SEP **SES** **TNM**

INSTITUTO TECNOLÓGICO DE CHIHUAHUA II

DETECCIÓN DE REGISTROS DUPLICADOS DE PACIENTES EN MÚLTIPLES FUENTES DE INFORMACIÓN

TESIS
PARA OBTENER EL GRADO DE

MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA

JOEL ADÁN SALDAÑA VILLALBA

DIRECTOR DE TESIS
M.I.S.C. JESÚS ARTURO
ALVARDO GRANADINO

CO-DIRECTOR DE TESIS
DR. HERNÁN DE LA GARZA
GUTIÉRREZ

CHIHUAHUA, CHIH. A DICIEMBRE DE 2017

Dictamen

Chihuahua, Chih., 14 de diciembre del 2017

LIC. OLGA REBECA CASTILLO CRUZ
JEFA DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
Presente.-

Por medio de este conducto el comité tutorial revisor de la tesis para obtención de grado de Maestro en Sistemas Computacionales, que lleva por nombre **"DETECCIÓN DE REGISTROS DUPLICADOS DE PACIENTES EN MÚLTIPLES FUENTES DE INFORMACIÓN"**, que presenta el (la) **C. JOEL ADÁN SALDAÑA VILLALBA**, hace de su conocimiento que después de ser revisado ha dictaminado la **APROBACIÓN** del mismo.

Sin otro particular de momento, queda de Usted.

Atentamente

La Comisión de Revisión de Tesis.

A. Alvarado G.

M.I.S.C. Arturo Alvarado Granadino
Director de Tesis

Dr. Hernán de la Garza Gutiérrez
Co-Director

M.C. Rafael Sandoval Rodríguez
Revisor

Leonardo Nevárez Chávez

M.C. Leonardo Nevárez Chávez
Revisor

SECRETARIA DE
EDUCACION PUBLICA
INSTITUTO TECNOLOGICO
DE CHIHUAHUA II
DIVISION DE ESTUDIOS DE
POSGRADO E INVESTIGACION

RESUMEN

En busca de mejorar la atención a sus pacientes, los Servicios de Salud del Estado de Chihuahua han decidido implementar un sistema de gestión de los pacientes de forma centralizada, con el cual las personas pueden acudir entre hospitales y clínicas basándose en un solo Expediente Clínico Electrónico, por medio del cual podrán ser atendidos con su historial clínico completo sin necesidad de cargar con él. Para garantizar esto se contempla un modelo de integración de los datos que valide que cada expediente sea único para cada habitante del estado.

Para llegar al objetivo de un crear un sistema de gestión médica, tenemos que manejar un gran volumen de información, la cual está seriamente ligada a una serie de datos específicos denominados “datos personales del paciente”, los cuales tienden a ser generados en más de una ocasión, ocasionando serios problemas de lentitud en el sistema y de espacio en las bases de datos sumándose al hecho de que la información duplicada n-veces elimina el objetivo de tener un registro único por paciente.

La lógica difusa es una herramienta muy usada para eliminar la repetición de registros en información personal, donde los ingresos de la misma están propensos a contener uno o más caracteres faltantes o sobrantes.

ABSTRACT

On the search to improve assistance, Servicios de Salud del Estado de Chihuahua has decided to implement a patient administration system on a centralized way, which will allow people to go between hospitals and clinics using one electronic clinical record, thru which patients will be assisted based on their medical history without having to carry it every time. To guarantee this, an integration data model is being contemplated to validate every record to be unique for each citizen.

To reach the goal of creating a medical administration system, we have to handle a big information volume, which will be directly linked to a specific data series called “patient personal data”, this data tends to be generated in more than one occasion causing serious problems of slowliness in the system and space issues on the data bases in addition to the fact that n-times duplicated information eliminates the goal of having one unique record per patient.

Fuzzy logic is a very commonly used tool to remove the repetition of records in personal information where their entries are prone to have more than one character extra or missing.

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CONTENIDO

I) INTRODUCCIÓN

1.1 Introducción al proyecto

Con la llegada de las Tecnologías de la Información, la humanidad buscó la manera de agilizar los procesos que mayor tiempo le consumían, junto aquellos donde la precisión es importante. Este hecho representó la computarización de procesos con diferentes grados de complejidad, disminuyendo considerablemente las actividades que realizaba una persona para completar una sola tarea, mientras se aumentaba la calidad y disminuía el tiempo de la misma.

Tal es el caso de los sistemas hospitalarios para llevar el registro de los pacientes, donde según las políticas de cada institución se podían ir generando expedientes en papel por cada área por la que fuera pasando el paciente. Esto empezó a cambiar con los programas informáticos, los cuales permitieron un acceso más rápido de los pacientes a las diferentes áreas, pero persistiendo el hecho de tener que ser ingresados a varios sistemas en la misma institución.

Para resolver este problema se diseñó un tipo de sistema enfocado a la información dentro de las bases de datos, denominado como Índice Maestro de Pacientes (MPI, por sus siglas en inglés), el cual se encarga de enlazar los registros pertenecientes a los mismos pacientes en los diferentes sistemas por medio de un identificador único, manipulando así a todos los pacientes registrados dentro la organización como un ente único sin duplicados.

Motivo por lo cual, los MPI son usados especialmente en instituciones de salud para mantener la integridad y veracidad de los datos personales de sus pacientes, evitando con ello posibles fragmentos en los expedientes, mientras se facilita el acceso y la transición entre departamentos.

Lamentablemente los registros duplicados siguen existiendo a pesar de los esfuerzos por erradicarlos. Las grandes cantidades de información manejada por las instituciones médicas

y el flujo constante de pacientes las 24 horas del día obstaculizan el llevar un registro integro de la información. La agilización de los trámites y acceso de los pacientes a las diferentes áreas de las unidades de salud, ocasiona en la mayoría de los casos registros incompletos por medio de los cuales es difícil la distinción entre los pacientes ya registrados y los de nuevo ingreso, propiciando la duplicación de los mismos.

Los MPI suelen llevar la búsqueda de pacientes duplicados por medio de una o varias herramientas de la Inteligencia Artificial como son las Redes Neuronales, los Algoritmos Inteligentes, la Lógica Difusa y Árboles de Decisión entre otras. De estas técnicas, la Lógica Difusa es una de las más usadas para encontrar las similitudes entre personas debido a la facilidad que tiene para computarizar el razonamiento aproximado que se utiliza en el mundo real.

Esta rama de la Inteligencia Artificial permite dar valores exactos a frases ambiguas o imprecisas como “Juan es alto”, “hace mucho calor”, etc. Estas frases son capaces de generar un pensamiento diferente tanto para el emisor como para el receptor, dependiendo de sus experiencias de vida. Con base en ello, podemos definir que la lógica difusa ayuda a simular el funcionamiento del cerebro humano dentro de una computadora.

1.2 Problemática

Los Servicios de Salud en el estado de Chihuahua distribuyen su atención a la población por medio de los organismos de Seguro Popular (SP) y del Instituto Chihuahuense de la Salud (ICHISAL), además del padrón de población abierta (PA), el cual sirve para atender al sector de la población que no cuenta con afiliación a ninguna de las entidades mencionadas.

Para llevar el control de los expedientes pertenecientes a la población, los Servicios de Salud cuentan con varios sistemas clínicos, como son el Sistema Hospitalario (SIHO) y el Sistema de Información para la Gerencia Hospitalaria (SIGHO) repartidos en los diferentes centros de salud pertenecientes a los diferentes organismos con los que cuenta para dar atención a la creciente población estatal.

Estos sistemas trabajan de manera autónoma, por lo que cada una de las unidades médicas cuenta con su propia versión del sistema que utiliza y sin contar con conexión hacia los otros centros de salud, generando un aislamiento de información entre unidades médicas y una falta de comunicación entre los diferentes niveles de atención médica existentes, lo que lleva a producir como mínimo una tabla de pacientes por cada uno de ellos y en algunos casos hasta 2 tablas, diferenciando una tabla para pacientes propios al padrón y otra para pacientes externos al mismo.

A esto hay que sumarle que no siempre un paciente se atiende en el centro de salud que le corresponda, hecho que ha ocasionado que una misma persona posea varios registros de expedientes clínicos tanto físicos como electrónicos en diferentes establecimientos de salud, ocasionando que el volumen de expedientes sea tal, que dejar un único expediente por paciente sin ayuda tecnológica resulte ser una tarea titánica y por demás poco confiable, es por eso, que debido a las labores que desarrollo en la Subdirección de Tecnologías de la Información y su correlación con el proyecto de un Expediente Clínico Electrónico, se me pidió dar una solución a la problemática de la duplicación de expedientes. Por este motivo se plantea la creación de un proyecto que permita mediante el uso de la lógica difusa detectar en primer lugar los duplicados obvios y en segundo lugar mostrar los candidatos a duplicados con un índice de certeza que facilite que los usuarios del sistema a diseñar, puedan consolidar varios expedientes de un mismo paciente en uno solo para formar un MPI.

