTA DE LOS ALGORITMOS PARA LA BIBLIOTECA DE CLASES DMCL
IV.- IMPLEMENTACIÓN DE LA BIBLIOTECA DE CLASES DCML
V.- PRUEBAS Y RESULTADOS
VI.- CONCLUSIONES Y TRABAJO FUTURO

UNIVERSIDAD AUTONOMA DE BAJA CALIFORNIA

FACULTAD DE CIENCIAS QUIMICAS E INGENIERIA

Dr. Juan Ramón Castro Rodríguez
Director de Tesis

Dr. José Luis González Vázquez
Secretario

Dr. Manuel Castañón Puga
Co-Director de tesis

Dr. Luis Enrique Palafox Maestre
Director

Agradecimientos

Deseo agradecer a

CONACYT (Consejo Naconal de Ciencia y Tecnología) por financiar mis estudios de Maetría en Ciencias e Ingeniería durante los dos años de duración de la misma. A la Universidad Autonoma de Baja California, por permitirme ser parte de la institución, brindarme las herramientas necesarias para mi propio desempeño y por apoyarme a cumplir mis metas. Agradezco profundamente al cuerpo docente de dicha institución por siempre brindarme su apoyo, aclarar mis dudas y estar siempre al pendiente de mi progreso durante el periodo académico. Agradezco profundamente a mi director de tesis el Dr. Juan Ramon Castro Rodriguez por su apoyo incondicional, su confianza, dedicación y entrega. Gracias por siempre estar disponible para resolver dudas y brindarme su tiempo para resolver cualquier problema. A mi co-director de tesis Dr. Manual Castañon Puga por brindarme su tiempo, confianza y apoyarme durante mi periodo académico. A mis sinodales: Dra. Carelia G Gaxiola Pacheco y Dra. Olivia Mendoza Duarte, gracias por la oportunidad, por el apoyo recibido durante la realización de esta tesis y por el tiempo que se han tomado para leer este documento. A mi madre, por ser un ejemplo en la vida a seguir, por darme su apoyo y estar siempre presente en todo momento. Gracias por entender mis pasos de vida, por brindarme tu ayuda cuando la necesité y por todo el amor que siempre me haz dado. A mi padre por brindarme su apoyo incondicional, por confiar en mí. Por enseñarme que siempre hay que buscarle el lado bueno

a la vida. A toda mi familia que de alguna u otra manera siempre han apoyado a mis padres y a mi a seguir adelante, siempre dando buenos consejos. A todos mis amigos por formar parte de mi vida, brindarme su cariño y confianza, por siempre darme animos para seguir adelante en mi camino. A esa persona que se ha torando especial para mi y me ha mostrado cosas importantes de la vida, me ha ensenãdo a descubrir las cosas que puedo lograr si me lo propongo. Muchas Gracias a todos. *José Antonio Martínez Maldonado*

Resumen

La medicina en la busqueda de mantener la salud del hombre ha buscado siempre la implementación de técnicas que faciliten el realizar su trabajo, hoy en día la inteligencia computacional ha sido utilizada dentro de varias areas como un poderoso arcenal de técnicas de procesamiento de datos obteniendo muy buenos resultados. Razón por la cual areas como la medicina han buscado el incoporarla dentro de ciertos procesos que realizan. Es aquí donde se pueden encontras los sistemas de soporte para la toma de decisiones los cuales ayudan a evaluar diferentes situaciones para al final mostrar posibles soluciones que pueden o no ser tomadas por el ser humano. Del analsis de algoritmos de inteligencia computacional surge Data Minning Class Library una implementación de algoritmos con la cual se busca realizar sistemas de soporte para la toma de decisiones. Algunos de los algoritmos encontrados dentro de la biblioteca son K-means, Gustafson-Kessel, Gath-Geva, Fuzzy C-means, también se cuenta con sistemas de inferencia difusa de mandanni y takagi-sugeno-kang, así como redes neuronales con propagación hacia atras y arboles de decisiones. Esta biblioteca de algoritmos esta orientada a ser usada por programadores como un recurso para desarrollar sistemas de toma de decisiones de una manera sencilla y como una herramienta para el procesamiento masivo de datos.

Abstract

The medicine in his constant effort of maintaining the health of the human searches for techniques that help him develop his work more easy. Todays computational intelligence has been use in different areas like a powerful group of data procesing techniques obtaining good results. Reason why medicine has been searching the way to incorporate it in different process in order to improve them. Decision support systems come in handy in evaluating critical situations found in processes so they can show possible solutions that the user of them can decide if is goin to apply it or not. This is the reason why we started developing the Data Mining Class Library, as a solution to create decision support system fast and easy. Some of the algorithms we implemented in this library are K-means, Gustafson-Kessel, Gath-Geva, Fuzzy C-means, also we incorporated fuzzy inference systems of mandann and takagi-sugeno-kang, neural networks with backpropagation and decision trees. This library of algorithms is oriented to be used by programers as a value resource in developing decision support systems and as tool for massive data procesing.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Dentro del área de la medicina constantemente se busca la incorporación de nuevas técnicas o herramientas que auxilien a conservar la salud del ser humano, durante el paso del tiempo se han incorporado sistemas capaces de realizar mediciones con el objetivo de conocer características propias de un determinado paciente para con ello conocer que lo aqueja y de esta manera tratarlo; sin embargo el proceso mediante el cual se toma una decisión de cómo tratar a un paciente queda en base al conocimiento del médico, siendo esto riesgoso para el paciente debido al gran número de variables que se tienen que revisar para estimar un diagnóstico acertado. De igual manera existe un nivel de complejidad al momento de realizar una recomendación sobre que tratamiento llevar, esto debido a ciertas consideraciones tales como: la sensibilidad a sustancias, efectividad del tratamiento, en algunos casos cambios alimenticios. Consideraciones que en ocasiones necesitan adaptarse a nuevos descubrimientos que se puedan presentar en el paciente.

Los errores de diagnóstico médico son considerados la octava causa de muerte de un paciente, esto por diferentes motivos, entre ellos destacan: la omisión de algún dato importante que el paciente evita mencionar al médico, la interpretación errónea de los síntomas

del paciente, tratamiento incorrecto (muy relacionada con la anterior), desconocimiento de la enfermedad debido al surgimiento de nuevos brotes de enfermedades o evolución de las mismas ocasionando que sean más resistentes a métodos de tratamiento convencionales.

El constante avance de la tecnología ha permitido que el ser humano tenga un ritmo de vida más cómodo y simple, sin embargo dentro de la ciencia médica siguen siendo pocas las aplicaciones que se han tenido en las cuales se involucre al paciente como responsable directo, es verdad que se ha contribuido a la mejora de aparatos de medición y herramientas que auxilian al médico dentro del desempeño de su trabajo, sin embargo la relación entre médico y paciente queda solo dentro del consultorio, dejando a un lado cualquier tipo de recomendación que se le puede hacer y que ayude en su tratamiento, aunque es entendible que el médico no puede estar disponible a cada momento para el paciente. Debido a la complejidad que existe de modelar un sistema capaz de realizar las funciones de tratamiento de un médico son pocas las aportaciones de los mismos dentro del área médica.

La recopilación de información es un factor muy importante dentro de cuestiones medicas, ya que permiten tener las herramientas necesarias para la formulación de un diagnostico exacto, pero debido a que es una gran cantidad de información por cada paciente es complejo el estar consultando múltiples registros del paciente propiciando que este proceso sea tardado y lento, en ciertas ocasiones se tienen sistemas que permiten automatizar un poco el proceso de consulta más aun siguen siendo ineficientes aparte de que solo permiten extraer información ya almacenada del paciente, sin embargo no es capaz de reaccionar de manera inteligente realizando recomendaciones preventivas, sugerencias de planes de tratamiento y diagnósticos.

Estos factores han hecho que la Ciencia y Tecnología traten de incursionar en la medicina con aplicaciones básicas que ayuden a los médicos al tratamiento de enfermedades o simplemente como auxiliar en la obtención de parámetros para la realización de diagnósticos.

1.1. Antecedentes

1.1.1. Inteligencia Computacional

El concepto de inteligencia computacional (IC) surge con Alan Turín al cuestionarse si las computadoras podrían llegar a ser inteligentes, fue así como en 1950 publico su prueba de inteligencia computacional, la cual consistía en, una persona realizando una serie de preguntas por medio de un teclado a una persona y a una computadora; si el interrogador era incapaz de distinguir entre uno u otro, se dice que dicha computadora es inteligente [13].

El concepto de IC se basa fundamentalmente en entender el cómo funciona la mente humana y a su vez llevar esto a un ente no biológico, durante muchos años ha sido un tema que ha mantenido ocupado a científicos y filósofos [14]. Esto llevo a que se definiera a la IC como el "estudio de mecanismos adaptativos que permitan o faciliten comportamientos inteligentes en entornos complejos o cambiantes", y siendo una sub-rama de Inteligencia Artificial adopto varios de sus paradigmas, entre los cuales están: Redes neuronales Artificiales, Computo Evolutivo, Inteligencia Colectiva y Sistemas Difusos [13].

Dentro de las técnicas de IC existen dos tipos de modelos de aprendizaje que ayudan a la resolución de problemas, estos son aprendizaje supervisado y no supervisado; el aprendizaje supervisado es aquel en el cual al sistema se le proporcionan las instancias "x" y "y", donde "y" representa la variable que el sistema debe de estimar y "x" es un conjunto de valores que ayudaran al sistema a determinar "y". En aprendizaje no supervisado al sistema no se le proporciona la variable a estimar "y", solo en base a la variable "x" deberá ser capaz de determinar la salida estimada [11].

1.1.2. Redes Neuronales

Las redes neuronales están basadas en la idea de cómo el cerebro es capaz de procesar la información, por esta razón manejan factores que se presentan biológicamente dentro del proceso de aprendizaje en el cerebro humano. Dentro de una neurona humana se tienen varias partes que se pretende imitar dentro del proceso, una de estas partes son las dendritas las cuales contienen lo que se conoce como sinapsis las cuales permiten obtener información de otra neurona, una vez pasada la información a las dendritas esta recorre el cuerpo de la neurona hasta llegar al cuerpo de la célula, dentro realiza un proceso sumatorio de todas las señales de entrada, una vez hecho esto se pasa a una función de activación que al cumplir un cierto límite envía una señal a la siguiente neurona [10].

Una de las ventajas de utilizar redes neuronales en algún problema es la facilidad con que son capaces de reducir el rango de hipótesis del mismo [11], esto quiere decir que la solución obtenida es más genérica dando la pauta para que al cambiar o incrementarse los datos se pueda seguir aplicando el modelo generado.

1.1.3. Lógica Difusa

Dentro del conjunto de técnicas se tiene que mencionar la lógica difusa ya que nos permite realizar un modelado del conocimiento impreciso, manejar incertidumbres y sobre todo, nos permite modelar la manera en la que el ser humano es capaz de expresarse. Incluso dentro de otras técnicas es preciso encontrar un poco de teoría de lógica difusa, esto debido a que permite la extracción de conocimiento con cierto grado de incertidumbre[10].

La lógica difusa es un conjunto de conceptos y teorías que se pueden ver reflejadas en un marco de trabajo o framework conocido como Sistemas de Inferencia Difusos el cual nos permite realizar todo un proceso de aprendizaje, consta de 3 componentes esenciales: Reglas,

base de datos la cual contiene el tipo de funciones de membrecía que se utilizaran y un mecanismo de razonamiento que en base a las reglas realiza un estimado de una posible solución [12].

1.1.4. Minería de Datos

Desde que el ser humano comenzó a hacer uso de medios de almacenamiento de información, siendo estos electrónicos o no, ha realizado actividades de minado de datos con el fin de encontrar conocimiento entre la información que se le presentaba. Con el avance de la tecnología, la información se ha ido incrementando considerablemente complicando así la busqueda de información relevante para un determinado problema. La minería de datos surge como una hábil colección de técnicas y métodos capaces de la detección y adquisición de datos relevantes para el problema a tratar [5].

Algunas de las principales tareas con las cuales la minería de datos cuenta son: aprendizaje máquina, pre-procesamiento de datos, clasificación, la cual se encarga de identificar la información en base a sus atributos, regresión, agrupamiento, también llamado clustering, este término se refiere al proceso de identificación de grupos, reglas de asociación que permiten la identificación de características que varios datos tienen entre si, y visualización centrado en la manera en la que dicha información es presentada al usuario final [3]. La minería de datos empezó como tal siendo utilizada en el mundo de los negocios con el fin de predecir situaciones futuras, tomar decisiones y extraer conocimiento, surge de la combinación de varias áreas tales como estadística, inteligencia artificial, computación gráfica, bases de datos y procesamiento masivo, esto en los años 80ś cuando se dieron las bases para el desarrollo de ella, gracias a Rakesh Agrawal, Gio Wiederhold, Robert Blum, Gregory Piatetsky-Shapiro y a las grandes empresas que empezaron a hacer uso de ella.

De las primeras aplicaciones que marcaron el inicio de la minería de datos se tiene la

desarrollada por la administración de hacienda estadounidense la cual era un sistema para detectar fraudes en la declaración y evasión de impuestos, este sistema se desarrollo con técnicas tales como lógica difusa, redes neuronales y técnicas de reconocimiento de patrones [4].

Con el avance de la tecnología la minería de datos fue tomando importancia dentro de varias áreas; como se menciono anteriormente una de ellas fue la de los negocios, la cual es capaz de adquirir grandes volúmenes de datos de los clientes con la finalidad de entender mejor sus necesidades o futuras peticiones, otras áreas importantes que han aprovechado de técnicas de minería de datos (debido a su gran manejo de información que por métodos convencionales no podrían manejar) son: medicina, ciencia e ingeniería, algunos ejemplos dentro de estos campos incluyen manejo de la información recolectada por satélites, determinación de patrones patógenos en un paciente, clasificación de ADN, entre otras [5].