El logro de este objetivo, representará grandes beneficios para el paciente en cuestiones de tiempo de atención y costos, mientras las instituciones se verán beneficiadas tanto en aspectos económicos como en calidad de atención y mejores diagnósticos médicos al tener acceso a los expedientes del paciente pertenecientes a otras instituciones de los mismos Servicios de Salud, con lo que se podrá tener registros reales de las cantidades de la población que acuden por centro de salud, permitiendo obtener estadísticas exactas con las cuales poder tomar mejores decisiones para el manejo de los recursos y consecuentemente tomar las previsiones logísticas necesarias.

1.3 Alcances y limitaciones

1.3.1 Alcances

- El sistema se conectará a las diferentes bases pertenecientes a los diferentes padrones para obtener los registros sobre los cuales se implementará el controlador difuso. Por ejemplo: el nombre, póliza, fecha de nacimiento de un paciente que pertenezca al Seguro Popular, para poder comparar a dicha persona con pacientes de otros padrones.
- El Índice Maestro de Pacientes (MPI), Si bien, como se mencionó con anterioridad se puede crear con varias herramientas de la Inteligencia Artificial, se desea probar que se puede crear uno utilizando solamente la lógica difusa.

1.3.2 Limitaciones

- Las actualizaciones constantes de las tablas no permiten llegar a un punto donde el sistema pueda comprobar el 100% del contenido en las mismas, por lo que se plantea el hecho de ejecutar la verificación de manera mensual, el cual incluso, es el tiempo en el que el SP envía sus tablas actualizadas y verificadas para su uso a los diferentes centros de salud.
- El controlador difuso solo será capaz de sugerir pacientes con similitudes pero nunca ligarlos. Este procedimiento de enlazar los registros será realizado por una persona con mayor conocimiento sobre los pacientes para disminuir los posibles casos de expedientes médicos con faltas al mismo, o posibles anexos de historiales médicos pertenecientes a otros pacientes.

1.4 Justificación

En el estado de Chihuahua existen 329 centros de salud, de los cuales solamente 31 se encuentran en la capital, y para el año 2017, en todos ellos se empezará a pedir la Clave Única de Registro de Población (CURP) como medio obligatorio de identificación, conforme a lo estipulado en los cambios presentados a las reformas de salud, debido a lo cual, los Servicios de Salud se vieron en la necesidad de evaluar los sistemas con los que cuentan actualmente, encontrando que los mismos no se usan adecuadamente, principalmente en el ingreso de los pacientes a los centros de salud, generando varias

inconsistencias como son la duplicación de los mismos junto a datos incompletos o erróneos en sus registros. Estas inconsistencias han sido causadas en mayor medida por los siguientes motivos:

- Sistemas como el SIGHO cuentan con validaciones para evitar la duplicación de registros de pacientes, las cuales son fácilmente burladas por los usuarios, esto debido a la alta demanda que se tiene a la hora de registrar a los pacientes, donde resulta más rápido cambiar el sexo de un paciente y duplicarlo, que seleccionar cuál de los pacientes ya registrados está solicitando su reingreso.
- Por causa de varios factores como falta de personal, falta de equipo médico o simplemente que el paciente se encuentre en un área geográfica diferente a la del centro que está afiliado, es necesario que acudan a una unidad médica diferente, dividiendo así, su expediente clínico en varias unidades médicas.
- Desconocimientos, prisas y/o falta de atención a la hora de ingresar a los pacientes han causado que los registros sean incorporados con caracteres numéricos o especiales dentro de los nombres propios, con faltas de ortografía e incluso con datos incompletos en los que usualmente se incluyen RFC y CURP, dificultando con ello que sean localizados cuando regresan al centro de salud y propiciando un registro duplicado para el mismo paciente.

Estas inconsistencias repercuten en costos como los siguientes:

- Los médicos carecen de elementos confiables y oportunos para la toma de decisiones referentes al tratamiento de los pacientes.
- Aumenta la posibilidad de duplicar estudios, medicamentos y recursos en general afectando a instituciones y pacientes por igual.

Este tipo de problemáticas se pueden disminuir considerablemente llevando un expediente clínico único centralizado que organice a la población estatal por igual.

1.5 Objetivo

Desarrollar un Índice Maestro de Pacientes utilizando solamente procedimientos de Lógica Difusa para detectar a los pacientes duplicados en los diferentes sistemas que actualmente son usados por los Servicios de Salud del Estado de Chihuahua. Con la finalidad de unificar sus registros y enlazar sus expedientes médicos en uno solo, creando así un Expediente Clínico Electrónico único para la población del estado perteneciente a uno de los diferentes padrones como son: Seguro Popular, ICHISAL y/o Población Abierta.

II) ESTADO DEL ARTE

2.1 Sistemas difusos en los MPI

Los sistemas de control difusos se han posicionado dentro de una gran parte de nuestras actividades diarias sin que nos diéramos cuenta de ello, debido principalmente a su gran facilidad de uso y versatilidad para afrontar los problemas cotidianos. Un ejemplo conciso de ello es dentro de la Fotografía, en la que los sistemas difusos pueden ayudar con el enfoque, el uso del flash y hasta disminuir el movimiento de la cámara para tomar mejores fotografías; si se tiene en cuenta que cada día la cantidad de personas que toman una foto aumenta, utilizando cámaras especializadas o celulares, se tiene un gran mercado para los controles difusos dentro de la Fotografía.

La perspectiva de los sistemas difusos para afrontar las situaciones de manera semejante a como lo haría un experto en el área, pero con una velocidad de procesamiento proporcionada por un sistema informático, los hacen perfectos para una gran cantidad de labores dentro y fuera de la industria. Motivo por lo que son ampliamente usados en actividades industriales como el control de la temperatura en calderas, así como el procesado de imágenes y reconocimiento de caracteres; en tareas del hogar se utiliza en el lavado de ropa, donde la lavadora puede decidir la cantidad de agua que entra, en qué momento y qué cantidad de detergente es introducido al sistema; mientras en actividades de oficina se pueden usar para controlar el ascenso y descenso de elevadores, etc.

El controlador difuso más famoso fue desarrollado por la compañía Hitachi en Japón alrededor de 1983 y es utilizado por el metro de la ciudad de Sendai desde 1987 para controlar la velocidad y el frenado de los trenes (Mendel, et al, 2014). Sus desarrolladores lograron pasar el conocimiento y las técnicas de manejo de un conductor experto a un algoritmo al que denominaron Automatic Train Operation (ATO), en la Figura 2.1 se muestra un diagrama de flujo representando el conocimiento de los conductores sobre el frenado. Las pruebas realizadas mostraron un manejo suave de los trenes en el que los pasajeros difícilmente notan los cambios de velocidad, proporcionando un frenado suave mientras se mantiene el confort de los viajantes (Yasunobu, et al, 2002).

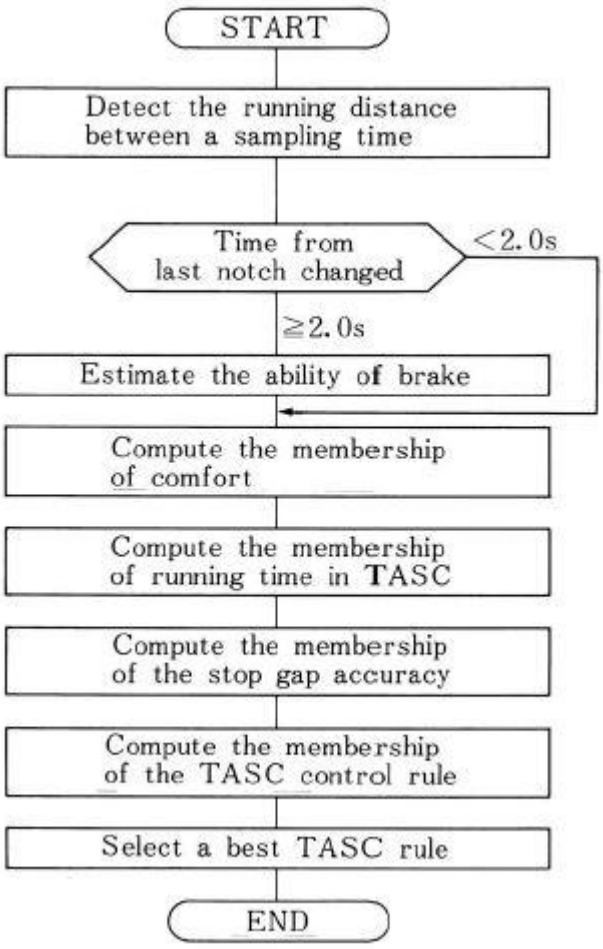


Figura 2.1. Diagrama de flujo del controlador difuso del frenado de los trenes.

En el campo de la medicina se cuenta con aplicaciones que ayudan a la detección de enfermedades como la Leucemia Linfoblástica Aguda (LLA), el cual es el cáncer infantil más común y consiste en que la médula ósea produce una cantidad muy grande de linfocitos (glóbulos blancos) inmaduros por lo que no pueden combatir las enfermedades eficazmente. El algoritmo para la detección de LLA (Ordaz, 2011) consta de cuatro etapas: Imagen de entrada, Segmentación, Clasificación, Reconocimientos y Salida. Donde al final del proceso su autor menciona tener una sensibilidad del 98% con un error del 3.3% usando la lógica difusa, mientras sin el uso de la misma, la sensibilidad disminuye hasta un 86% y el grado de error aumenta a 22.2%.

De manera semejante, los controladores difusos se encuentran en los MPI desde antes del 2004, compitiendo fuertemente en el diseño de los mismos contra sistemas determinísticos y sistemas probabilísticos, con los que usualmente son confundidos. En el artículo "Patient Data Matching Software: A Buyer's Guide for the Budget Conscious" (Sujansky y Jones, 2004) se hace una comparativa de cuatro herramientas de MPI en donde cada una de ellas cuenta con diferentes herramientas para encontrar información duplicada, y entre ellas está una con controladores difusos.

2.2 Los Master Patient Index en México

Existen una gran variedad de programas, como el Enterprise Master Patient Index, que forma parte de la plataforma ehCOS para instituciones de salud, dicho paquete de programas ha sido diseñado para gestionar todos los procesos administrativos y hospitalarios de los centros de salud de una manera amigable para el usuario, entre ellos un MPI capaz de compartir los datos de los pacientes entre las diferentes áreas de una manera gráfica y sencilla para el usuario. De manera similar trabajan el Oracle Healthcare Master Person Index, el Sun Master Person Index, este tipo de sistemas funcionan basándose en herramientas como las redes neuronales, la lógica difusa, arboles de decisión, así como algoritmos inteligentes, algoritmos fonéticos y algoritmos determinantes entre otras.

Motivos como altos costos y falta de adaptación a las necesidades de las instituciones médicas, han propiciado el desarrollo de varios proyectos de investigación para solventar la problemática de la falta de un MPI adecuado a ellas. Sistemas como el de Detección de Personas Duplicadas en un Sistema de Gestión de Información Médica (Soberats, 2011), desarrollado por el Centro de Informática Médica en La Habana, Cuba, es un programa desarrollado en Java durante el 2011 como un módulo para el Sistema de Información Hospitalaria denominado alas HIS. Se basa en búsquedas difusas para procesar los resultados por medio de algoritmos inteligentes, redes neuronales y árboles de decisión.

Asimismo en el artículo titulado "Desarrollo de un Índice Maestro de Pacientes Utilizando Estándares y Software Open Source" (Martínez, et al, 2015), se describe del desarrollo de un Índice Maestro de Pacientes que se utiliza actualmente en el Hospital

Regional Antonio J. Scaravelli de la provincia de Mendoza en la República Argentina. Este MPI utiliza Hibernate Search para indexar las bases de datos y permitir búsquedas de texto completo directamente sobre las mismas, para lo cual se utiliza el algoritmo fonético Double Methafone, basado en sus resultados aplica una técnica para elegir una cantidad de resultados los cuales son comparados con el algoritmo de Jaro-Winkler. Una vez comparados los registros una persona identificará los duplicados y los enviará a una segunda persona para su vinculación.

Mientras en México uno de los pocos sistemas creados es el Sistema Evaluador de Calidad de Datos en MySQL (Ortiz, 2015), el cual fue desarrollado por la Facultad de Ingeniería de la UNAM y trabaja indexando los registros con base en algoritmos fonéticos, para luego realizar búsquedas por medio de algoritmos de distancia y vincularlos por medio del sistema FEBRL.

Por parte de los Servicios de Salud del Estado de Chihuahua, este proyecto es el primer avance en tener un MPI para la realización de sus labores.

III) MARCO TEORICO

3.1 Lógica Difusa

La lógica difusa o borrosa, tiene sus orígenes en las filosofías de la antigua Grecia, en las cuales se creía en la existencia de varios grados de pertenencia o veracidad de las cosas de acuerdo a los pensadores sofistas de los tiempos. El término de Lógica Difusa fue finalmente definido por quien se le conoce como el padre de misma, el Dr. Lotfi Zadeh en 1965 y quien la definió como: “Conforme la complejidad de un sistema aumenta, nuestra capacidad para ser precisos y construir instrucciones sobre su comportamiento disminuye hasta el umbral más allá del cual, la precisión y el significado son características excluyentes”.

Bajo este concepto Zadeh representó la forma de expresar y pensar de un ser humano para afrontar los problemas y las decisiones que enfrenta en su vida diaria, en las que se expresa de una manera ambigua con frases del tipo: “Juan conduce muy rápido”, “José es alto”, “Me gusta la comida bien caliente”, donde las palabras “muy”, “es” y “bien” sirven de cuantificadores a las expresiones para que el interlocutor pueda evaluarlas basándose en los conocimientos que ha adquirido sobre las variables lingüísticas “rápido”, “alto” y “caliente”.

Bajo la lógica difusa las variables lingüísticas son aquellas palabras (verbos, adjetivos o etiquetas lingüísticas) que demuestran un grado de incertidumbre o membresía del sujeto con el cual vamos a trabajar hacia un adjetivo, mientras los valores o etiquetas lingüísticas son aquellas clasificaciones en las que podemos dividir a las variables lingüísticas, dígase bajo, normal y alto como posibles valores de una variable denominada “altura de una persona”. El hecho de que la variable “altura de una persona” pueda caer en cada uno de los valores según la interpretación de cada oyente, convierte a la lógica difusa en una lógica multivalor, en la que la variable puede pertenecer a cada uno de los diferentes conjuntos difusos con un valor de pertenencia distinto, con lo que se da a entender, la lógica difusa funciona como una extensión a la lógica clásica donde sólo se permiten el pertenecer

a uno de los extremos de la verdad, dígase falso o verdadero, negro o blanco. La Figura 3.1 muestra el esquema general de un controlador difuso.

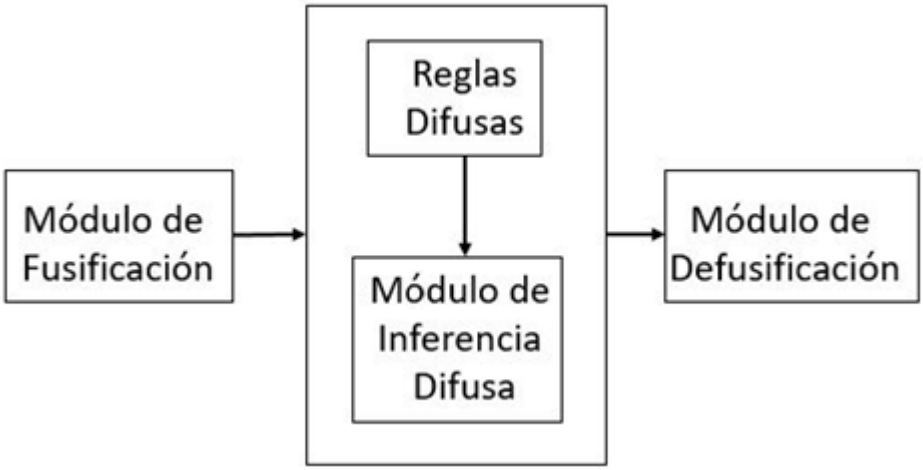


Figura 3.1. Esquema general de un controlador difuso.

3.1.1 Fusificación

La función más importante sobre las variables lingüísticas es la fusificación, la cual se define como el proceso de pasar los valores reales de las variables a valores difusos, para ser usados en los conjuntos difusos, lo que lo convierte en un paso de suma importancia durante el desarrollo del sistema (Barros, et al, 2017).

Los conjuntos difusos ya mencionados fueron desarrollados por Zadeh con la finalidad de representar de manera cuantitativa el conocimiento lingüístico, el cual funciona de manera cualitativa, lo que permite a los conjuntos contener elementos de manera parcial, es decir que para un elemento x, donde $x \epsilon A$, el valor de x se medirá por medio de un número real comprendido entre 0 y 1, incluyéndolos, bajo la función $\mu_A (x)$: $X \rightarrow [0,1]$, hecho que permitió a Zadeh inferir que la lógica difusa es muy precisa a pesar de trabajar con información imprecisa.

De esta manera la lógica difusa permite que cada variable evaluada pertenezca simultáneamente a uno o varios conjuntos a la vez, difiriendo el valor o grado de pertenencia

de la misma hacia ellos, razón por la cual se suele referir a los conjuntos difusos como subconjuntos difusos. El valor de pertenencia puede ser uno de los siguientes casos.

- $\mu_A(x) = 0$, se considera que x no pertenece al conjunto.
- $0 < \mu_A(x) < 1$, x pertenece de manera parcial al conjunto.
- $\mu_A(x) = 1$, x pertenece totalmente al conjunto.

La pertenencia de cada variable se define por una función de membresía propia (μ_A) del conjunto hacia el universo caminando por una transición gradual de 0 a 1, al contrario de la lógica clásica donde el cambio se produce de manera instantánea entre los valores. La función representa gráficamente al conjunto difuso, y dependerá de la manera en que se enfrentará al problema. Dentro de las principales formas geométricas con las que cuentan, las siguientes fueron utilizadas en el proyecto.