De los campos anteriores siendo medicina el campo de aplicación tratado en este documento, es notoria la evolución que ha tenido la misma al ir introduciendo aparatos electrónicos capaces de realizar mediciones con el fin de obtener información para un tratamiento o diagnóstico, esto ha propiciado que se tengan grandes cúmulos de datos los cuales no son aprovechados en su totalidad. Otro de los factores que afectan a la medicina es el crecimiento poblacional el cual no solo incrementa el número de información almacenada, sino también los recursos invertidos en el tratamiento de los pacientes, por ello es que la minería de datos surge como una alternativa para el aprovechamiento de los grandes volúmenes de información y como auxiliar del médico en el tratamiento del paciente propiciando con ello el manejo eficiente de los recursos [6].

La minería de datos ha sido aplicada dentro del área médica con tareas como por ejemplo evaluar la efectividad de un tratamiento en un paciente, esto mediante la asociación de causas, síntomas y medicamentos existentes. Otras de las tareas que ha comenzado a realizar es el

manejo de la administración con el fin de mantener un control sobre los pacientes que ingresan o salen, los de alto riesgo y aquellos con enfermedades especiales, con el objetivo de mantener un orden y optimizar la utilización de recursos. También se ha aplicado a la administración de clientes, la cual es ampliamente utilizada en el área de negocios, en medicina se utiliza con el objetivo de conocer las necesidades de los pacientes, preferencias y demandas para de esta manera proporcionar un servicio de calidad [7].

1.1.5. Sistemas de Soporte para la toma de decisiones

A los sistemas computacionales cuyo entorno está compuesto por técnicas para la toma de complejas decisiones son llamados sistemas para la toma de decisiones abreviado SSD (o en ingles *DSS*: Decision Support Systems), su funcionalidad es brindar ayuda a un experto a la toma de complejas decisiones dentro de un área particular, tomando información almacenada de dicha problemática a tratar.[15] Un SSD cuenta con 3 componentes principales los cuales son: un sistema de administración de base de datos, el cual le permite tener acceso a la base de datos con la cual será alimentada, sistema de administración de modelo base, cuya funcionalidad es procesar la información de tal manera que pueda utilizarla para realizar la toma de decisión y un sistema de administración y generación de diálogos, el cual funge como comunicador al usuario final [15].

Desde el uso de registros electrónicos en el sector médico son bastas las aplicaciones que han comenzado a surgir con el fin de llevar un control eficiente de los pacientes, es por ello que uno de los sistemas que más están siendo utilizados son los SSD clinicas con los cuales los médicos pueden detectar enfermedades con mayor rápidez, llevar el control de sus pacientes, detectar posibles reacciones en medicamentos recetados, recomendaciones de un tratamiento e incluso como auxiliar en la realización de un diagnóstico. Dichos sistemas son capaces de realizar funciones básicas tales como manipulación de base de datos electrónica

de pacientes, proporcionar planes de tratamiento, diagnósticos aproximados, manejo de la información, alertas sobre interacciones entre algunos medicamentos y recordatorios sobre cuidados preventivos [8].

Existen también dentro de este campo otros sistemas que han estado presentes realizando tareas de recomendación en sitios de compras, sitios de películas, entre otros, estos sistemas son denominados Sistemas de recomendación (SR), los cuales utilizan dos metodologías para realizar la recomendación, estas son: modelo de usuario y perfil de usuario, el primero obtiene información monitoreando la actividad del usuario dentro del entorno proporcionado, el segundo requiere la creación de una cuenta en la cual se proporcionen las preferencias de dicho usuario [13].

Existen diferentes tipos de sistemas de recomendación, algunos de ellos son:

- Recomendación Colaborativa: un usuario comparte sus intereses con otro basado en los intereses mutuos de ambos.
- Recomendación basada en contenido: Utiliza la información de un determinado objeto con el que el usuario interactuó para recomendar otros.
- Recomendación basada en conocimiento: Utiliza información del perfil para recomendar objetos basándose en los intereses.
- Híbridos: Mezclan diferentes tipos de recomendaciones [13].

La aplicación de estos sistemas de recomendación ha empezado a presentarse en el sector médico, esto debido a la implementación de registros electrónicos médicos los cuales sirven para que estor SR puedan ser capaces de recomendar a un determinado paciente información médica basándose en su condición y que sea entendible para el mismo.

1.2. Protocolo

1.2.1. Planteamiento del Problema

Dentro del área de la médicina constantemente se busca la incorporación de nuevas técnicas o herramientas que auxilien a conservar la salud del ser humano, durante el paso del tiempo se han incorporado sistemas capaces de realizar mediciones con el objetivo de conocer características propias de un determinado paciente y de esta manera poder conocer que lo aqueja y de esta manera tratarlo, sin embargo el proceso mediante el cual se toma una decisión de cómo tratar a un paciente queda en base al conocimiento del médico, siendo esto riesgoso para el paciente, esto debido al gran número de variables que se tienen que revisar para estimar un diagnostico acertado. De igual manera existe un nivel de complejidad al momento de realizar una recomendación sobre que tratamiento llevar, esto debido a ciertas consideraciones tales como: la sensibilidad a sustancias, efectividad del tratamiento, en algunos casos cambios alimenticios que en ocasiones necesitan adaptarse a nuevos descubrimientos que se puedan presentar en el paciente.

Los errores de diagnóstico médico son considerados la octava causa de muerte de un paciente, esto por diferentes motivos, entre ellos destacan: la omisión de algún dato importante que el paciente evita mencionar al médico, la interpretación errónea de los síntomas del paciente, tratamiento incorrecto (muy relacionada con la anterior), desconocimiento de la enfermedad debido al surgimiento de nuevos brotes de enfermedades o evolución de las mismas ocasionando que sean más resistentes a métodos de tratamiento convencionales.

El constante avance de la tecnología ha permitido que el ser humano tenga un ritmo de vida más cómodo y simple, sin embargo dentro de la ciencia médica siguen siendo pocas las aplicaciones que se han tenido en las cuales se involucre al paciente como responsable directo, es verdad que se ha contribuido a la mejora de aparatos de medición y herramientas

que auxilian al médico dentro del desempeñ o de su trabajo, sin embargo la relación entre médico y paciente queda solo dentro del consultorio, dejando a un lado cualquier tipo de recomendación que se le puede hacer y que ayude en su tratamiento, aunque es entendible que el médico no puede estar disponible a cada momento para el paciente. Debido a la complejidad que existe de modelar un sistema capaz de realizar las funciones de tratamiento de un medico son pocas las aportaciones de los mismos dentro del área médica.

La recopilación de información es un factor muy importante dentro de cuestiones medicas ya que permiten tener las herramientas necesarias para la formulación de un diagnostico exacto, pero debido a que es una gran cantidad de información por cada paciente es complejo el estar consultando múltiples registros del paciente propiciando que este proceso sea tardado y lento, en ciertas ocasiones se tienen sistemas que permiten automatizar un poco el proceso de consulta más aun siguen siendo ineficientes aparte de que solo permiten extraer información ya almacenada del paciente, sin embargo no es capaz de reaccionar de manera inteligente realizando recomendaciones preventivas, sugerencias de planes de tratamiento y diagnósticos.

Estos factores han hecho que la Ciencia y Tecnología traten de incursionar en la medicina con aplicaciones básicas que ayuden a los médicos al tratamiento de enfermedades o simplemente como auxiliar en la obtención de parámetros para la realización de diagnósticos, a su vez permiten al paciente llevar un control de su condición.

1.2.2. Justificación

Como se ha mencionado anteriormente el ser humano ha empezado a buscar dentro del área médica y por necesidad de herramientas que lo ayuden tanto a mantener o recuperar la salud, de estas problemáticas es por las que en conjunto con la ciencia y tecnología se han desarrollado tantas nuevas maneras de tratar o detectar enfermedades, así como de dispositivos que facilitan la obtención de parámetros con el fin de ayudar en el diagnostico o

avance de resultados de algún tratamiento.

La inteligencia computacional ha avanzado significativamente durante el transcurso de los años propiciando de esta manera que muchas de sus técnicas tales como sistemas expertos, lógica difusa, redes neuronales y minería de datos sean ampliamente explotados en varios campos de conocimiento, ya sea en aprendizaje, lingüística, reconocimiento de patrones, extracción de conocimiento, entre otras. Dando como resultado que la veamos envuelta dentro de nuestro ritmo normal de vida, incluso aunque no se percate uno de ella.

Algunas de estas técnicas ya han sido utilizadas dentro del campo medico para atacar problemas muy específicos o conocer datos cuantitativos de algún paciente. Se han realizado sistemas capaces de ayudar en la detección del cáncer de pecho basándose en métodos bayesianos [1], también se han presentado trabajos para la clasificación de patrones de electroencefalogramas (EEG), uno de ellos utilizando una red neuronal de impulsos [2].

Sin embargo las técnicas de IC que se aplican a las diferentes problemáticas son desarrolladas de forma muy específica ocasionando de esta manera que el campo de posibles técnicas a emplear sea limitado, siendo que dichas técnicas son genéricas, esto debido a la falta de software que tenga incluida la variedad de técnicas de IC orientadas principalmente a la toma de decisiones.

Existen dentro del campo médico herramientas de soporte para la toma de decisiones, entre estas herramientas nos interesa en particular los SSD los cuales utilizan métodos de IC para determinar un diagnostico aproximado, evaluar la efectividad de un tratamiento y supervisar la evolución de un paciente, dichos sistemas son de gran ayuda en minimización de errores, ahorro de recursos y sobre todo ayudan a la toma de decisiones.

Las técnicas que se pretende utilizar en este presente trabajo son las pertenecientes a IC y serán utilizadas para el minado de datos para su aplicación en un SSD, algunas de las principales razones por las cuales se pretende hacer uso de estas técnicas es debido a que

se enfocan a: descubrimiento de conocimiento, adaptación en base a los datos, detección de patrones que a simple vista o por algún método de clasificación no es posible encontrar de una manera tan eficiente, permiten trabajar con grandes cúmulos de datos y permiten tener mejores resultados que con otras técnicas, cabe mencionar que dichos sistemas son elaborados para ser utilizados por personas especialistas, el enfoque que se le desea dar en este presente trabajo será para cualquier tipo de usuario.

Una de las ventajas que puede tener la incorporación de estas técnicas al desarrollo de un sistema, es la pronta identificación y estimación de posibles patrones, propiciando con esto que los tiempos y recursos invertidos en la solución de un determinado problema se vean reducidos drásticamente.

1.2.3. Hipótesis

Los Algoritmos de Inteligencia Computacional para el minado de datos biomédicos proporcionan eficientes resultados que auxilian la toma de decisiones médicas.

1.2.4. Objetivos

Objetivos Generales

Analizar e implementar algoritmos de inteligencia computacional para su uso en la toma de decisiones médicas.

Objetivos Particulares

- Analizar algoritmos de Inteligencia Computacional para el minado de datos biomédicos.
- Analizar algoritmos difusos, neuronales, genéticos y de análisis estadístico para la toma

de decisiones.

- Implementar algoritmos con mejor desempeño y aproximaciones para toma de decisiones clínicas.

1.2.5. Metas

- Realizar una búsqueda de los diferentes algoritmos utilizados en minería de datos con aplicaciones biomédicas.
- Análisis de algoritmo K-means y aplicación en toma de decisiones.
- Análisis de algoritmo Gustafson-Kessel y su aplicación en toma de decisiones.
- Análisis de algoritmo Gath-Geva y su aplicación en toma de decisiones.
- Análisis de algoritmo Fuzzy c-means y su aplicación en toma de decisiones.
- Análisis de algoritmo sustractivo y su aplicación en toma de decisiones.
- Análisis de algoritmo de red bayesiana y su aplicación en toma de decisiones.
- Análisis de algoritmos de redes neuronales con aprendizaje hacía atrás y su aplicación en toma de decisiones.
- Análisis de algoritmos de redes neuronales de kohonen y su aplicación en toma de decisiones.
- Análisis de algoritmos de arboles de decisión y su aplicación en toma de decisiones.
- Análisis de algoritmo genético y su aplicación en toma de decisiones.
- Análisis de sistemas difusos de mamdani y sugeno y su aplicación en toma de decisiones.

- Análisis de algoritmo de vectores de soporte.
- Implementar algoritmo K-means.
- Implementar algoritmo Gustafson-Kessel.
- Implementar algoritmo Gath-Geva.
- Implementar algoritmo Fuzzy c-means.
- Implementar algoritmo substractivo.
- Implementar algoritmo de redes neuronales con aprendizaje hacía atrás.
- Implementar algoritmo de redes neuronales de kohonen.
- Implementar algoritmos de arboles de decisión.
- Implementar algoritmo genético.
- Implementar algoritmo de vectores de soporte.
- Implementar sistemas difusos de mamdani y sugeno.
- Seleccionar algoritmos y realizar pruebas con casos de estudio.

1.2.6. Metodología

Se planea realizar una investigación de tipo descriptiva esto ya que se realizara un análisis sobre las diferentes técnicas que se presentan dentro de IC y su aplicación para minado de datos, con el fin de ser implementados en una aplicación para la toma de decisiones.

Se utilizara el método deductivo ya que antes de poder realizar las implementaciones es necesario conocer como están conformados, de donde surgen los algoritmos, teoría del minado de datos, IC y sus técnicas, así como de los SSD.

Debido al rumbo que se llevara dentro de esta investigación se realizará un análisis sobre IC lo cual nos permitirá tener una noción de por qué son utilizadas dentro de otras áreas, partiendo de ello se realizará un estudio de mayor alcance dentro de minería de datos, el cual se convertirá en el tema principal de este trabajo, abarcando desde las técnicas totalmente basadas en estadística hasta llegar a las utilizadas en inteligencia computacional tales como redes neuronales, algoritmos genéticos, lógica difusa y aprendizaje maquina.

Se tendrá la siguiente clasificación:

- Técnicas basadas en estadística.
- Técnicas basadas en agrupación.
- Técnicas basadas en lógica difusa.
- Técnicas de redes neuronales.
- Técnicas de algoritmos genéticos.

Dentro de cada una de ellas se realizará un análisis sobre los algoritmos más representativos para su futura implementación. También se realizará un análisis sobre los SSD los cuales servirán como base para este trabajo ya que los algoritmos estarán incorporados en un sistema de esa clase, lo siguiente será revisar los sistemas que actualmente está presentes en medicina y son conocidos como sistemas de soporte de decisiones clínicas, se revisara la manera en la que operan, que funciones y características presentan, al igual que su desempeño dentro del área.

Los recursos de consulta que se usaráran en este trabajo son:

- Libros de los diferentes temas abordados.
- Artículos.

- Revistas Electrónicas.
- Internet: Sitios con enfoque científico, universidades, etc.

Dado que el área de enfoque de las diferentes técnicas que se presentaran en este documento es el médico, se trabajara con información perteneciente a condiciones médicas o en su caso mediciones realizadas a pacientes con el objetivo claro de realizar alguna de las tareas ya mencionadas.

Con cada uno de las técnicas se pretende realizar pruebas proporcionándole un conjunto de datos con el objetivo de determinar ciertos patrones en la información que nos puedan decir en algunos casos un diagnóstico y en base a el poder realizar una recomendación sobre lo que se puede hacer en dicho caso, ya sea dar sugerencias de un plan de tratamiento, observaciones de que sería lo más aconsejable a seguir o simplemente orientar en cómo llevar un control sobre su condición.

El aplicar dichos algoritmos surge de la necesidad dentro del área médica por realizar un procesamiento con la información de los pacientes que es recabada durante una intervención y sobre todo el auxiliar al médico o paciente con la enfermedad; tomando como base la información del paciente, así como mediciones que se hayan realizado, en base a ella se puede efectuar una serie de recomendaciones que le serian muy útiles al paciente, ya sea como medidas preventivas o como una simple manera de llevar un control.

La IC ha demostrado dentro de varias áreas que puede brindar grandes ventajas en cuanto a la búsqueda y descubrimiento de información, esto debido a las distintas técnicas propuestas que incluye, en las cuales se pretende que la computadora sea capaz de procesar los datos de una manera inteligente, tomando esta idea podemos aplicar dichas técnicas a diferentes problemáticas, es por ello que nos hemos sumado a realizar una implementación de estas técnicas para su incorporación a un SSD esto con el fin de ayudar a los pacientes y médicos

a llevar un proceso más controlado.

Estas técnicas tomaran los datos de los pacientes tanto personal como la recabada por aparatos de medición médicos, estos datos serán la base para que estas técnicas puedan encontrar patrones escondidos en dicha información y con esto poder realizar sugerencias al paciente o médico

El realizar un análisis completo sobre las diferentes técnicas existentes dentro de IC nos permitirá tener un amplio espectro sobre las capacidades de cada técnica, en base a esto nos dispondremos a realizar implementaciones de los algoritmos que puedan ser aplicados a un proceso de toma de decisión.

Teniendo los algoritmos ya implementados nos dispondremos a realizar pruebas con diferentes bases de datos para revisar su eficacia y eficiencia, esto nos permitirá tener información para comprobar las aproximaciones que nos generen al momento de realizar las diferentes tareas. Lo siguiente será empezar a desarrollar combinaciones de algoritmos para analizar si los resultados mejoran al incorporar un conjunto de técnicas para un mismo problema.

Organización de este documento

Este documento se encuentra organizado de la siguiente manera, el primer capitulo de la tesis es una breve Introducción donde se hace un resumen del tema tratado así mismo se incorpora el protocolo de tesis, en el sisguiente capitulo se presenta la información recabada sobre el tema de tesis. El tercer capítulo es una descripción detallada sobre la solución, dentro del cuarto capítulo se explica a fondo el desarrollo de la solución, en este caso se presentan diagramas de clases, diagramas de paquetes, tablas de algoritmos implementados y ejemplos de arquitectura des un sistema de decisión. El quinto capítulo muestra los experimentos y resultados obtenidos. El ultimo capítulo es la conclusión y cierre de este documento, así como

trabajo futuro.

Capítulo 2

Marco teórico

2.1. Minería de Datos

La minería de datos es una colección de tecnicas que tienen por propocito la extracción de información valiosa de una base de datos en particular [6], esto con el proposito de descubrir patrones que se mantienen ocultos o predecir futuros resultados. La míneria de datos es un subproceso de la disciplina del descubrimiento de conocimiento en bases de datos (en ingles *Knowledge Discovery in Databases (KDD)*), la cual se enfoca en convertir datos crudos en información valiosa. Los diferentes procesos que se realizan dentro de la mineria de datos se pueden observar en la figura 2.1. La adquisición de datos en diferentes areas ha tenido un crecimiento constante razón por la cual las tecnicas de míneria de datos han sido aplicadas con exito en una variedad de problemas [4].

Algunas de las principales tareas que la míneria de datos realiza son [21]:

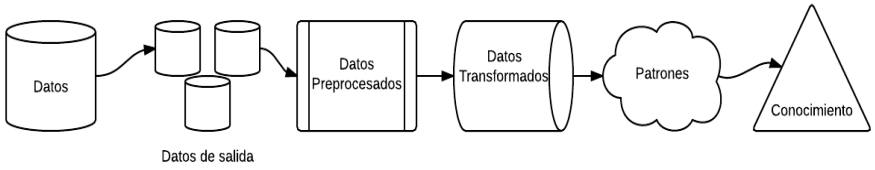

Figura 2.1: Procesos de la Mineria de Datos

2.1.1. Agrupamiento

Agrupamiento (en ingles *Clustering*): consiste en particionar los puntos en grupos llamados "clusters", los puntos en el mismo grupo comparten caracteristicas similares para esto se basa en la distancia del punto a los demás. Existen diferentes maneras de crear los grupos que dependen del paradigma de clusterización que se este empleando [21].

Técnicas de Agrupamiento

Existen diferentes clasificaciones dependiendo de como se van realizando las particiones que dan origen a un agrupamiento [2], algunas de los agrupamientos más utilizados son:

- Por particiones: estos tipos de algoritmos de agrupación construyen k particiones de un conjunto de datos en particular de n objetos. Tiene por objetivo minimizar una función objetivo, tal como la sumatoria de las raices de las distancias por la media.
- Jerarquico: crea una descomposición jerarquica de los datos, ya sea aglomerativa o divisiva. La descomposición aglomerativa trata a cada dato como un posible centro del grupo la primera vez que itera. Demanera iterativa une los centros de los grupos creando así nuevos grupos, donde cada dato tiene la misma distancia. La descomposición divisiva lo hace de de manera inversa comienza con un solo grupo al cual va dividiendo hasta formar grupos con caracteristicas similares.

- Agrupamiento basado en densidad: el agrupamiento es generado incorporando una función objetivo de densidad. La función objetivo esta defiida como un numero de objetos en un vecindario.

- Agrupamiento basado en cuadricula: este tipo de agrupamiento es usado en datos espaciales. Una cuadricula divide los datos en celdas para formar grupos. Este tipo de agrupamiento no depende en la distancia de los datos a diferencia de los anteriores.

- Agrupamiento Basado en Modelos: su principal objetivo es encontrar modelos que encajen con los datos de entrada. Incorpora algunas caracteristicas semejantes que las tecnica de agrupamiento basa en densidad con la diferencia de que incorpora modelos enteros.

Algunos algoritmos de Agrupamiento son:

- *K-means*: También conocido como C-means duro, cuyo objetivo es el encontrar clusters basados en la similitud de los datos mediante el calculo de la distancia entre cada dato. La función utilizada en esta técnica para el calculo de las distancias es la distancia euclidiana [17] [23]. Dicha técnica ha sido utilizada en diferentes áreas, tales como segmentación de imagenes, compreción de voz y mineria de datos.

- *Gustafson-Kessel* (GK): Es muy similar al anterior con la particularidad que utiliza otro metodo para calcular la distancia, este metodo se conoce como la distancia de Mahalanobis, este metodo se adapta de tal manera que forma elipsoides de los puntos, encerrando todo aquel que este cercano al punto central del grupo. Al ser un proceso iterativo, en cada iteración busca estimar los parametros como el punto central del grupo, matriz de covarianza y los datos encerrados en el elipsoide [19] [28].

- *Gath-Geva*: Es una extensión del algoritmo de Gustafson-Kessel, incorpora el tamaño

y la densidad como parametros a evaluar del grupo. A diferencia de GK no incorpora una función objetivo a ser minimizada como parametro de parada [12].

- *Fuzzy C-means* (FCM): Es uno de los algoritmos difusos más populares, convirtiendose en uno de los más modificados en la literatura. Igual que los anteriores algoritmos utiliza la distancia para evaluar las caracteriticas de los datos, incorpora un parametro para asignar el grado de fusificación a los datos, permitiendo de esta manera que un dato pueda pertenecer a multiples grupos [16] [11].

- *Subtractive* (SC): Esta basado en el método de la montaña, el cual cuadricula los datos con el proposito de obtener los centros de los grupos, para considerar un cuadro un centro necesitaba estar rodiado por multiples puntos dato. Este algoritmo permite utilizar radios de aceptación y rechazo. De esta manera evalua que es un centro y cuantos datos estan dentro de ese grupo.

2.1.2. Clasificación

Consiste en asignar una etiqueta a un dato que no se encuentra etiquetado, formando de esta manera conjuntos de clases. Existen diferentes tipos de clasificadores debidamente categorizados en: Clasificación Probabilistica, Arboles de decisión y Maquinas de Vectores de Soporte [21], siendo estos tres de los más importantes para esta tesis.

Algunos ejemplos de clasificadores probabilisticos son:

- *Bayes:* Como su nombre lo dice, utiliza el teorema de bayes para realizar la calsificación, este se basa principalmente en determinar para cada clase la función de probabilidad de densidad. Dicho teorema asume que cada atributo es independiente.

- *Naive Bayes:* A diferencia del anterior, este puede procesar problemas con un largo

número de dimensiones. Suponiendo que cada atributo es independiente este teorema descompone la probabilidad en un producto de probabilidades.

- *Vecino Proximo:* Este utiliza los datos para determinar la densidad y en base a ellas formar las clases.

Arboles de Decisión

Los arboles de decisión han sido utilizado en varias áreas como una poderosa herramienta dentro de sistemas para la toma de decisiones, con el objetivo de auxiliar a expertos a tomar la mejor decisión para resolver un problema especifico. Los algoritmos de arboles de decisión crean una colección de nodos estategicamente evaluados por un atributo predictor que encuentra la mejor ramificación que separe los datos en sub-grupos más homogeneos. Esto se realiza multiples veces iterando hasta clasificar todos los datos. El proceso comienza dividiendo un conjunto de datos en dos, uno es usado como entrenamiento y el segundo para probar si el arbol entrenado realmente puede desempeñar su función. Una vez que se le proporcionan los datos de entrenamiento al arbol, este selecciona un atributo raíz. Con base a la raíz se obtienen los nodos que mejor clasifican a los datos, esto continua hasta que obtiene el mejor atributo correspondiente a los datos. Uno de los algoritmos más utilizados es el dichotomiser iterativo 3 (ID3) introducido por Quinlan en 1986. Años más tarde se realizo una actualizacoón al mismo y fue llamado C4.5. Estos algoritmos calculan la ganancia de la información de cada atributo como recurso para ayudarlos a construir el arbol.

- *ID3:* Este algoritmo genera los nodos utilizando una busqueda de los datos codiciosa y de arriba a abajo. El calculo de la ganancia de informacion determina al atributo más valioso que clasifique los datos. Este calculo ayuda minimizando los niveles del arbol.

- *C4.5:* Este algoritmo también fue desarrollado por Ross Quinlan, quien lo propuso para superar las limitaciones del ID3 tales como la sensivilidad a conjunto de datos grandes. Utiliza un radio de ganancia de información para evaluar el atributo divisorio de cada nodo.

Maquinas de Vectores de Soporte

El objetivo principal de las Maquinas de Vectores de Soporte *en ingles: Support Vector Machines (SVM)* es encontrar la mejor función que permita clasificar los datos basandose en sus caracteriticas. Algunas de las ventajas que aporta SVM son que no presentan problemas con el número de dimensiones, requieren de pocos ejemplos para su entrenamiento, hoy en día cuentan con muchas técnicas eficientes para realizar sus entrenamientos. Existen diferentes problemas que pueden ser atacados por SVM, algunos de ellos pueden ser lo que se conocen como Linealmente separable con lo cual quiere decir, que basta con dibujar una linea entre las clases para realizar una separación de ellas, claro que en el campo es dificil encontrar datos de esta manera, razón por la cual SVM tambíen puede ser utilizada en problemas no lineales, para esto hace uso de kernels, los cuales ayudan a darle otro enfoque a los datos pudiendo de esta manera realizar una separación. Algunos de estos métodos lo que hacen es ver a los datos en 3 dimensión para poder saber si estan sobre puestos unos con otros. Existen diferentes tipos de kernel que ayudan a resolver dichos problemas [29].

2.1.3. Sistemas de lógica difusa

Un sistema de lógica difusa (en ingles: *Fuzzy Logic Systems (FLS)*) es un sistema que intenta manejar datos con cierta incertudumbre. En el mundo real no existe una representación logica simple para “0” y “1” o “Falso” y “Verdadero”, por ejemplo cuando alguien se expresa sobre el clima, dicen “menos frio” o “más frio”, este tipo de expreciones se conocen como

variables linguisticas dentro de FLS; en lugar de utilizar dichas expresiones dentro de FLS se asigna un grado de pertenencia que va desde el 0 hasta el 1, indicando de esta manera que tan acertado esta a la verdad. Para realizar la transformación de variable linguistica a valor númerico se utliza un fussificador el cual deacuerdo a la entrada hace la transformación a un valor númerico. Otros componentes importantes son las reglas y la máquina de inferencia los cuales son responsables de obtener posibles decisiones, por ultimo tenemos el defusificador, el cual hace la transformación de valores difusos a valores reales [15][20].