- Función Pi o Trapezoidal.- Es la más común debido a la simplicidad de uso y facilidad para representar una gran cantidad de valores intermedios dentro del universo. Está definida mediante 4 variables a, b, c y d, donde cada variable representa un cambio de movimiento de los valores de la misma.

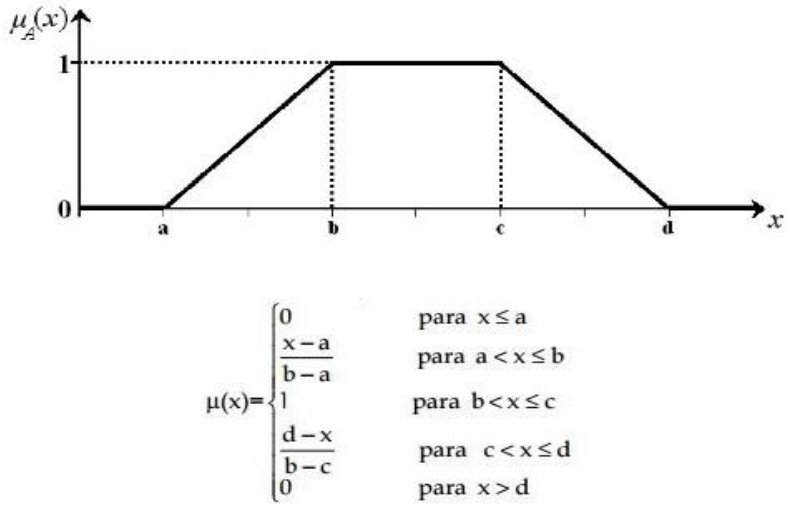
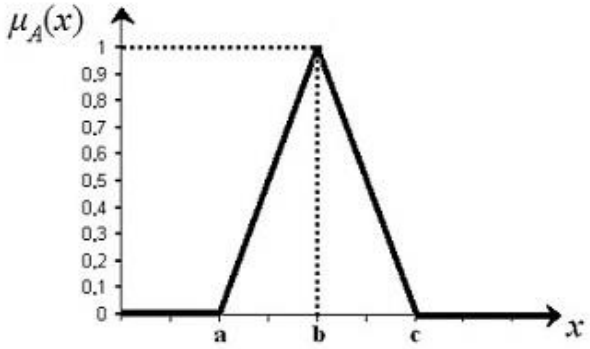


Figura 3.2. Función difusa Trapezoidal.

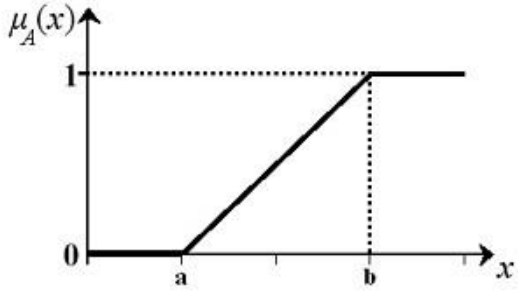
- Función Lambda o triangular.- Al igual que la función trapezoidal, se utiliza para obtener valores difusos intermedios dentro del universo, su diferencia radica en los puntos de pertenencia al universo, mientras Pi cuenta con un rango de valores para representar la membresía total al conjunto, la función lambda soló cuenta con un punto de inferencia y se define mediante 3 variables.



$$\mu(x)=\begin{cases}0 & \text{para } x \le a \\ \dfrac{x-a}{m-a} & \text{para } a < x \le m \\ \dfrac{b-x}{b-m} & \text{para } m < x \le b \\ 0 & \text{para } x > b\end{cases}$$

Figura 3.3. Función triangular difusa.

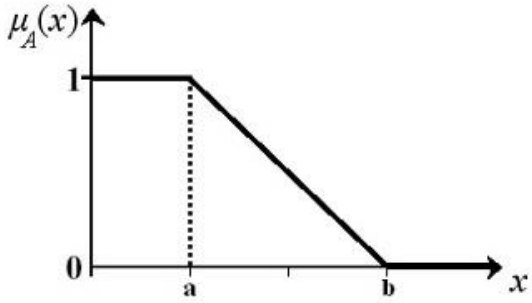
- Función Gamma o rampa mayor que.- Semejante en diseño y facilidad a la función trapezoidal, la función gamma se usa para especificar valores al final del universo mediante 3 variables, se caracteriza por el crecimiento acelerado conforme más cerca se encuentra a de b.



$$\mu_A(x) = \begin{cases} 0 & \text{para } x \le a \\ \dfrac{x-a}{b-a} & \text{para } a < x \le b \\ 1 & \text{para } x > b \end{cases}$$

Figura 3.4. Función mayor que.

- Función L o rampa menor que.- Conocida como 1 – función gamma, la función L representa valores al inicio del universo.



$$\mu_A(x) = \begin{cases} 1 & \text{para } x \le a \\ \dfrac{b-x}{b-a} & \text{para } a < x \le b \\ 0 & \text{para } x > b \end{cases}$$

Figura 3.5. Función menor que.

La sencillez de estas funciones no implica el perder exactitud a la hora de realizar los cálculos necesarios. Una serie de conjuntos bien ubicados dentro del universo y siguiendo unas sencillas reglas como, “abarcar todos los puntos del universo” o “la suma de los puntos donde se intersectan 2 conjuntos no debe ser mayor a 1” proporcionan un sistema difuso

muy robusto capaz de solventar la mayoría de las situaciones. En la siguiente figura se muestra en su totalidad un conjunto difuso, aplicando los términos vistos al universo de una persona alta.

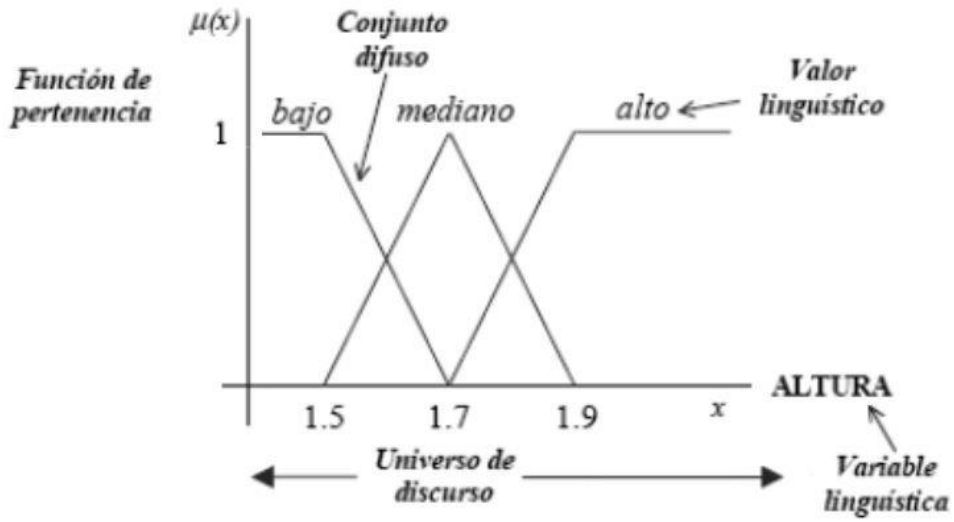


Figura 3.6. Ejemplo de un conjunto difuso.

3.1.2 Reglas de Control Difuso

Las salidas obtenidas de los conjuntos difusos son evaluadas por una serie de reglas difusas del tipo SI-ENTONCES, por las cuales se decidirá el proceso a seguir, ya sea una siguiente iteración por los mismos u otros conjuntos difusos, o bien, la terminación del proceso.

Figura 3.7. Ejemplo de una regla difusa con salida a otro conjunto difuso.

Estas reglas se componen de un antecedente, el cual puede ser una o más expresiones lógicas conectadas por los operadores Y, O o NO y de un consecuente, como se muestra en la siguiente figura.

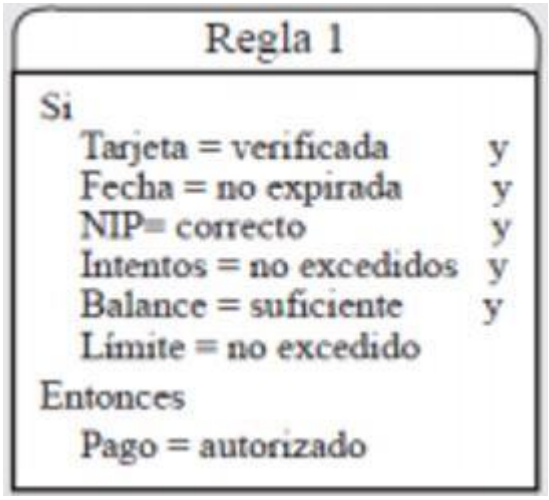


Figura 3.8. Regla difusa con antecedentes compuestos.

Las reglas son gestionadas por el motor de inferencia, el cual se encarga de ordenar, insertar, modificar y/o eliminar las reglas para poder relacionar los conjuntos de entrada y salida con lo que se generará el conocimiento esperado, donde, si la base de conocimientos (reglas del control difuso difuso) cuenta con inconsistencias, el sistema se comportará de manera errática.

3.1.3 Defusificación

La salida obtenida durante el proceso de inferencia, es el área resultante de los conjuntos difusos donde estos se sobreponen unos a otros, proveniente de la aplicación de las diferentes reglas difusas, resultando en una serie de valores difusos a convertir en un solo valor real para su interpretación en el problema, a este proceso se le denomina defusificación.

El método más usado para esta conversión, es el método del centroide o centro de gravedad, en el que se obtiene el centro del área solapada como la representación real de los conjuntos difusos.

IV) DESARROLLO

4.1 Modelo de negocio

El sistema para la detección de registros duplicados de pacientes en múltiples fuentes de información, tiene como finalidad distinguir de entre los registros contenidos en las diferentes bases de datos pertenecientes a los Servicios de Salud del Estado de Chihuahua, a los que correspondan a una misma persona, con la finalidad de crear una base de datos única perteneciente a todo el estado de Chihuahua.

Por medio de las pláticas con el subdirector del departamento Tecnologías de la Información, oficina encargada del proyecto, se optó por dividir el proyecto en 2 partes principales, la primera de ellas es la búsqueda de registros de pacientes duplicados por medio de Lógica Difusa, método sobre el cual se enfoca este documento de tesis; la segunda parte trata de la interfaz gráfica por la cual una persona podrá determinar si realmente son la misma persona o solo contienen similitudes entre sí (Anexo A).

Como plan de desarrollo se tomó de base el modelo en cascada (Pressman, 2010), el cual permite avanzar en forma lineal por cada una de las fases de desarrollo de manera rigurosa. Este enfoque sugiere avanzar de forma secuencial y sistemática por las diferentes etapas empezando por la determinación de los requerimientos, continuando por la planeación, modelado, construcción y despliegue del proyecto (Figura 4.1) para poder entregar un producto terminado y en funcionamiento.

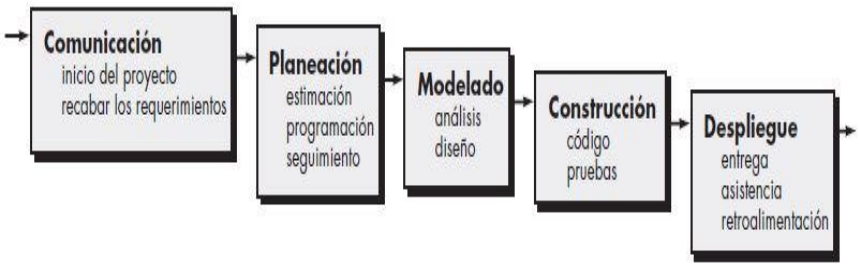


Figura 4.1. Etapas del modelo en cascada.

Con base en las mismas pláticas y al esquema de trabajo elegido, se definieron los objetivos de cada etapa como se muestran en el diagrama siguiente.

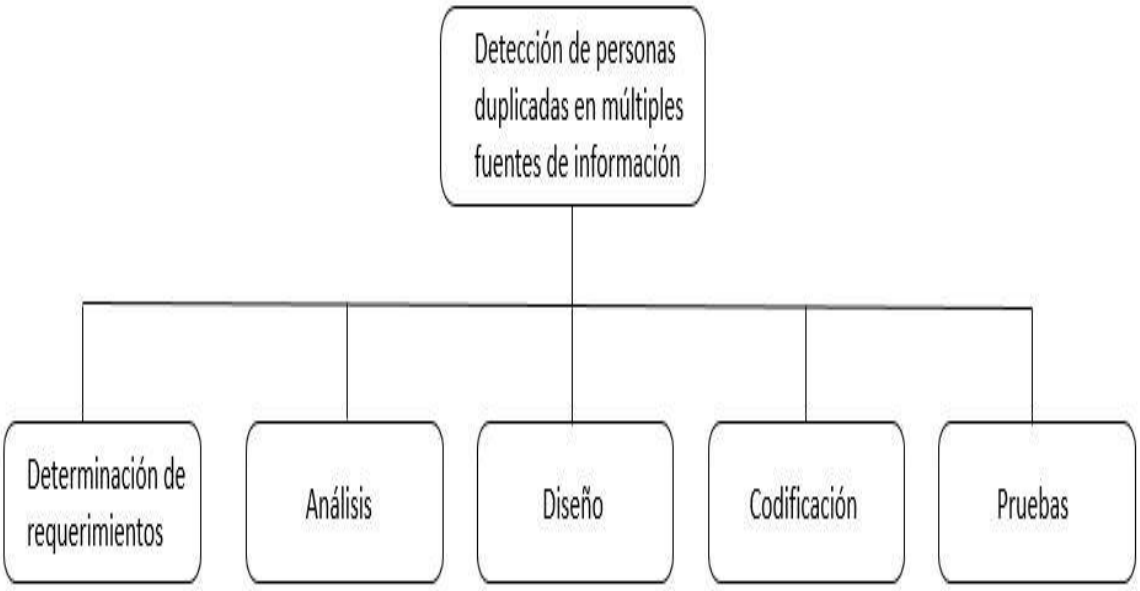


Figura 4.2. Diagrama de objetivos.

4.2 Determinación de requerimientos

De la etapa anterior se organiza la información obtenida para obtener los puntos clave que debe tener el software, así como la metodología a seguir para su desarrollo y los objetivos generales que se deben cumplir según lo requerido por los Servicios de Salud del Estado de Chihuahua.

Basados en la revisión de la información, y las pláticas con el personal de la Subdirección, se definieron una serie de puntos fundamentales a cubrir durante el desarrollo del software:

- Los datos de los pacientes se encuentran distribuidos entre cada una de las instituciones de salud con las que cuentan los Servicios de Salud del Estado de Chihuahua, una conexión estable y recuperación de caídas a las mismas, además de un control sobre las bases de datos ya cotejadas constituyen uno

de los requisitos más indispensables para que este sistema pueda cumplir su principal función.

- El éxito de todo sistema depende en gran medida de un buen funcionamiento, un procedimiento adecuado de comparación entre pacientes se vuelve indispensable para la aprobación del mismo. Este método debe ser capaz de solventar las diferentes situaciones de incongruencias que se encuentre en los datos personales a la hora de comparar los mismos con los de otro paciente.
- El sector poblacional es un factor de crecimiento constante en cada una de las instituciones de salud, es por ello que el sistema debe de contar con un control sobre los pacientes aun después de comparados y sin importar su resultado final, con el fin de llegar a un verdadero expediente clínico único por paciente.
- Uno de los requerimientos más solicitados es una pantalla de decisión final, en la cual una persona podrá catalogar a un registro como duplicado o no, esto con el fin de asegurar un porcentaje mayor de efectividad y evitar posibles problemas legales que pudieran surgir por una mal práctica médica ocasionada por una mala relación de los expedientes clínicos.
- Una interfaz intuitiva y fácil de usar no sólo aumenta el éxito y desempeño del sistema, también ayuda a los usuarios a estar más concentrados en su trabajo y disminuir los posibles errores de uso, por ello es de suma importancia una serie de pantallas que no sean desgastantes al usuario y que incluyan comandos representativos de sus funciones.

En el siguiente diagrama de casos de uso, se muestran los puntos mostrados anteriormente de acuerdo a la funcionalidad que se espera del sistema.

Figura 4.3. Diagrama de casos de uso.

De igual manera, durante las pláticas se nos proporcionó la siguiente imagen (Figura 4.4), la cual representa la idea que tiene el personal de los Servicios de Salud para la pantalla donde trabajarán los usuarios finales y determinarán el proceso final pertinente a cada registro evaluado.