El proceso básico de un FLS comienza con los datos crudos los cuales son transformados por el fusificador después llevados a la máquina de inferencia donde en conjunto con las reglas los evaluan y devuelven salidas difusas, estas a su vez son transformadas por un defusificador en salidas crudas 2.2.

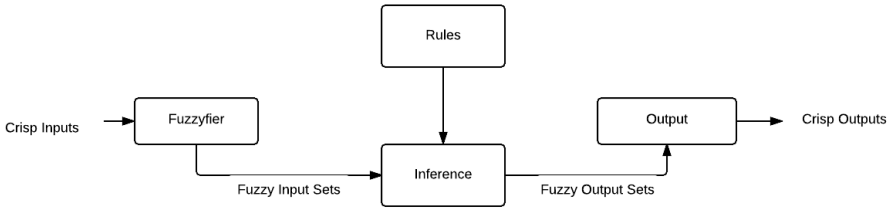

Figura 2.2: FLS Structure

Algunos modelos difusos explorados dentro de la literatura son:

Mamdani

Fue introducido por primera vez como controlador de una máquina de vapor. Las entradas eran un conjunto de variables linguisticas obtenidas de expertos humanos. En este model tanto las entradas como salidas son conjuntos. La base de conocimiento es la siguiente:

$$R^k : IFx_1 isF_1^k and ... and x_i isF_i^k and ... and x_n isF_n^k THEN y_1 isG_1^k(x) ... y_j isG_j^k(x) ... y_m isG_m^k(x) \forall k = 1, ..., r \tag{2.1}$$

Las funciones de membresia en los antecedentes de las reglas tienen la siguiente forma:

$$\mu_{F_i^k}(x'_{i,p}) \forall p = 1, ..., q k = 1, ..., r i = 1, ..., n \tag{2.2}$$

Las funciones de membresia en los consecuentes tienen la siguiente forma:

$$\mu_{G_i^k}(y_{j,p}) \forall p = 1, ..., q k = 1, ..., r i = 1, ..., n \tag{2.3}$$

y la fuerza de disparo es la siguiente:

$$\alpha_{k,p} = \prod_{i=1}^{n} \mu_{F_i^k}(x'_{i,p}) \tag{2.4}$$

Utilizando la defuzificación de Center of Sums

$$\hat{y_j} = \frac{\sum_{k=1}^{r} a_k A_{G_i^k} C_{G_i^k}}{\sum_{k=1}^{r} a_k A_{G_i^k}} \tag{2.5}$$

Takagi-Sugeno-Kang

Propuesta por Takagi, Sugeno and Kang para manejar conjuntos difusos en la entrada y una función en la salida. La estructura de la regla utilizada por este tipo de modelo es:

$$R^k : IFx_1 isF_1^k and ... and x_i isF_i^k and ... and x_n isF_n^k THEN y_1 isg_1^k(x) ... y_j isg_j^k(x) ... y_m isg_m^k(x) \forall k = 1, ..., r \tag{2.6}$$

Las funciones de membresia en los antecedentes de las reglas tienen la siguiente forma:

$$\mu_{F_i^k}(x'_{i,p}) \forall p = 1, ..., q k = 1, ..., r i = 1, ..., n \tag{2.7}$$

y una fuerza de disparo:

$$\alpha_{k,p} = \prod_{i=1}^{n} \mu_{F_i^k}(x'_{i,p}) \tag{2.8}$$

2.2. Sistemas de Soporte para la Toma de Decisiones

Los SSD han tomado gran popularidad en años recientes debilo a las multiples aplicaciones que se tienen como herramientas valiosas para humanos expertos en una área particular. Este tipo de sistema tienen por objetivo ayudar dentro de procesos cognitivos deficientes, esto lo realizan incorporando diferentes recursos de información, conocimiento relevante y una estructura de las posibles decisiones[10].

Los inicios de SSD empezaron a surgin en 1960 cuando investigadores compenzaron a explorar el uso de modelos computacionales cuantitativos como asistentes en planeaciones o decisiones, algunos trabajos representativos de dicha area surgieron con investigadores como Raymond 1966, Turban 1967, Urban 1967 y Holt y Huber 1969. El primer registro que se tiene de un sistema para la toma de decisiones asistido por computadora es el de Ferguson y Jones en 1969, la cual consistia en un sistema planeador de producción. Muchos de los sistemas

desarrollados fueron realizados por universidades, teniendo un incremento considerable en 1980 ocasionando que dicha area se expandiera considerablemente [25].

Algunos componentes principales de los SSD[10]:

- Sistema de Manejo de Base de Datos (en ingles *Data Management System (DBMS)*): Gestiona y almacena información sobre la problematica a la que el sistema se va a enfrentrar, de igual manera proporciona acceso al usuario a dicha información.
- Sistema de Manejo del Modelo Base (en ingles *Model-base Management System (MBMS)*): Transforma la información contenida en la base de datos por DBMS en información valiosa para la toma de decisión.
- Sistema Manejador y generador de Dialogos (en ingles *Dialog Generation and Management System (DGMS)*): Se encarga de gestionar la interacción con el usuario final que operara el SSD. Dicho componente debe contar con interfaces faciles de usar debido a que este tipo de sistemas no estan orientados a personas expertas en computadoras.

Dada las diferentes problematicas que existen y las distintas maneras de evaluar los problemas surgen categorias de los SSD cada uno de ellos con su propia manera de asistir a los usuarios. De las principales categorias que han surgido con los años son [25]:

- Impulsado por Modelo: Se enfoca en la manipulacion financiera, optimización y/o simulación de modelos. A diferencia de otra categoria, estos sistemas no requieren de grandes conjuntos de datos para la toma de decisión. Se encuentran limitados en datos y parametros.
- Impulsada por comunicación: Hace uso de redes y técnologias de la comunicación para facilitar la toma de decisiones colaborativas o la comunicación.

- Impulsada por datos: Se enfocan en la manipuación de datos internos, externos y de tiempo real de una determinada compañia.
- Impulsada por documentos: Utiliza almacenamiento computacional y técnologias de procesamiento para analizar y recuperar documentos. Esta especialmente centrado en utilización con politicas, procedimientos, especificaciones de producto, catalogos y documentos corporativos historicos.
- Impulsada por conocimiento: Utilizado por especialistas como herramientas que le proporcionan sugerencias y recomendaciones para la resolución de problemas especificos. Contienen datos especializados para la tarea que pretenden resolver.

2.2.1. SSD Clinicas

SSD Clinicos permiten a medicos, enfermeros y pacientes disponer de información especifica, inteligente, filtrada y apropieda al tiempo. Este tipo de sistemas son muy utilizados ya que son de bajo costo, ayudan a eficientizar los procedimientos clinicos e incrementan los beneficios para el paciente. De sus funciones principales son el ahorro de recursos, prevención, indicadores de peligro, recomendaciones de tratamientos, etcetera. Como todo sistema de soporte de Decisiones los clinicos tratan de auxiliar a un experto en la dificil tarea de tomar una decisión en lugar de completamente remplazarlo. Por más que dicho sistema sea muy bueno, su operabilidad siempre debe ser por algun usuario experto [3].

Muchos de los sistemas de este tipo han optado por incorporar técnicas de inteligencia artificial, esto debido a su desempeño en tareas dentro de diferentes áreas, dentro de medician se ha utilizado en sistemas para la prescripción de medicamentos, vigilancia clinica, entre otros. Los SSD se utilizan casi en cualquier área donde se manejen datos a gran cantidad. La dificultad que estos sistemas han encontrado no radica en los problemas a atacar, si no en

integrarlos dentro del proceso del experto clinico, incluso aunque su fiabilidad y precisión ha sido demostrada en varias ocasiones muchos expertos siguen sin querer cambiar su manera de trabajar e integrarlas[8].

Los SSD Clinicos presentan dos categorias principales: basados en conocimiento y no basados en conocimeinto. Los basados en conocimiento en su mayoria incorporan conjuntos de reglas tipo"IF-THEN", un motor de inferencia y un mecanismo para comunicarse. Al correr este tipo de sistemas hacen uso de las reglas contenidas en una base de datos, después aplicando estas a los datos introducidos generar salidas las cuales son externalizadas haciendo uso del mecanismo de comunicación. La mayoria de los SSD Clinicos tienen que lidiar con vagedad, imprecisión e incertidumbre en los datos, razón por la cual se utilizan diferentes técnicas capas de procesarlos [3].

Con el constance crecimiento de datos dentro de áreas como la médicina se busca que se utilicen diferentes técnicas que sean capaces de procesar dichos datos y arrojar patrones importantes que realmente nos digan sobre un determinado comportamiento. Esto ha hecho que se utilize mineria de datos en diferentes sistemas medicos para tareas como detección de patrones, clasificación de datos, procesamiento de grandes cantidades de datos, entre otros. Esto ha hecho que los SSD Clinicos empiecen a utilizar técnicas de mineria de datos. Dentro de las principales diferencias entre sistemas de conocimiento basados en reglas "IF-THEN" y aquellos con mineria de datos, es que aquellos con reglas deben ser proporcionado con hechos y reglas asociados a ellos mientras que los que utilizan mineria de datos son capaces de descubrir patrones y aplicar lo descubrierto al conjunto de datos [14].

Las consideraciones principales a la hora de implementar un SSD Clinico radican en primordialmente que problema se quiere atacar o que área se desea explorar, seguido de esto tenemos a que usuario se hará llegar la información pertinente o resultado del sistema y finalmente que tanto control el usuario tendra sobre la información, esto depende del tipo de

usuario que realmente vaya a utilizar el SSD Clinicas. Estos sistemas clinicos pueden involucrarse en varias tareas dentro del proceso clinico. Tareas tales como cuidados preventivos, diagnosticos, monitoreo del tratamiento, seguimiento, alertas, entre otras [5].

2.2.2. Aplicaciones

Algunas de las aplicaciones que hoy en día se pueden encontrar sobre los SSD son muy variadas. Dentro del area médica se puede encontrar en conjunto con algunos dispositivos de sensado tales como un oxímetro de pulsos, el cual funciona para detectar cambios en la condición del paciente, la tarea que desempeña un SSD con este dispositivo consiste en alertas o mantener recordatorios al médico. De igual manera es usada en laboratorios en procedimiento de evaluación de resultados de alguna determinada prueba realizada a un paciente [**?**]. Una de las principales tareas y que recientemente ha cobrado mucho interes dado la dificultad que implica su realización es el diagnostico, los sitemas que se enfocan en atacar esta problematica simplemente sirven de asistencia a los médicos, son incapaces de tomar la decision por su cuenta. Existen otro tipo de sistemas enfocados a la evaluación de los planes de tratamiento, los cuales se enfocan en buscar inconsistencias que puedan presentar, un ejemplo de ello se puede presentar al meter una orden para transfución sanguinea el sistema puede ser capaz de analizar y señalarle al medico si es que los niveles de hemoglobina se encuentran arriba del nivel permitido. Con el constante interes por almacenar datos, se ha convertido en una tarea muy dificil el buscar información dentro de los miles de datos que se tienen, por ello los SSD son usados para ayudar a encontrar información valiosa o necesaria que ayude a la solución de un determinado problema, un ejemplo de ello puede ser un sistema Web que a base de preguntas pueda filtrar la información para de esta manera otorgar lo más relevante [8].

Capítulo 3

Propuesta de los algoritmos para la biblioteca de clases DMCL

3.1. Propuesta

Las ciencias de la salud en años recientes ha buscado incorporar nuevas tecnologias que le auxilien a mantener su principal objetivo, el cual es mantener la salud del ser humano, para ello han desarrollado técnicas médicas. Con ayuda del computo inteligente se han incorporado sistemas para la administración, manejo de pacientes y como soporte dentro de ciertos procesos medicos. Dentro de los sistemas que se han comenzado a utilizar destaca la participación de las técnicas de mineria de datos. La aplicación de estas técnicas dentro del área médica se debe en gran parte a que permiten la extracción de patrones, descubrimieto de conocimiento y classificación de información. Existen varias aplicaciones con muy buenos resultados, algunas de ellas son: detección de cancer de pecho, detección de tumores cerebrales, cancer de pulmon, diabetes y extracción de características [26]. La colección y almacenamiento de datos ha sido realmente aprovechado por las técnicas de mineria de datos,

las cuales han sido usadas como un medio para ayudar en la toma de decisiones dentro de procesos medicos [27].

Las ciencias de la salud cuentan con multiples procesos que se basan en la recolección de datos con el fin de entender una patologia. Estos datos son recolectados realizando examenes fisicos, examenes invasivos y consultando el historial clinico. Esto con el fin de ayudar al médico a realizar un diagnostico o incluso elaborar un plan de tratamiento [9]. El procesos de diagnostico es muy complicado tanto para medicos expertos como principiantes, esto dado que a cada paso del diagnostico se deben realizar decisiones tales como la realización de examenes que realmenten aporten información valiosa para la formulación de un buen diagnostico. Esto va de la mano con realmente entender la enfermedad que se desea tratar. El proceso por el cual se acepta o no un diagnostico se basa primordialmente en el método cientifico. El diagnostico es un proceso iterativo que depende de nuevos descubrimientos, esto permite que sea evaluado de nuevo a cada paso para así tener el mejor diagnostico posible.

Los sistemas antes mencionados y que han empezado a ser utilizados dentro de algunas áreas son los llamados sistemas de soporte para la toma de decisiones (SSD), los cuales ayudan al ser humano en el proceso de toma de decisión dentro de una determinada tarea. Estos sitemas estan conformados por fuentes de información o conocimiento. Dentro de estas fuentes incluso se encuentran algunas técnicas provinientes del área de inteligencia artificial. Estos sistemas han sido utilizados para el cumplimiento de tareas tales como monitoreo de pacientes, alertas o advertencias médicas de algunos procesos o contra-indicaciones, manejo de datos, entre otras más [7].

En recientes años los sistemas de soporte para la toma de decisiones han aumentado dentro de los diferentes campos de aplicación, esto ha hecho que se involucren diferentes técnicas como nucleo de estos sistemas, algunas provinientes del área de inteligencia computacional las cuales tienen por ventaja el procesamiento de grandes volumenes de información, la busqueda

de conocimiento y relaciones en los datos o caracteristicas [24].

El análisis de algoritmos de inteligencia computacionas y de sistemas de soporte para la toma de decisione nos ha llevado a realizar la implementación de los algoritmos k-means, fuzzy c-means, gath-geva, gustafson-kessel, subtractivo, así como maquinas de inferencia difusa, redes neuronales y arboles de decisión, todos ellos contenidos en la biblioteca DMCL la cual tiene como objetivo auxiliar en la implementación de SSD's, así como siendo un servicio para el procesamiento de datos masivos. Se orienta a desaroolladores de sistemas inteligentes. Dicha biblioteca se encontrara alojada en un servidor con recursos optimos para el procesamiento de la información, los desarrolladores se conectaran a esta para procesar su información obteniendo como respuesta datos procesados. La biblioteca estara desarrollada bajo el lenguaje Microsoft C# debido a que con ello entramos dentro de un mercado de aplicaciones numerosas y de fácil implementación al ser un lenguaje orientado a objetos.

Capítulo 4

Implementación de la biblioteca de clases DCML

4.1. Data Mining Class Library

En este capítulo se describe a detalle el marco de trabajo para el desarrollo del software denominado Data Mining Class Library (DMCL). El cual fue desarrollado para ser utilizado como una biblioteca de funciones para aplicaciones de toma de decisiones medicas. Dicha biblioteca esta formada de técnicas de minera de datos enfocandonos principalmente a la formación de grupos (en ingles *clustering*), de igual manera se incluyen ciertas técnicas para ayudar en la toma de decisiones. Estas técnicas son arboles de decisiones y Sistemas de Inferencia difusa. Todas estas técnicas parten de Inteligencia Computacional siendo esta área nuestra zona principal de enfoque. La biblioteca tiene la posibilidad de ser utilizada no solo en médicina, dado que son técnicas muy generales puedes ser usadas en multiples trabajos donde se requiera de un gran procesamiento de datos. Dentro de diferentes áreas tales como Ingeniería, Ciencia, Médicina, solo por mencionar unas, se han empezado a utilizar

diferentes softwares que con ayuda de técnicas de inteligencia computacional ayudan a un usuario experto a tomar decisiones complicadas o simplemente a realizar mejor su trabajo informandolo de tareas importantes. DMCL esta pensada en ser utilizada para ser usada por programadores que tengan por objetivo el desarrollo e implementación de este tipo de sistemas o para tareas de detección de patrones, procesamiento de datos, descubrimiento de conocimiento, entre otras.

4.1.1. Biblioteca de clases DMCL

DMCL es una biblioteca de clases desarrollada en su totalidad en C#. El proposito principal es construir Sistemas de soporte para la toma de decisiones haciendo uso de la programación orientada a objeto como medio de implementación. La figura 4.1 muestra el contenido de la binlioteca DMCL.

Diseño DMCL

DMCL esta conformado por una colección de paquetes conformados por clases 4.2. Los paquetes estan organizados de manera que dentro de cada uno residen clases pertenecientes a un tipo de técnica especifica.

Como podemos ver en la figura 4.2 los algoritmos se encuentran divididos en Clustering, DSS y Fuzzy Inference System principalmente, dentro de cada uno de ellos se encuentran diferentes clases cada una de ellas es un algoritmo propio del tipo de técnica especificado por el titulo del paquete. El contenido de cada paquete esta explicado a detalle en la tabla 4.1. Con el uso de la prgramación orientada objeto y las propiedades con las cuales cuenta se pueden realizar muchas variantes. El uso de la herencia dentro de la biblioteca permite tener algoritmos que comparten caracteristicas pero su funcionamiento es diferente, con esto hacemos más eficiente la implementación de estas técnicas.

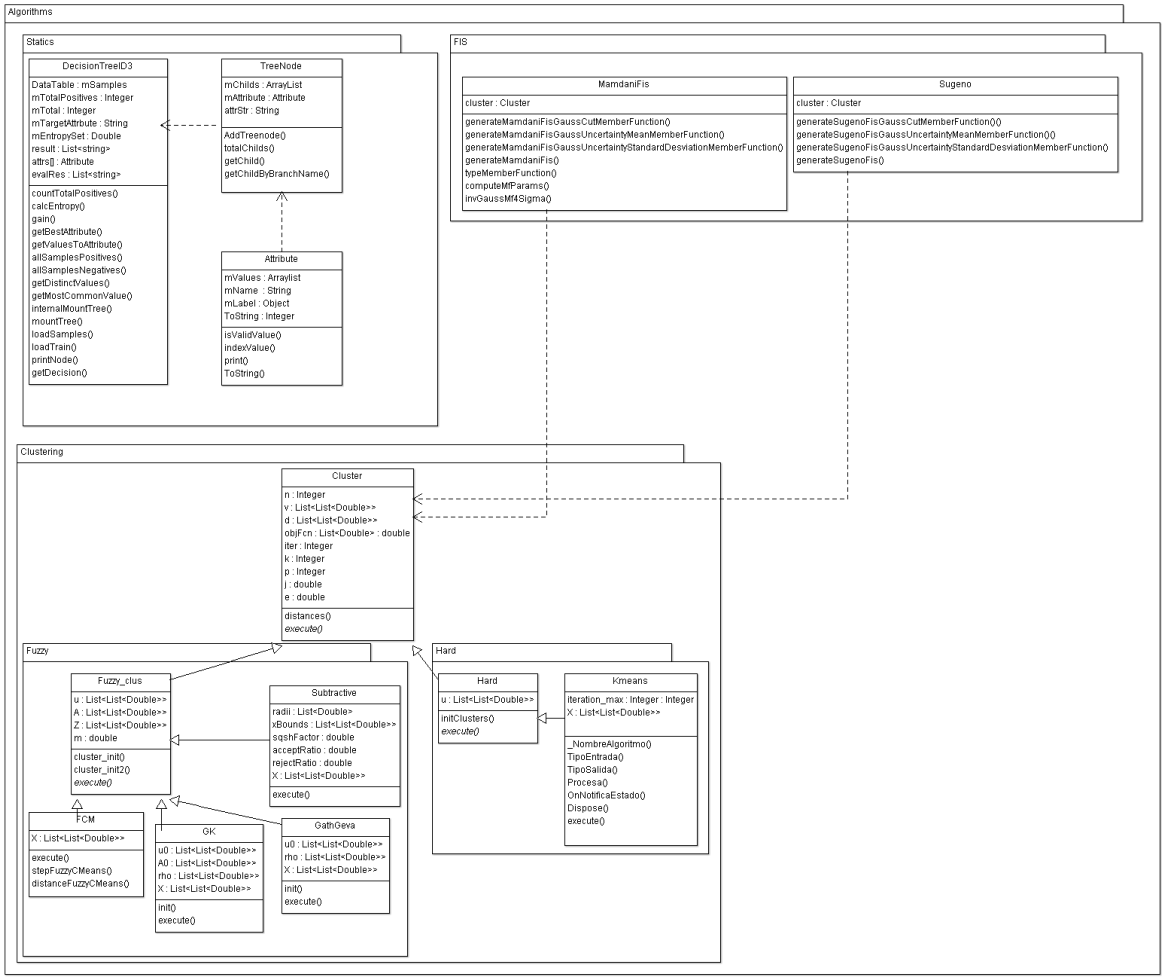

Figura 4.1: Diagrama de Clases de DMCL

La estructura mostrada en la tabla 4.1 comienza con un paquete principal llamado algoritmos donde se encuentran Clustering, DSS y Fuzzy Inference Systems. Dentro de Clustering tenemos dos carpetas y una super clase de la cual heredan ciertos métodos las clases contenidas en las carpetas, dichas se especializadas en la formación de grupos: Hard y Fuzzy, dentro de hard tenemos dos clases unicamente una super clase llamada Hard_clust que hereda de Cluster y Kmeans la cual hereda de Hard_clust. Dentro del paquete Fuzzy tenemos 5 clases: una super clase llamada Fuzzy_clus y 4 algoritmos implementados que heredan de dicha super clases, estos son: Fuzzy C Means, GustafsonKessel, Gath-Geva y Subtractve. En la figura 4.3 podemos observar un diagrama UML de clases de dicho paquete. De igual manera se pueden

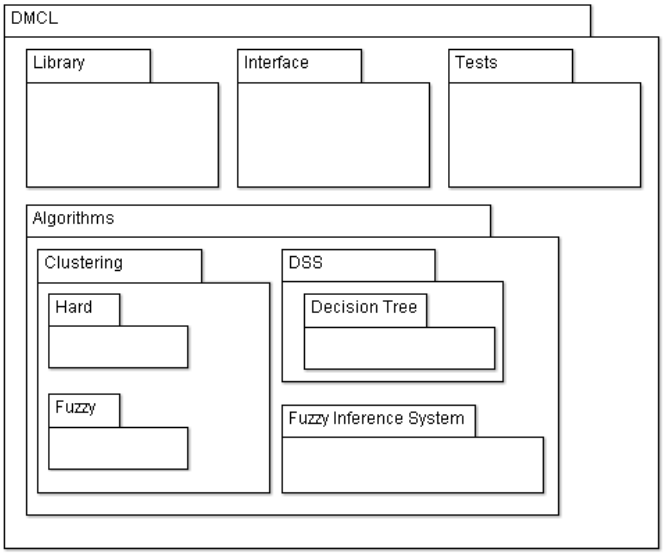

Figura 4.2: Diagrama de paquetes UML de DMCL

observar los metodos utilizados por dichas clases y como por versatilidad se opta por usar la herencia para realizar una abstracción de lo que se conoce como cluster en mineria de datos y los tipos que pueden existir de ellos. Dichos tipos dependen de un calculo muy utilizado el cual es la distancia, razón por la cual se implementan dichos metodos en la superclase.

También se tiene un paquete especializado en clasificación llamado DSS dentro tenemos el paquete Decision Tree y dentro tenemos las clases: Attribute, Decision Tree y Tree Node, estas en conjunto se enfocan en la creación de arboles para generar decisiones o simplemente para la clasificación de datos. Dicha estructura se puede observar en la figura 4.4 la cual muestra un diagrama UML de clases donde se muestran las clases y su dependencia entre ellas, ya que para la creación de un arbol es necesario la utilización de la clase Attribute la cual se encarga de contener los datos que se quieren evaluar y sus respectivas etiquetas o nombre del dato. La clase Tree Node es la encargada de contener los datos caracteristicos del arbol que sirven de nodo para el mismo.

Por ultimo tenemos el paquete Fuzzy Inference System el cual contiene solo dos clases: MamdaniFis y SugenoFis las cuales crean sistemas de inferencia difusa de Mamdani con

Tabla 4.1: DMCL Clases

Paquete		
Clustering	Cluster	
	Hard	Hard_clus
		Kmeans
	Fuzzy	Fuzzy_clus
		Fuzzy C Means
		Gustafson Kessel
		Gath-Geva
		Subtractive
DSS	Decision Tree	Attribute
		DecisionTreeID3
		TreeNode
Fuzzy Inference System	MamdaniFis	
	SugenoFis	