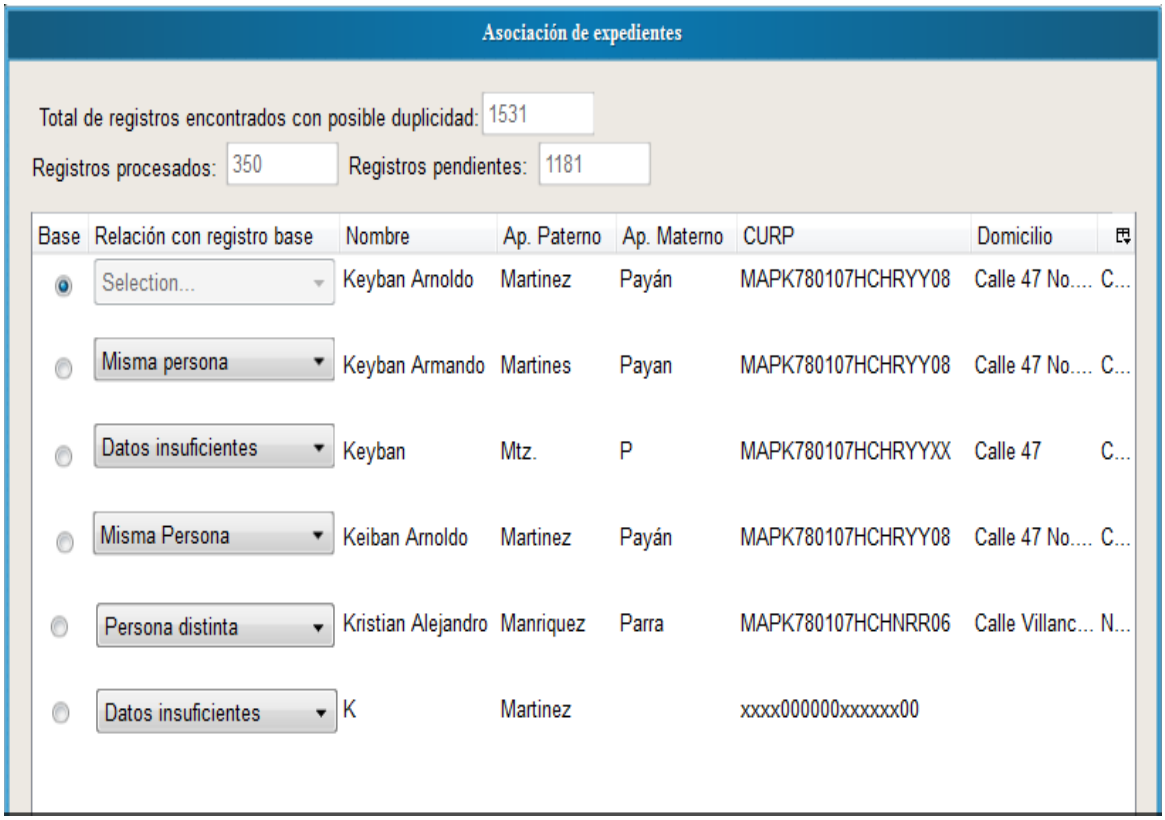


Figura 4.4. Boceto de la pantalla final.

4.3 Análisis

Una de las principales funciones de toda unidad médica, es proporcionar atención médica a quien lo solicite, sin importar el padrón con el que cuente la persona, motivo por el cual, estas personas deben estar registradas en dicha unidad. La siguiente tabla muestra los padrones a los que una persona puede pertenecer.

Tabla 4.1. Padrones de Afiliación.

Nombre	Descripción
SP	Seguro Popular.
ICHISAL	Instituto Chihuahuense de la Salud.
PUPA	Padrón Único de Población Abierta.

Durante la etapa de análisis podemos encontrar los posibles problemas que pueden existir a la hora de desarrollar el proyecto, y de ahí partir a una posible solución de los mismos. Uno de los aspectos más sobresalientes a tomar en cuenta y que debe solventar el

sistema difuso, es el origen de la información a comparar, misma que se encuentra distribuida en los diferentes sistemas informáticos con los que se cuenta para llevar el registro de los pacientes en cada unidad médica (ver tabla 4.2).

Tabla 4.2. Bases de datos a comparar.

Sistema o padrón	Descripción	Gestor de BD
ICHISAL	Instituto Chihuahuense de la Salud.	ORACLE
SIHO	Sistema Hospitalario.	ORACLE
SIGHO	Sistema Integral de Gestión Hospitalaria.	SQL Server

Cada uno de estos sistemas cuenta con bases de datos separadas y sin conexión a las demás (en cada una de las unidades médicas en las que son utilizados), en la tabla 4.3 se desglosan las unidades en las que se utiliza el sistema SIHO.

Tabla 4.3. Unidades Médicas que utilizan el sistema SIHO.

Clave	Descripción
1	HOSPITAL CENTRAL.
2	HOSPITAL INFANTIL.
3	HOSPITAL GENERAL DE JUÁREZ.
4	HOSPITAL HOSAME.
5	HOSPITAL DE DELICIAS.
6	HOSPITAL DE LA MUJER DE CUAUHTEMÓC.
7	HOSPITAL DE JIMÉNEZ.
8	HOSPITAL DE PARRAL.
9	HOSPITAL DE LA MUJER JUÁREZ.

Además de dicha fragmentación en la información, los mismos sistemas pueden aceptar más de un registro correspondiente a la misma persona y el ingreso de pacientes pertenecientes a otros padrones, generando una aglomeración de expedientes dentro de cada una de las bases de datos diferentes aumentando la dificultad para su agrupación. Como ejemplificación de ello, se tiene en la Figura 4.5 un listado TOP de los pacientes más repetidos en el Hospital Central de Chihuahua, agrupados por Apellido Paterno, Apellido Materno y Nombre.

APATERNO	AMATERNO	NOMBRE	TOTAL
MASCULINO	DESCONOCIDO	(Null)	250
GALLEGOS	GARCIA	ISABEL	235
PEREZ	ALCANTARA	FLOR ZERENA BERENICE	79
PARRA	GOMEZ	HORACIO	73
MARTINEZ	DELGADO	CLARA	62
ENRIQUEZ	BAILON	ALBERTO	50
SUAREZ	MARQUEZ	CARMEN	42
CRUZ	GUTIERREZ	IGNACIO	42

Figura 4.5. Pacientes TOP por nombre en el Hospital Central.

Mientras que en la Figura 4.6 se realiza la misma agrupación de datos pero, anexando la póliza de derechohabiencia de los pacientes a la búsqueda. Como se visualiza en la misma, algunos de los totales disminuyeron a causa de esta agrupación (la columna de póliza se omite por razones de seguridad).

APATERNO	AMATERNO	NOMBRE	TOTAL
MASCULINO	DESCONOCIDO	(Null)	249
GALLEGOS	GARCIA	ISABEL	235
PEREZ	ALCANTARA	FLOR ZERENA BERENICE	79
PARRA	GOMEZ	HORACIO	71
MARTINEZ	DELGADO	CLARA	61
ENRIQUEZ	BAILON	ALBERTO	50
SUAREZ	MARQUEZ	CARMEN	41
CRUZ	GUTIERREZ	IGNACIO	41

Figura 4.6. Pacientes TOP del Hospital Central por nombre y póliza.

Esta disminución también se puede deber por otras causas, una de ellas, la gran cantidad de nombres que se comparten entre personas, para lo cual, estudios del Instituto Nacional Electoral (INE) en abril 2015 marcan a Hernández como el apellido más común en México, siendo Hernández Hernández Juan el nombre más usado por los mexicanos, con un total de 2,943 personas denominadas de esa manera (INE, 2015), pudiendo ser esta una de las causas del por qué se encuentran registros con el mismo nombre y diferente número de póliza.

Sin embargo, la causa que podemos definir como la principal (para esta diferencia) y la que más afecta al desarrollo del sistema difuso, es la discordancia en los datos dentro de los registros, por tanto, el estado en que se encuentra la información a analizar, marca la pauta de los conjuntos difusos a diseñar.

Una serie de datos corruptos puede interrumpir el buen flujo del sistema o arrojar resultados fuera de lo esperado. Con el fin de poder definir las reglas difusas sobre las cuales van a trabajar los conjuntos difusos, se vuelve fundamental el análisis de la información en su estado actual y el posible tratado de la misma antes de ser evaluada por las reglas.

Dichas incongruencias se pueden catalogar de las siguientes maneras:

- Datos erróneos en uno o más campos.

Una revisión a los datos contenidos en las diferentes tablas arrojó una serie de errores cometidos a la hora de guardar la información. Uno de los sucesos más comunes a la hora de ingresar datos, es el escribir mal alguno de los nombres por diferentes causas como desconocimiento de la forma en que se escribe por falta de atención al escribirlos. A continuación se muestra un ejemplo de datos erróneos en la base de datos del hospital Central.

AARMENDARIZ
AMENDARIZ
AREMENDARIZ
AREMNDARIZ
ARMANDARIZ
ARMANDRIZ
ARMEDARIZ
ARMENADRIZ
ARMENDAIZ
ARMENDARIZ
ARMENDRARIZ
ARMENDRIZ
ARMENDARIS

Figura 4.7. Diferentes maneras de escribir el apellido Armendáriz.

- Caracteres especiales dentro de los nombres.

Otro caso muy común es agregar caracteres especiales o numéricos a los nombres propios. En la imagen siguiente se muestra una búsqueda por apellido paterno que arrojó varios datos con caracteres no alfabéticos del lado izquierdo, mientras que en el lado derecho se visualiza cómo quedarían los mismos datos sin dichos caracteres.

0RRANTIA	RRANTIA
0RTEGA	RTEGA
0RTIZ	RTIZ
A JURISDICCION	A JURISDICCION
A LA BREVEDAD AL	A LA BREVEDAD AL
ABAROA	ABAROA
ACEVED0	ACEVED
ACEVEDO	ACEVEDO
ACOSTA	ACOSTA
ACOSTA NO USAR	ACOSTA NO USAR
ACTIVIDAD EN EL PARQUE	ACTIVIDAD EN EL PARQUE
ADULTO MAYOR	ADULTO MAYOR
AGENDA REVISADA	AGENDA REVISADA
AGENDAR SOLO	AGENDAR SOLO
AGUILAR	AGUILAR
AGUIRRE	AGUIRRE
AGUIRRE)	AGUIRRE
AGURRE	AGURRE
AK0SIMA	AKSIMA
ALARCON	ALARCON
ALCANTAR	ALCANTAR
ALDABA	ALDABA

Figura 4.8. Búsqueda de datos con caracteres especiales.