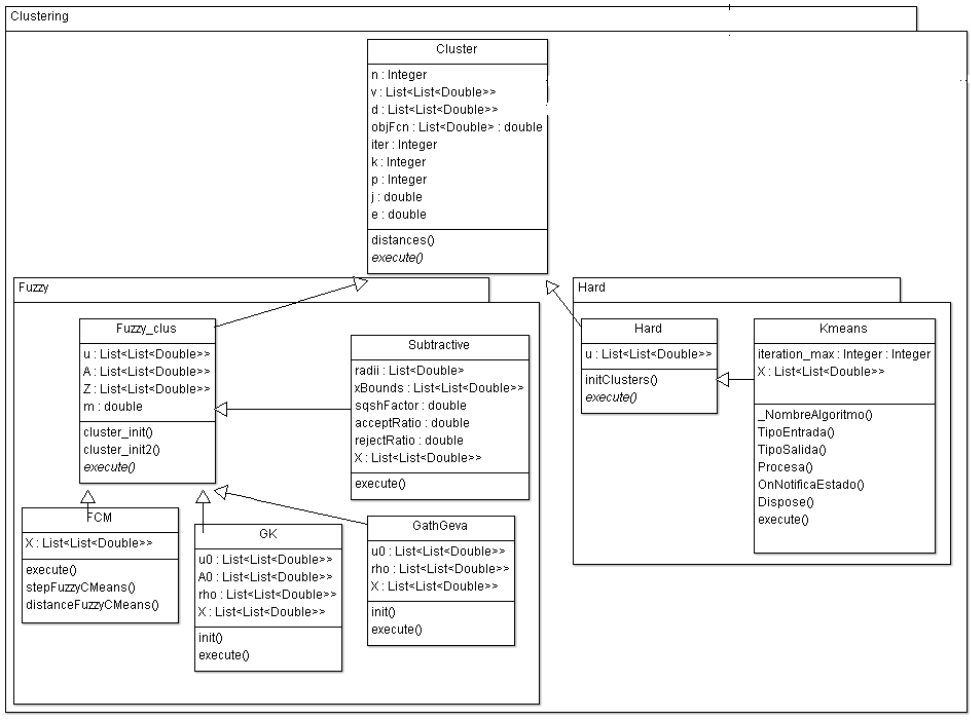

Figura 4.3: Diagrama UML de clases del paquete Clustering

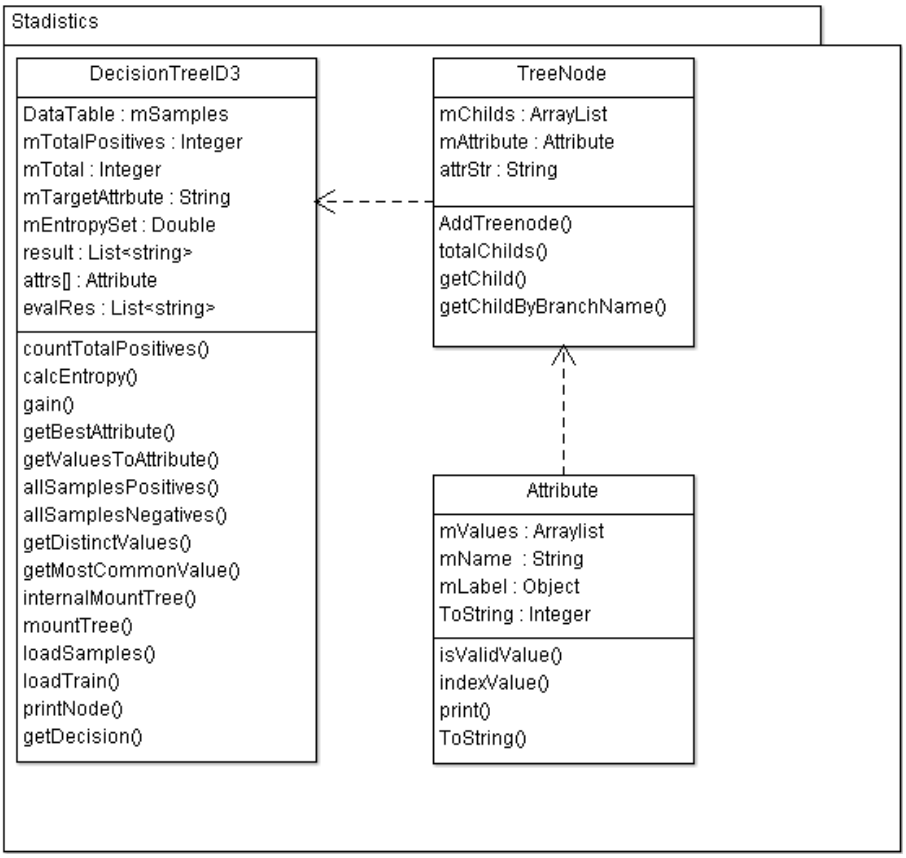

Figura 4.4: Diagrama UML del Paquete Stadistics

conjuntos difusos o Sugeno con una función a la salida. En la figura 4.5 se puede observar un diagrama UML de clases donde se ve la implementación de dichas clases, cabe mencionar que ambas clases hacen uso de alguna técnica de clusterización, esto debido a que dichas implementaciones utilizan la agrupación para determinar las reglas más optimas que se deben utilizar por la maquina de inferencia.

En la tabla 4.2 se puede observar las tareas que los algoritmos antes mencionados pueden desempeñar en un ambito cualquiera, asi como las tareas que se han ralizado dentro de sistemas computacionales médicos al incorporar estos algoritmos.

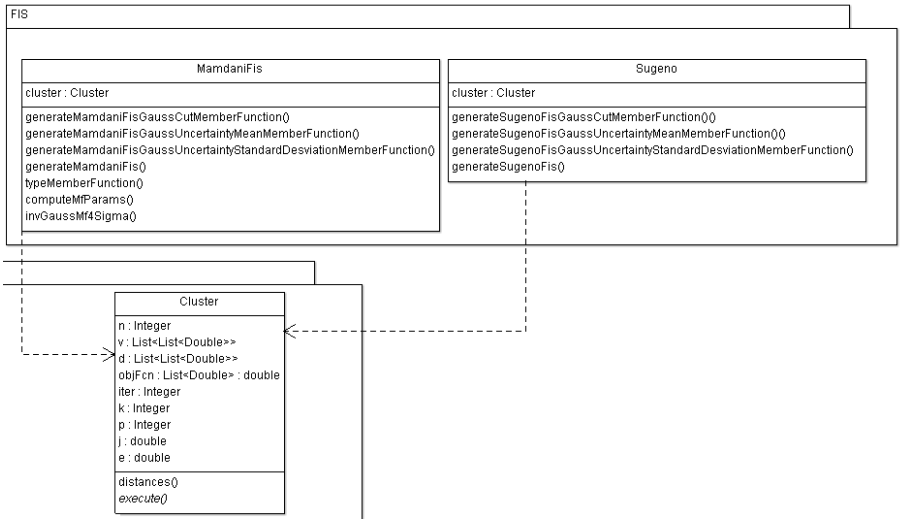

Figura 4.5: Diagrama de Clases UML del paquete FIS

Tabla 4.2: Algoritmos y sus aplicaciones

Algoritmo	**Tarea desarrollada**	**Aplicaciones Medicas**
K-means	Agrupamiento	Clasificación de Señales
Gustafson-Kessel	Agrupamiento Difuso	Extracción de reglas difusas
Fuzzy C-Means		Particiones difusas
Gath-Geva		
Subtractive	Agrupamiento	Pre-procesamiento de datos
Decision Tree ID3	Clasifiacion , Toma de decisiones	Diagnosticos
Fuzzy Logic System	Evaluacion, Toma de decisiones	Evaluación de tratamientos

4.1.2. Implementación de los algoritmos de DMCL

La implementación de cada uno de los algoritmos contenidos dentro de la biblioteca y debido a que esta desarrollada utilizando el paradigma orientado a objetos, para su uso es necesario la creación o instaciar un objeto con el nombre de la clase dado. Para dicha implementación se debe tomar en cuenta que algoritmo se quiere utilizar y que parametros se le pueden dar al algoritmo. Si hablamos de agrupamiento un parametro que no puede faltar es el número de grupos que deseamos tener a la hora de que el algoritmo realice la creación de grupos. De igual manera podemos tener otros parametros.

Algoritmo de Agrupamiento

Para la implementación de un algoritmo de agrupamiento podemos ver el listado siguiente 4.1 en el cual se puede observar que necesitaremos una entrada el cual puede ser un vector o una matriz de datos. A su vez proporcionamos un número de grupos que deseamos tener, si es todo lo que deseamos enviarle al algoritmo este utilizara los parametros que estar decalrados por default.

Listing 4.1: Object-oriented coding example

```
//parameters
int K = 4;
X = getData();
// Creacion del objeto K-means
Kmeans Km = new Kmeans(X, K);
Km.execute();  // Ejecucion del Algoritmo
Km.V // Obtener Grupos formados
```

Algoritmo de Sistema de Inferencia Difusa

La implementación de un FIS se puede observar en el listado siguiente 4.2, para su utilización los datos de entrada y de salida tienen el formato de una matriz. Dado que una particularidad de este tipo de sistemas es que manejan incertidumbre, este parametro se le puede proporcionar a la funcion de membresia utilizada por el modelo difuso. Dado que apra este tipo de implementaciones es requerido una técnica de agrupamiento esto con el objetivo de obtener reglas optimas para la maquina de inferencia difusa, en este caso estamos utilizando Fuzzy C-means. Puede utilizarse cualquiera de las técnicas de agrupamiento incorporadas en DMCL. Por ultimo se crea la instancia haciendo uso de objeto modelo (MamdaniFis

o SugenoFis) y uno de sus métodos que van relacionados a los parametros que queramos utilizar.

Listing 4.2: Object-oriented coding example of Fuzzy Inference System

```
//parameters
Input inputList = new Input();
Outpur outputList = new Outpur();
inputList.Add(inputs); // vector de datos
outputList.Add(outputs); // vector de datos
uncertainty = 0.0;
//Tecnica de agrupamiento para determinar
//las reglas del modelo difuso
FuzzyCMeans m = new FuzzyCMeans();
//Modelo difuso
MamdaniFis genFis = new MamdaniFis(m); // Fuzzy Model

//Creacion del FIS
Mamdani fis =
genFis.generateMamdaniFisGaussUncertainty
MeanMemberFunction(inputList, outputList, uncertainty);
```

Algoritmo de Toma de Decisiones

En el listado 4.3 se puede observar una implementación de un arbol de decision tipo ID3 el cual requiere de los nombres de las clases que se usan para buscar los datos y formar el nodo. También se necesita proporcionar las salidas del algoritmo, en este caso seran 3. Lo siguiente es entrenar el arbol con datos de entrenamiento y a su vez la etiqueta de cada columna.

Pasamos las posibles salidas al algoritmo y por utlimo iniciamos el proceso proporcionandole los datos de entrenamiento, los datos que deseamos clasificar y lo que deseamos encontrar.

Listing 4.3: Object-oriented coding example of Decision Tree

```
//parametros
//clases
String[] columns = new string[] { "Sepal_len", "Sepal_wid",
   "Petal_len", "Petal_wid", "class" };
// Posibles Salidas
String[] answer = new String[] { "1", "2", "3" };
// Creacion del objeto ID3
DecisionTreeID3()  id3 = new DecisionTreeID3();
// Datos a Evaluar
DecisionTree.Attribute[] attribute = new
   DecisionTree.Attribute[] { sl, sw, pl, pw };
// Datos para entrenamiento
DataTable samples = id3.loadSamples(columns, data);
// Pasar las posibles salidas al algoritmo
id3.Answers = answer;
// Generacion de los nodos del arbol
TreeNode root = id3.start(samples, "class", attribute);
// Imprimir Arbol
id3.printNode(root, "");
```

4.1.3. Implementación de la biblioteca Externa de Redes Neuronales

Dentro de la lista de algoritmos que se incluirian dentro de DMCL se encuentran redes neuronales con propagación hacia atras y kohonen. Dichas implementaciones se realizaron utilizando una biblioteca llamada NeuronDotNet la cual esta completamente desarrollada bajo el lenguaje Microsoft C#. Dicha biblioteca permite la creación de redes neuronales con diferentes parametros de entrada tales como utilizar diferentes funciones de aprendizaje, indicar las conecciones entre las capas de la red, indicar el numero de neuronas que se tienen en la entrada, en la capa o capas ocultas y en la salida, permite inicializar los pesos de la red mediante varias funciones y finalmente tener control sobre el entrenamiento de dicha red. Para la utilización de dicha biblioteca se creo una plantilla dentro de DMCL con los parametros que se utilizan más a la hora de realizar la implementación de una red neuronal. Dichos parametros son los mencionados anteriormente. Para utilizar esta plantilla es necesario proporcionar las entradas, salidas, número de neuronas, número de ciclos que va a iterar la red, factor de aprendizaje y datos de entrenamiento para realizar el aprendizaje de la red, dichos datos proporcionan patrones y su resultado. Una vez que se proporcionan todos los datos mencionados anteriormente se realiza el entrenamiento y por la obtención de resultados proporcionando las entradas a evaluar por la red. La implementación de la plantilla se puede observar en el listado 4.4.

Listing 4.4: Object-oriented coding example of Neural Network

```
//Parametros para configurar la Red.
double[] sampleInput = new double[] { 1d, 2d, 3d, 4d,
   5d, 6d, 7d };
double[] sampleOutput = new double[] { 0.01d, 0.02d,
   0.06d, 0.08d, 0.10d, 0.12d, 0.14d };
```

```
        double learningRate = 0.3d;
        int neuronCount = 10;
        int cycles = 100;
        TrainingSet trainingSet = new TrainingSet(1, 1);
        BackPropagationNetworkTemplate network = new
           BackPropagationNetworkTemplate(sampleInput,
           sampleOutput
learningRate, neuronCount, cycles, trainningSet );

        // Entrenamiento de la red.
        network.startTraining();

          // Evaluacion y resultados de la red.
        network.getResult(sampleInput);
```

4.1.4. Ejemplo de Arquitectura del Sistema de Decisiones

Los SSD se encuentran principalmente constituidos por fuentes de información o conocimiento siendo este algun tipo de mecanisco capas de generar o inferir los datos necesarios para ayudar a un experto a tomar una decisión. Este mecanismo interno se encuentra conformado por algunas técnicas de inteligencia computacional algunas de estas se encuentran incorporadas dentro de la biblioteca mencionada anteriormente. Algunas de las aplicaciones que se pueden encontrar en la literatura desarrollaban su función dentro de áreas como médicina, ingenieria, ciencia, negocios, etc. El enfoque dentro de esta tesis se encuentra ligado a la médicina, razón por la cual se han consultado en la literatura diferentes sitemas y funconalidades de los mismos. Una área que incorpora sistemas dentro de médicina es la

de Sistemas de Información de Salud, la cual se enfoca en sistemas de salud que ayudan en la recolección, analisis y evaluación de la calidad de datos para finalmente transformarlos en recursos valiosos dentro de la toma de decisiones.

Algunas de las aplicaciones que se pueden desarrollar utilizando la herramienta desarrollada son los diagnosticos médicos los cuales utilizando los datos del paciente obtendrian caracteristicas mediante algun algoritmo de agrupamiento y con la ayuda de un arbol de deciones o un sistema de inferencia difusa podrian tomar una decisión que ayude a clasificar una determianda patologia encontrada en base a los datos del paciente, en la imagen 4.6 se puede observar el modelo de dicha aplicación de ejemplo.

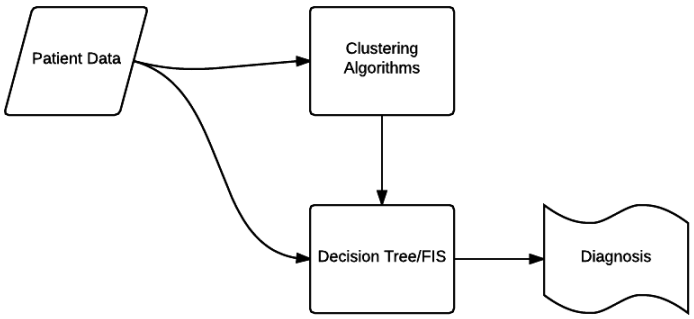

Figura 4.6: Ejemplo de Aplicación de diagnostico médico

Otras de las posibles aplicaciones que la herramienta podria tener es en el desarrollo de aplicaciones de recomendacion, algunos de ellos se pueden encontrar en sistemas de video, tiendas en linea, pequeños sistemas de salud, entre otros. Esta aplicación de recomendación tomaria como base los articulos visitados o vistos y las preferencias del usuario previamente almacenadas para en base a ellas y la colección de articulos que tiene hacer una clasificación sobre cuales recomendar mediante un arbol de decisiones o un sistema de inferencia difusa. El modelo se puede observar en la imagen 4.7 siguiente.

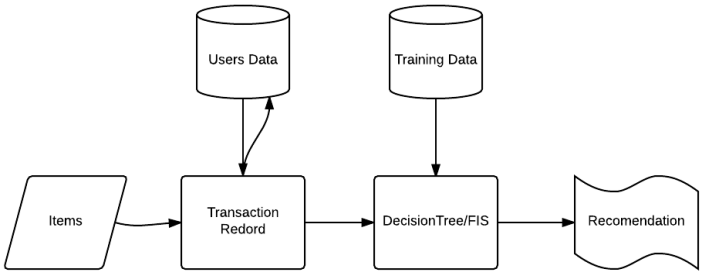

Figura 4.7: Ejemplo de Aplicación de recomendación

4.1.5. Pruebas

DMCL fue probada utilizando base de datos de referencia tales como Colesterol, Iris, Motocicleta, Cancer, Dermatologia, y otros casos base. Como otro punto de referencia tomamos la aplicacion de Matlab®, la cual ya contiene varios de los algoritmos implementados en DMCL, se probaron las bases de datos mencionadas y se evaluo si se obtenian los mismos resultados proporcionandoles los mismos parametros de entrada.

Las bases de datos que se utilizaron para probar los algoritmos son las siguientes:

- Colesterol

  La cual contiene un muestreo de 21 mediciones con 264 muestras de sangre, los cuales seran usados para su clasificación dentro de 3 tipos de colesterol, los cuales son: LDL, VLDL y HDL para la identificación de algun tipo de patron que sepueda presentar en los datos. Razón por la cual se utilizo en algunos algoritmos de agrupamiento, dado que nos permiten encontrar grupos basados en caracteristicas semejantes. Esto nos da de entrada con las dimensiones de 21*264 y las salidas de 3*264, esta base de datos puede ser encontrada en algunos programas de Aprendizaje Máquina e igual en el repositorio de la universidad de Irvine.

- Dermatologia

Esta base de datos cuenta con 34 atributos de los cuales 33 son valores lineales y 1 nominal, dicha base de dato tiene como objetivo el diagnostico del erythemato-squamous el cual es una enfermedad dermatologica. Dicha base de datos incluye los siguientes atributos:

- Erythema
- Escala
- Bordos precisos
- Picason
- Koebner
- Papula poligonal
- Papula Folicular
- Mucosa
- Rodilla y codo
- Cuero cabelludo
- historia familiar
- Incontinencia melanica
- Eosinophils
- PNL
- Fibrosis de la dermis papilar
- Exocytosis
- Acanthosis
- Hyperkeratosis

- Parakeratosis
- Crestas epiteliales
- Elongamiento de las crestas epiteliales
- Adelgazamiento de la epidermis suprapapilar
- Pústula espongiforme
- Microabesos munro
- Hipergranulos focal
- Desaparición de la capa basal
- Vacuolización y daño de la capa basal
- Espongiosis
- Aspecto de los dientes de sierra de retes
- Enchufe de cuerno folicular
- Paraqueratosis perifolicular
- Inflamación monoluclear
- Infiltrado de banda
- Edad

Este es un problema de clasificación, en la que los pacientes pueden caer dentro de 6 denominaciones diferentes: Psorioasis, Dermatitis Seboreic, Liquen plano, Pitiriasis rosada, Dermatitis Cronica y Pitiriasis rubta pilaris. Cada ellas fueron númeradas del 1 al 6 debidamente para la evaluación de dicha base de datos para su debida implementación [13].

- Enfermedad del Corazón

Esta base de datos contiene 76 atributos, pero los datos importantes y mayormente utilizados en la literatura son 14. Con ellos se pretende estimar la presencia de enfermedad en el corazón, comenzando con la clase 0: no presenta hasta la clase 4 [1].

- Edad
- Sexo
- Tipo de dolor de pecho
- Presión de sangre en reposos
- Colesterol
- Azucar en sangre
- Electrocardiograma en reposos
- Ritmo cardiaco maximo
- Angina inducida por ejercicio
- Depresión ST inducida por ejercicio
- Curva de pico durante ejercicio
- Numero de buques principales
- Ritmo cardiaco
- Enfermedad cardiaca diagnosticada (A encontrar)

- Iris

Base de datos que cuenta con 150 datos de 4 características de una muestra de 1000 flores, estas características son:

- Largo del Sepalo en cm
- Ancho del sepalo en cm

- Largo del petalo en cm
- Ancho del petalo en cm

Con estas características se pretende estimar dentro de que denominación de las 3 disponibles las 150 flores son categorizadas. Este es un claro ejemplo de un problema de clasificación donde se cuentan con solo 3 clases para ello [22].

- Lentes

  Esta base de datos fue extraida del libro [18], la cual contaba con 24 datos cada uno con 4 atributos diferentes. Con estos datos se pretende estimar que tipo de lentes es recomendable que un usuario utilice. Los atributos a considerar son los siguientes:

  - Edad
  - Prescripción de anteojos
  - Astigmatismo
  - Proporción de producción de lagrimas

- Jugada de Tenis

  Esta base de datos fue extraida del libro [18], la cual contaba con 14 datos y 3 atributos diferentes, los cuales tienen como objetivo estimar si dichos atributos son los indicados para que un juego de tenis se realice al aire libre. Dicha base de datos tiene dos posibles respuestas “si” y “no”. Los atributos mencionados son:

  - Estado (clima)
  - Humedad
  - Viento

Capítulo 5

Pruebas y Resultados

5.1. Pruebas

Algoritmo de agrupamiento

En la tabla 5.1 se pueden observar los grupos formados con el algoritmo K-means por la libreria DMCL y el mismo algoritmo integrado en Matlab®.

Tabla 5.1: Comparación entre DMCL y Matlab® con K-means.

	Grupo #1	Grupo #2	Grupo #3
Matlab®	0.3322	0.5235	0.2159
	0.3820	0.6057	0.2454
	0.3725	0.5903	0.2389
	0.3118	0.4896	0.2015
	0.2612	0.3926	0.1736
DMCL	0.3322	0.5234	0.2158
	0.3820	0.6057	0.2454
	0.3725	0.5903	0.2389
	0.3117	0.4896	0.2014
	0.2611	0.3926	0.1735

Se realizo una prueba similar utilizado el algorimto Fuzzy C-means, con los mismos para-

metros de entrada que el anterior, a excepción del parametro defuso el cual se dejo el utilisado por defecto de 2. El número de grupos a encontrar era de 3. En la tabla 5.2 se pueden observar los resultados obtenidos de las dos implementaciones.

Tabla 5.2: Comparación entre DMCL y Matlab® con Fuzzy C-means.

	Grupo #1	Grupo #2	Grupo #3
Matlab®	0.3149	0.4434	0.2126
	0.3617	0.5120	0.2416
	0.3525	0.4988	0.2352
	0.2952	0.4150	0.1984
	0.2489	0.3396	0.1708
DMCL	0.3148	0.4431	0.2125
	0.3615	0.5117	0.2415
	0.3524	0.4985	0.2351
	0.2950	0.4148	0.1983
	0.2488	0.3394	0.1707

La siguiente prueba fue con el algoritmo subtractive, el cual es un algoritmo difuso para agrupamiento. En esta prueba se obtuvieron solamente 2 grupos, a diferencia de los dos anteriores esta técnica utiliza radios para formar los grupos. Los resultados obtenidos se puedes observar en la tabla 5.3.

Tabla 5.3: Comparación entre DMCL y Matlab® con Subtractivo

	Cluster #1	Cluster #2
Matlab®	-0.3592	0.3399
	0.8262	0.5588
	1.0030	0.1464
	0.1098	-0.0744
	1.2785	0.6150
DMCL	-0.3592	0.8262
	0.3398	0.5588
	1.003	0.1463
	0.1098	-0.0743
	1.2785	0.6150

Sistema de Inferencia Difusa Modelo Mamdani

Se realizaron pruebas con el sistema de inferencia difusa de Mamdani utilizando en conjunto una técnica de agrupamiento para determinar las reglas a utilizar para el sistema, en este caso se utilizo Fuzzy C-Means. El dataset a probar fue el de colesterol con 3 reglas a calcular, la incertidumbre proporcionada fue de 0, la función de membresia a utilizar fue Gauss. Como se puede ver en 5.4 se obtiene las entradas del sistema las, el numero de reglas generadas y las salidas del mismo, por cuestones de visualización solo se pusieron las 5 primeras entradas de un total de 21. Al final aparecen las reglas generadas y su interacción con las entradas.

Tabla 5.4: Resultado de Modelo Mamadani

	DMCL FIS Mamdani	Matlab
Entrada01	sigma=0.0637, mean1=0.2273, mean2=0.2273, sigma=0.0553, mean1=0.3106, mean2=0.3106, sigma=0.0732, mean1=0.4049, mean2=0.4049	0.0638 0.2274 0, 0.05531 0.3106 0 0.07329 0.405 0
Entrada02	sigma=0.0746, mean1=0.2590, mean2=0.2590, sigma=0.0649, mean1=0.3563, mean2=0.3563, sigma=0.0856, mean1=0.4669, mean2=0.4669	0.07467 0.2591 0, 0.06497 0.3564 0 , 0.08567 0.4669 0
Entrada03	sigma=0.0730, mean1=0.2524, mean2=0.2524, sigma=0.0633, mean1=0.3462, mean2=0.3462, sigma=0.0845, mean1=0.4563, mean2=0.4563	0.07306 0.2525 0, 0.06338 0.3463 0, 0.08455 0.4563 0
Entrada04	sigma=0.0606, mean1=0.2129, mean2=0.2129, sigma=0.0522, mean1=0.2891, mean2=0.2891, sigma=0.0702, mean1=0.3810, mean2=0.3810	0.06064 0.213 0, 0.05228 0.2892 0, 0.07021 0.381 0
Entrada05	sigma=0.0484, mean1=0.1822, mean2=0.1822, sigma=0.0404, mean1=0.2474, mean2=0.2474, sigma=0.0505, mean1=0.3066, mean2=0.3066	0.04849 0.1822 0, 0.04045 0.2474 0, 0.05052 0.3066 0
Salida01	sigma=28.7545, mean1=35.1013, mean2=35.1013, sigma=27.1162, mean1=42.9970, mean2=42.9970, sigma=45.7390, mean1=142.7713, mean2=142.7713	28.75 35.1 0, 27.12 43 0 , 45.74 142.8 0
Salida02	sigma=29.8161, mean1=101.5116, mean2=101.5116, sigma=37.2189, mean1=177.2724, mean2=177.2724, sigma=25.9445, mean1=92.3218, mean2=92.3218	29.82 101.5 0, 37.22 177.3 0, 25.94 92.32 0
Salida03	sigma=10.7374, mean1=43.3431, mean2=43.3431, sigma=9.3480, mean1=44.8987, mean2=44.8987, sigma=6.9701, mean1=30.8023, mean2=30.8023	10.74 43.34 0, 9.348 44.9 0, 6.97 30.8 0

Sistema de Inferencia Difusa Modelo TSK

De igual manera se utilizo un sistema de inferencia difusa tipo Takagi-Sugeno-Khan en conjunto con el algoritmo de agrupamiento subtractivo el cual es considerado semi-supervizado, esto debido a que solo se proporciona un radio el cual utiliza para discriminar datos. Se utilizo el dataset de colesterol, un radio de 0.5 , una función de membresia gausiana y una incertidumbre de 0.1. Como se pede ver en el listado 5.5 se obtienen 21 entradas, 3 salidas y un 5 reglas. Por cuestiones de visualización solo se pusieron 5 entradas, por la misma razón solo se puso una salida. Al final se puede observar las 5 reglas y su interacción con las entradas.

Arbol de Decisiones ID3

El arbol de decisiones ID3 fue probado utilizando diferentes base de datos benchmark entre ellas se encuentra Iris, Dermatologia, Enfermedad del corazón, Lentes y Juego de Tenis. Dichas base de datos tienen como objetivo estimar la clasificación de los datos contenidos. Dichos pueden tener tanto clasificaciones como simples afirmaciones o negaciones, o hasta resultados más complicados como estimar enfermedades.

La base de datos del iris se puede observar en la siguiente tabla 5.6, como se puede observar se probo proporcionandole la mitad del dataset. Y con unos datos de prueba mostrados a continuación se pretendio obtener la clasificación correcta.

Otra de las bases de datos probadas fue la base de datos de enfermedades del corazón, la cual se puede observar en la tabla mostrada a continuación 5.7.