Una variante al caso mencionado es llenar todos los campos con caracteres no alfabéticos, caso que se ilustra a continuación, donde se encontraron en la base de datos del Hospital Infantil de Chihuahua 6,016 registros, en los que los campos pertenecientes al nombre son llenados con el carácter pipe |. Afortunadamente los registros se pudieron rescatar al aparecer en el campo de domicilio los datos referentes al paciente.

Paterno	Materno	Nombre	VCHDOMICILIOTRABAJO
|	|	|	AVALOS | DE LA PAZ | CARLOS DANIEL
|	|	|	CHICO | SALAS | LUIS MISAEL
|	|	|	PRIETO | RODRIGUEZ| ERNESTO
|	|	|	ZAPATA | REYES| RAUL ENRIQUE
|	|	|	RODRIGUEZ | VILLA | FRANCISCO

Figura 4.9. Casos en los que se ingresa | en lugar del nombre.

- Sexo Incorrecto.

Un caso muy común que se encontró, es el de personas registradas con un sexo diferente al que usualmente ingresaríamos si sólo nos guiáramos por los nombres.

JAQUEZ LOPEZ MARTIN LUCIO	M
BENITEZ MARQUEZ ALMA DELIA	M
JAQUEZ BENITEZ IVAN ALONSO	M
JAQUEZ BENITEZ CINDY JOHANA	M
JAQUEZ BENITEZ LUIS CARLOS	M
SOTELO BRISEÑO LUIS GUILLERMO	M
SOTELO CHAVARRIA JOSE LUIS	M
SOTELO MORENO LUIS GERARDO	M
QUIÑONEZ CALZADILLAS JUAN MANUEL	M
SALCIDO VALDEZ ERIKA MARGARITA	M
QUIÑONEZ SALCIDO JUAN ALEJANDRO	M
LOYA OLIVAS FERNANDO	M
VAZQUEZ GUTIERREZ LAURA VERONICA	M
LOYA VAZQUEZ EVA ZULEMA	M
PORRAS ORTIZ GUILLERMO ARMANDO	M
BLANCO CASTILLO MARGARITA	M
PORRAS BLANCO GUILLERMO ARMANDO	M
LAZALDE MORALES IRMA LETICIA	F
LAZALDE ANCHONDO MARIO	F
. MORALES ROSA EMMA	F
LUJAN LAZALDE ESTIBALY LETICIA	F
LUJAN LAZALDE MELISSA	F

Figura 4.10. Sexo incorrecto.

Por medio de pláticas de retroalimentación, se nos informó que este suceso se debía principalmente a una validación para no ingresar pacientes duplicados en el sistema SIHO, la cual era posible saltearse cambiando el sexo de los pacientes, y de la cual se aprovechaban los usuarios para evitar estar buscando a los pacientes correctos de entre la lista de pacientes ya registrados.

- Campo usados para usos distintos.

De entre las incongruencias detectadas, se encontró que especialmente el campo destinado a la CURP lo utilizaban para notas varias, como cuando el paciente se va sin pagar o es referido de algún lugar en especial, ejemplos que se agruparon en la siguiente imagen como referencia.

CURP	CANTIDAD
.	31
/09101985	1
ACCIDENTADO	15
ASILO DE ANCIANOS	4
ASILO DE ANSIANOS	2
BALEADO	9
BATALLON 66	3
CERESO DE DELICIAS	1
CIRCO	1
DE LA FERIA DE DELIC	1
DESCONOCIDO	5
DETENIDO	11
FOR DE GUERRERO	3
FUERON TODOS LOS DAT	1
GOL POR SOLDADOS	1
GOLPEADO	3
GOLPEADO DE REINGRES	1
JORNALERO	5
NO DIERON MAS DATOS	2
S	78
SE FUGO 371 DEBE	1
VETERANO D LA REV	1

Figura 4.11. Campo CURP usado para notas.

- Registros inutilizables.

Además de las situaciones anteriores, existen registros en los cuales no se puede identificar a quién pertenecen los registros. Un ejemplo de ésto se aplica a bebes recién nacidos que tienen que ser ingresados al sistema antes de que sean registrados o se tenga un nombre para ellos, motivos por los cuales, el personal del sistema opta por crearles un expediente basados en la fecha y hora en la que nacieron, tal cual como se ve en la Figura 4.12 donde se utilizan palabras asemejando la fecha en los campos destinados al nombre propio de los pacientes dentro del sistema SIHO.

VCHAPELLIDOPATERNO	VCHAPELLIDOMATERNO	VCHNOMBRE	CHRSEXO	DTMFECHANACIMIENTO	CHRRFC	CHRCURP
DE JULIO DEL	(Null)	AA	M	1975-10-10 00:00:00	2E1A751010	(Null)
ABRIL DEL	PM	AA	M	2015-04-30 00:00:00	3A4A150430	(Null)
DE JULIO DEL	PM	OCHO	M	1984-10-10 00:00:00	0E1O841010	(Null)
DE JULIO DEL	(Null)	AA	M	1968-10-10 00:00:00	2E1A681010	(Null)
DE ABRIL	PM	AA	M	1955-02-24 00:00:00	2E7A600224	(Null)
DE AGOSTO	PM	AA	M	1984-01-01 00:00:00	1E1A840101	(Null)
DE DICIEMBRE	(Null)	AA	M	1989-01-01 00:00:00	2E1A890101	(Null)
DE ENERO	PM	AA	M	1984-10-10 00:00:00	0E8A841010	(Null)
DE ENERO	(Null)	AA	M	1960-01-01 00:00:00	3E1A600101	(Null)
DE ENERO	(Null)	AA	M	1980-01-01 00:00:00	(Null)	(Null)

Figura 4.12. Datos erróneos en todos los campos.

Como se observa, el análisis a las bases de datos, muestra varias incoherencias en la información contenida. Para la solución que se propone, algunas situaciones se pueden solventar manteniendo la integridad de los datos originales, esto se logró copiando las bases de datos a un servidor donde las mismas serían agrupadas por sistema (Anexo C), lo que nos permite modificar los registros que necesitamos mediante una serie de consultas SQL. En el caso de los nombres con caracteres especiales dentro de los nombres (ver Figura. 4.8), se pueden eliminar dichos caracteres extraños dentro de los diferentes campos ejecutando un simple instrucción, la cual se muestra en la figura siguiente, para el caso del apellido paterno.

```
select distinct cPaterno,dbo.Caracteres(cPaterno)
from CTL_Pacientes
where cPaterno like '%[^a-z]%'
```

Figura 4.13. Query para eliminar caracteres y números del apellido paterno.

El resto de las mismas, como el campo CURP usado para notas y sexo incorrecto en los pacientes se dejan intactos.

4.4 Diseño

El análisis de los datos a utilizar es de gran importancia, dado que ayuda a definir el proceso a realizar. Una vez que se comprendió la información con la que se cuenta y su estado, se empieza a diseñar el conjunto difuso en su totalidad para satisfacer las necesidades requeridas por los Servicios de Salud del Estado de Chihuahua.

4.4.1 Variables Lingüísticas

Las variables difusas nos permiten la transición entre los conjuntos de una manera gradual, con lo que se pueden obtener ciertos grados de pertenencia o parcialidad a uno o varios conjuntos a la vez. Dichas variables deben ser nombradas con base en la información relevante al problema a resolver.

Dado que el proceso de comparación entre los conjuntos difusos se centra en la información personal de los pacientes, y estos pueden regresar a la unidad médica días o años después, entonces para las variables difusas se tienen que considerar aquellos datos que no sufran modificaciones con el paso del tiempo y además sirvan de identificación personal.

Los datos que se encontraron con persistencia en el tiempo y que se encuentran en la mayoría de las tablas para poder realizar el proceso de comparación, son los mostrados en la tabla 4.4.

Tabla 4.4. Variables lingüísticas a comparar.

Campo
Nombre completo
CURP
RFC
Póliza
Fecha de Nacimiento

Como se mencionó en el Marco Teórico, los conjuntos difusos a pesar de evaluar el lenguaje natural, trabajan sobre valores numéricos, por lo cual se tiene que dar numerales a cada una de las variables creadas.

Estos valores se fueron obteniendo mediante una serie de pruebas, las cuales se iniciaron con la idea de que la suma de todas ellas llegara a 100 puntos, dividiendo en partes iguales a cada variable para dar el valor deseado.

El segundo paso para obtener los valores fue, aumentar o disminuir el valor de cada variable según la fortaleza de la misma para identificar a la persona como única, siendo el CURP y RFC los campos sobresalientes en ese aspecto, y donde el análisis efectuado en la primer etapa del desarrollo mostró una falta de cultura por parte de la población para aprenderse o cargar documentos con esos datos, disminuyendo la unicidad que tienen dichos campos para identificar como única a la persona y con ello el valor dado.