De igual manera se utilizo la base de datos de los lentes. Se realizo una prueba muy sencilla para comprobar si estaba clasificando correctamente las entradas. Dicha prueba se puede observar en la tabla siguiente 5.8

Tabla 5.5: Resultado de Modelo TSK

	DMCL FIS TSK	Matlab
Entrada01	sigma=0.1742, mean1=0.2346, mean2=0.2593, sigma=0.1742, mean1=0.3158, mean2=0.3491, sigma=0.1742, mean1=0.1613, mean2=0.1783 sigma=0.1742, mean1=0.3736, mean2=0.4129 sigma=0.1742, mean1=0.3131, mean2=0.3460	0.1743 0.247 0, 0.1743 0.3325 0, 0.1743 0.1699 0, 0.1743 0.3933 0, 0.1743 0.3296 0
Entrada02	sigma=0.2062, mean1=0.2654, mean2=0.2934, sigma=0.2062, mean1=0.3619, mean2=0.4000, sigma=0.2062, mean1=0.1820, mean2=0.2011 sigma=0.2062, mean1=0.4279, mean2=0.4729 sigma=0.2062, mean1=0.3574, mean2=0.3950	0.2063 0.2795 0, 0.2063 0.381 0, 0.2063 0.1916 0, 0.2063 0.4505 0, 0.2063 0.3763 0
Entrada03	sigma=0.2056, mean1=0.2588, mean2=0.2860, sigma=0.2056, mean1=0.3519, mean2=0.3889, sigma=0.2056, mean1=0.1779, mean2=0.1967, sigma=0.2056, mean1=0.4145, mean2=0.4582, sigma=0.2056, mean1=0.3488, mean2=0.3856	[0.2057 0.2725 0, 0.2057 0.3705 0, 0.2057 0.1874 0, 0.2057 0.4364 0, 0.2057 0.3673 0
Entrada04	sigma=0.1722, mean1=0.2184, mean2=0.2414, sigma=0.1722, mean1=0.2933, mean2=0.3241, sigma=0.1722, mean1=0.1528, mean2=0.1689, sigma=0.1722, mean1=0.3442, mean2=0.3805, sigma=0.1722, mean1=0.2934, mean2=0.3243	0.1723 0.2299 0, 0.1723 0.3088 0, 0.1723 0.1609 0, 0.1723 0.3624 0, 0.1723 0.3089 0
Entrada05	sigma=0.1244, mean1=0.1879, mean2=0.2077, sigma=0.1244, mean1=0.2524, mean2=0.2790, sigma=0.1244, mean1=0.2871, mean2=0.3173, sigma=0.1244, mean1=0.2871, mean2=0.3173, sigma=0.1244, mean1=0.2429, mean2=0.2685	0.1244 0.1978 0, 0.1244 0.2658 0, 0.1244 0.1413 0, 0.1244 0.3022 0, 0.1244 0.2557 0
Salida01	-2515621.0086 -14338623.9747 17783454.9004 25685524.2190 -44212170.2884 1373247.6542 2614077.5433 -13046600.3002 44710559.0288 13547996.5176	6673 -2.654e+04 3.558e+04 -2.106e+04 4185 242 -3445 -1.22e+04 5.031e+04 -5.241e+04

Tabla 5.6: Clasificación de datos de base de datos Iris y sus respectivas salidas

Datos	Predecida	Dato real
5.5,2.6,4.4,1.2	2	2
4.8,3.1,1.6,0.2	1	1
6.1,3,4.9,1.8,3	2	3
5.8,2.7,3.9,1.2,2	2	2
6.1,3,4.6,1.4,2	2	2
5,3.2,1.2,0.2,1	1	1

5.2. Discusión de Resultados

La implementación de los algoritmos de Inteligencia computacional dentro de DMCL fueron probados utilizando bases de datos benchmark. Los resultados obtenidos fueron co-

Tabla 5.7: Clasificación de base de datos de enfermedades del corazón

Datos	Predecida	Dato real
62,1, 2,128,208,1,2,140,0,0,1,0,3	0	0
57,1,4,110,201,0,0,126,1,1.5,2,0,6	0	0
58,1,4,146,218,0,0,105,0,2,2,1,7	1	1

Tabla 5.8: Recomendación de lentes a usuarios

Datos	Predecida	Dato real
Jove, miope, si, reducida	duros	duros

rroborados usando la aplicación Matlab, la cual ya contiene algunos de los algoritmos implementados en DMCL, asi mismo las bases de datos utilizadas son bastante conocidas dentro de la literatura y cada una de ellas cuenta con las salidas que deberian obtener los algoritmos a probar. Algunos de los algoritmos requerian de parametros para que se ajustaran de tal manera que fueran capaces de encontrar de manera adecuada los datos. Una de las principales decisiones por las que surge DMCL es para tratar de acelerar la implementación de sistemas inteligentes dentro del área médica, sin embargo viendo que existen muy pocas soluciones de librerias de Inteligencia computacional se pretende en un futuro expandir las capacidades de la libreria. El probar con diferentes bases de datos nos demuestra que dicha implementación funciona de manera adecuada con diferentes problematicas. Al realizar una investigación sobre los algoritmos más utilizados dentro de la literatura pertenecientes a inteligencia computacional nos arrojo ciertos nombres claves de algoritmos, algunos de los cuales presentamos en esta libreria. Tomamos los algoritmos más probados, robustos y clasicos que se encontraron. De igual manera en la busqueda realizada dentro de Sistemas de Soporte para la toma de decisiones se encontro que varios de los algoritmos mencionados en este documento participaban en su mayoria como motores de estos sitemas o facilitaban su funcionamiento mediante combinaciones de técnicas lo cual nos facilito un poco el seleccionar los algoritmos indicados para agregar a DMCL.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusión

La medecina en la busqueda incansable por mantener la salud del ser humano ha buscado el implementar nuevas y mejores técnicas dentro de sus miles procesos. Hoy en día con el gran incremento del uso de las computadoras y el constante crecimiento de ellas han hecho que varias áreas incluyendola pongan sus ojos en ella, ya sea optimizar procesos, auxiliar en procesos, descrubrir conocimeinto, entre otros; por esta razón se ha buscado dentro de ŕeas como inteligencia computacional de valiosas técnicas que ayuden a problematicas dentro de la misma, la mayoria de estas problematicas tienen que ver con el analisis y procesamiento de datos.

Inteligencia computacional cuenta con un gran número de técnicas capaces de ayudar a procesar y analisar grandes cumulos de datos, entender estos datos y aportar datos caracteristicos que ayuden a descifrar mejor una cierta problematica. Existen dentro de la literatura diferentes tipos de sistemas que buscan el auxiliar de alguna manera al ser humano a la toma de decisiones, estos sistemas ya mencionados en esta tesis auxilian al ser humano a tomar

decisiones que dependen de muchas variables, una de la razones por las cuales se obto por tomarlos como base es debido a que DMCL surge como una solución de las pocas existentes que permite la implementación de sistemas de toma de decisiones de manera más rápida ya que proporciona varias técnicas ampliamente utilizadas dentro de este tipo de sistemas. Dichos algoritmos implementados no solo pueden servir para la toma de decisiones si no para la creación de sistemas inteligentes, procesamiento de datos e incluso clasificación de los mismos.

El desarrollo de sistemas inteligentes dentro de cualquier área proporciona un gran apoyo dentro de diferentes tareas que se pretenden realizar, especialmente se han visto siendo integradas dentro de áreas como ciencia, ingenieria quimica, bioingenieria, telecomunicaciones y más. Uno de los principales motivos por los cuales se han empezado a utilizar estas técnicas de manera muy recurrente se debe a que existen grandes cumulos de datos almacenados por diferentes procedimientos que se realizan y los cuales no son debidamente utilizados, ya que el ser humano por si solo no podria analizar todos los datos de manera eficiente. Todo esto nos ha llevado a realizar la implementación de los algoritmos dentro de la libreria DMCL ya que se pretende en un futuro realizar aplicaciones de esta indole para la resolución de tareas de recomendación en diferentes áreas.

Con las comparaciones realizadas en los diferentes experimentos se puede observar como dichas ténicas presentan robustez sin importar la implementación que dichas tengan, en ese caso fueron realizadas en dos lenguajes de programación muy diferentes, cada uno de ellos con propiedades muy especiales. Los resultados obtenidos nos mostraron que la implementación de DMCL era buenas ya que era capaz de replicar los resultados obtenidos en la implementación que posee Matlab®, sin embargo en cuestiones de tiempo ciertos casos no fueron tan buenos como otros. Esto se debe en parte a la biblioteca matemática elaborada por nosotros, ya que Matlab® cuenta con los mejores algoritmos matematicos incorporados en su núcleo.

La implementación de esta biblioteca se pudo dar debido a un análisis de los algoritmos de IC que más fueron encontrados en la literatura siendo utilizados en SSD así como sus diferentes aplicaciones a un basto número de problemas. Dentro de la literatura se ha observado que su eficacia en resolución de problemas ha tenido buenos resultados, razón por la cual se les ha encontrado en un gran número de sistemas utilizados hoy en día. Si bien en análisis de los algoritmos mencionados en esta tesis fueron concentrados ampliamente a sistemas de toma de decisiones médicas, este mismo análisis nos ha mostrado que la aplicación de estos algoritmos se encuentra presente en un gran número de áreas, en las cuales han sido utilizados para elaboración de sistemas inteligentes, servicios, optimizaciones, entre otras. Lo cual posicione a la biblioteca DMCL como una buena herramienta para ser utilizada hoy en día para resolver un gran número de problemas que requieran del uso de IC.

6.2. Trabajo Futuro

Los algoritmos implementados dentro de DMCL son de los más conocidos y probados, sin embargo existen aun más algoritmos que benefician enormemente a los sitemas de toma de decisiones que no estan incluidos en DMCL, uno de los posibles trabajos a futuros sera expandir el repertorio de algoritmos incorporados. La implementación de esto algoritmos se hizo desde cero, implementando incluso una libreria matemática para el funcionamiento de ls algoritmos, sin embargo varias de las operaciones que se realizan no son del todo obtimas provocando que la eficiencia de la biblioteca decaiga un poco, por ello se pretende implementar los algoritmos más eficientes que se encuentren en la literatura o incluso incorporar una bibliotca matemática bastante robusta que se encuentre en el mercado. Dado que el procesamiento de datos masivo es muy costoso se planea alojar la biblioteca en un servidor que cuente con tarjetas graficas que puedan ser utilizadas para ayudar al procesamiento de los mismos. Gracias al análisis realizados para la elaboración de DMCL se ha decidido aplicarla

a un gran número de problemas que se encuentran en diferentes áreas no relacionadas con medicina.

Bibliografía

[1] David W. Aha and Dennis Kibler. Instance-based prediction of heart-disease presence with the cleveland database. *Aritificial Intelligence in Medicine*.

[2] Periklis Andritsos. Data clustering techniques. *University of Toronto: Deparment of Computer Science*, 2002.

[3] P.K. Anooj. Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. *Journal of King Saud University - Computer and Information Science*, pages 1319–1578, 2011.

[4] Eta S Berner. Clinical decision support systems: Theory and practive. *Springer Science and Business Media*, pages 207–270.

[5] Eta S. Berner. Clinical decision support systems: State of the art. *Agency for Healthcare Research and Quality*, 2009.

[6] Sameer Verma Bijan Fazlollahi, Mihir A. Parikh. Adaptive decision support systems. *Decision Support Systems*, 20:297–315, 1997.

[7] Sameer Verma Bijan Fazlollahi, Mihir A. Parikh. Adaptive decision support systems. *Decision Support Systems*, 20:297–251, 1997.

[8] Enrico Coiera. Clinical decision support systems. *Taylor and Francis Group*, 2015.

[9] John B. Wong Daniel B. Mark. Decision-making in clinical medicine. *Harrisons Principles of International Medicine*, (19), 2015.

[10] Marek J. Druzdzed and Roger R. Flynn. Decision support systems. *Encyclopedia of Library and Information Science*, 2002.

[11] M. N. Ahmed S. Yamany N. Mohamed A. A. Farag and T. Moriarty. A modified fuzzy c-means algorithms for bias field estimation and segmentation of mr date. *IEEE Transactions on Medical Imaging*, 21(3):193–199.

[12] Gath and A. B. Geva. Unsupervised optimal fuzzy clustering. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 11(7):773–781.

[13] G. Demiroz H. A. Govenir and N. Ilter. Learning differential diagnosis of eryhemato-squamous diseases using voting feature intervals.

[14] J. Michael Hardin and David C. Chhieng. Data mining and clinical decision support systems. *Springer New York*, pages 44–63, 2007.

[15] Eiji Mizutani Jyh-Shing Roger Jang, Chuen-Tsai Sun. Neuro-fuzzy and soft computing a computational approach to learning and machine intelligence. *Prentice-Hall*, 1997.

[16] M. Naji K. M. Bataineh and M. Saqer. A comparison study between various fuzzy clustering algorithms. *Jourdan Journal of Mechanical and Industrial Engineering*, 5(4):335–343.

[17] Prof. Fakhreddine Karray Khaled Hammouda. A comparative study of data clustering techniques. *Cite Seer*, pages 1–21.

[18] García Martinez, Servente, and Pasquinig. Sistemas inteligentes. *Nueva Libreria*, 2003.

[19] Andre Lemos Maurilio Ignacio and Walmir Caminhas. Fault diagnosis with evolving fuzzy classifier based on clustering algorithm and drift detection. *Mathematical Problems in Engineering*, pages 3–4.

[20] Jerry M. Mendel. Uncertain rule-based fuzzy logic systems: Introduction and new directions. *Prentice Hall PTR*, 2001.

[21] Wagner Meira Jr Mohammed J. Zaki. Data mining and analysis: Fundamental concepts and algorithms. *Cambridge University Press*, 2014.

[22] D.W. Murphy P.M. Aha. Uci repository of machine learning databases [http://www.ics.uci.edu/ mlearn/mlrepository.html]. *Irvine, CA: University of California, Department of Information and Computer Science*, 1994.

[23] H. Bryan Riley Nikita Gurudath. Drowsy driving detectionby eeg analysis using wavelet transform and k-means clustering. *Procedia Computer Science*, 34:400–409, 2014.

[24] Daniel J. Power. Decision support systems: Concepts and resources for managers. *Greenwood Publishing Group*, page 251, 2002.

[25] Daniel J. Power. Decision support systems: a historical overview. *Handbook on Decision Support Systems 1*, pages 166–185, 2008.

[26] N. V. Natasha Sriraam and H. Kaur. Data mining techniques and medical decision making for urological dysfunction. *In Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, pages 2506–2516, 2008.

[27] Wenwei Yu U. Rajendra Acharya. Data mining techniques in medical informatics.

[28] H. Verma and R. K. Agrawal. Intuitionist gustafson-kessel algorithm for segmentation of mri brian image. *Advances in Intelligent and Soft Computing*, 131:133–144.

[29] Vipin Kumar J. Ross Quinlan Joydeep Ghosh Qiang Yang Hiroshi Motoda Geoffrey J. McLachlan Angus Ng Bing Liu Philip S. Yu Zhi-Hua Zhou Michael Steinbach David J. Hand Dan Steinberg XindongWu. Top 10 algorithms in data mining. *Springer-Verlag London*, pages 1–37, 2007.