El campo de Póliza, aunque exclusivo por paciente, es un campo que comparte la raíz del mismo con la familia del asegurado, provocando similitudes con otros pacientes. También puede cambiar con el tiempo según diferentes factores o incluso, un paciente puede contar con más de un número de póliza, reduciendo la ponderación final que se le da a dicho campo. Al final, la ponderación para cada variable quedó como se muestra en la siguiente tabla.

Tabla 4.5. Variables lingüísticas y su valor máximo.

Campo	Ponderación
Nombre completo	30
CURP	20
RFC	20
Póliza	10
Fecha de Nacimiento	10

4.4.2 Conjuntos difusos y sus funciones de pertenencia

Para cada una de las variables lingüísticas, se diseñó un conjunto difuso con el mismo nombre, el cual se dividió a su vez en subconjuntos representativos a frases comunes que se usan para comparar datos, los cuales fueron: No repetido, Poco Probable Repetido, Probablemente Repetido y Repetido. El total de conjuntos y subconjuntos creados se muestra en la tabla siguiente.

Tabla 4.6. Conjuntos difusos.

Variable Lingüística	**Etiqueta Lingüística**	**Representa**
Nombre	NN	No Repetido.
	NPP	Poco Probable Repetido.
	NP	Probablemente Repetido.
	NR	Repetido.
Fecha de Nacimiento	FN	No Repetido.
	FP	Probablemente Repetido.
	FR	Repetido.
CURP	CN	No Repetido.
	CPP	Poco Probable Repetido.
	CP	Probablemente Repetido.
	CR	Repetido.
RFC	RN	No Repetido.
	RPP	Poco Probable Repetido.
	RP	Probablemente Repetido.
	RR	Repetido.
Póliza	PN	No Repetido.
	PP	Probablemente Repetido.
	PR	Repetido.

Cada uno de los conjuntos difusos contiene los valores que van desde 0 hasta la ponderación máxima que se le otorgó a la variable con el mismo nombre, mientras los subconjuntos o etiquetas lingüísticas abarcan solamente una parte de los valores contenidos en ese rango, y son capaces de entrelazarse para no excluir valor alguno. Gráficamente, los conjuntos difusos se muestran como en la figura siguiente.

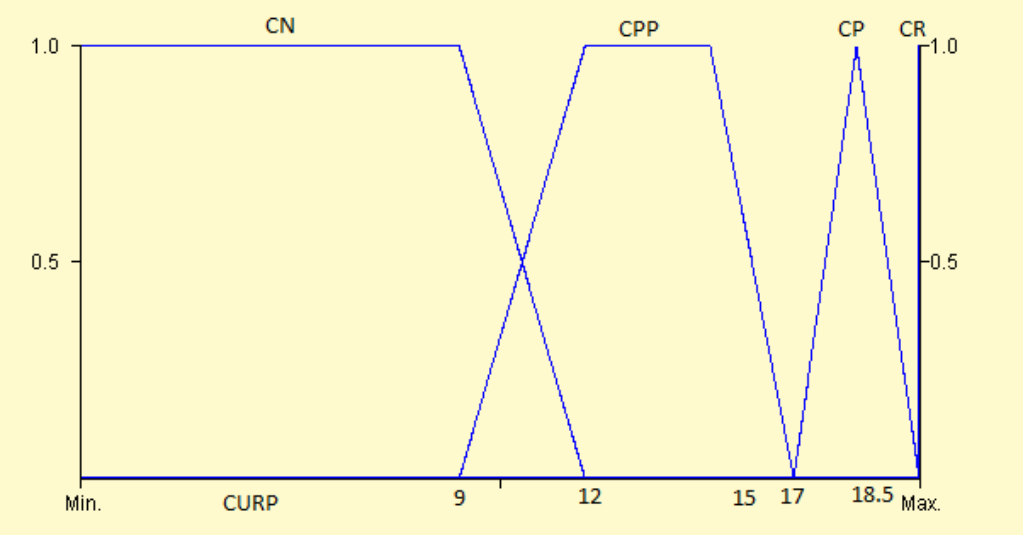


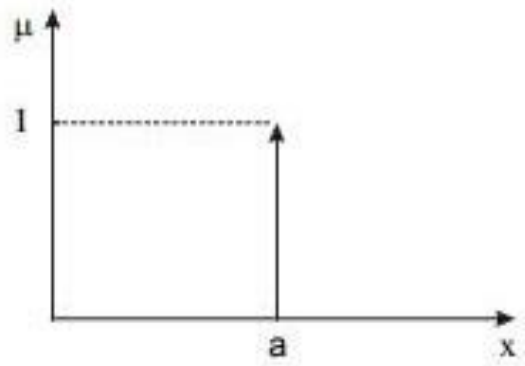
Figura 4.14. Conjunto Difuso CURP.

La imagen anterior ilustra cómo el conjunto difuso es seccionado en su totalidad por los subconjuntos difusos sin dejar valores sueltos. Los rangos de cada uno se obtuvieron primeramente mediante el análisis realizado a la información, del cual se obtuvieron premisas como las siguientes:

- Un CURP bien escrito es único, por lo tanto si 2 CURP son iguales se está comparando a la misma persona.
- Se notó que en ocasiones, cuando el prefijo del CURP es igual a otro y la homoclave es parecida, en realidad se está hablando de la misma persona, caso contrario a cuando el prefijo es igual y la homoclave totalmente distinta, situación que se da cuando se comparan personas diferentes.

Premisas que se aplican de igual manera al RFC y vienen a formar una etiqueta lingüística por sí mismas, donde la primera de ellas, al ser un único valor, nos permite crear

la etiqueta CR (CURP Repetido) bajo la función de pertenencia Singleton, la cual nos dice que si se cumple la condición, el valor de pertenencia será igual a 1.



$$A(x) = \begin{cases} 1 & x = a \\ 0 & x \neq a \end{cases}$$

Figura 4.15. Función de pertenencia Singleton.

La segunda premisa habla sobre el parecido que pueden tener los datos a comparar, hablando del RFC, este punto sería el prefijo o primeros 10 caracteres, donde a partir de la homoclave o últimos 3 caracteres, las comparaciones toman caminos separados, indicando con ello que pasado ese punto, la variable sólo obtendrá valores para un conjunto difuso.

Por lo anterior, se separaron los campos en prefijo y homoclave, para lo cual se calculó el porcentaje de pertenencia de cada uno con respecto al campo completo, obteniendo que el prefijo del RFC es mayor a un 76% para esos 10 caracteres y de un 23% para la homoclave. El porcentaje obtenido del prefijo se traslada como límite inferior del conjunto difuso creado a partir de la premisa 2.

Dichos valores se fueron ajustando por medio de pruebas heurísticas con el fin de disminuir la cantidad de registros que eran clasificados como probables repetidos erróneamente, hasta quedar en un 70% para el RFC y un 85% para el campo CURP. La siguiente figura muestra a la función de pertenencia para el conjunto difuso CURP Probablemente Repetido o CP.

$$f(x) = \begin{cases} 0 & para\ x < 17 \\ \dfrac{x-17}{18.5-17} & para\ 17 < x \leq 18.5 \\ \dfrac{x-20}{18.5-20} & para\ 18.5 < x \leq 20 \\ 0 & para\ x \geq 20 \end{cases}$$

Figura 4.16. Función triangular para CP.

Además de los conjuntos ya descritos, se crearon los siguientes conjuntos para la evaluación de las comparaciones.

Tabla 4.7. Conjuntos difusos de evaluación entre comparaciones.

Variable Lingüística	Etiqueta Lingüística	Representa	Rango
Evaluación 1	Ev1N	No Repetido.	0-10
	Ev1PP	Poco Probable Repetido.	8-30
	Ev1P	Probablemente Repetido.	20-40
	Ev1R	Repetido.	40
Evaluación 2	Ev2N	No Repetido.	0-42
	Ev2PP	Poco Probable Repetido.	39-58
	Ev2P	Probablemente Repetido.	50-66
	Ev2R	Repetido.	60-90
Evaluación 3	Ev3N	No Repetido.	0-7
	Ev3P	Probablemente Repetido.	5-15
	Ev3R	Repetido.	11-20
Evaluación 4	Ev2N	No Repetido.	0-42
	Ev2PP	Poco Probable Repetido.	39-58
	Ev2P	Probablemente Repetido.	50-66
	Ev2R	Repetido.	60-90

Para estos conjuntos el valor del límite superior es la suma de los conjuntos a comparar, en el caso de la Evaluación 1: CURP (20) + RFC (20), Evaluación 3: Fecha de Nacimiento (10) + Póliza (10), mientras en las Evaluaciones 2 y 4 se utilizan todos los conjuntos, dando un total de 90.

4.4.3 Base de conocimiento

En la base de conocimientos se cuenta con la información necesaria para definir los cursos de acción a tomar por medio de reglas lingüísticas, las cuales se obtuvieron por medio del análisis realizado y son el resultado de todas las combinaciones posibles de las variables de entrada aunadas a una posible salida.

Para reducir la cantidad de variables en los antecedentes, se agruparon las variables en pares para su comparación, un ejemplo es la siguiente regla difusa.

SI (*Fecha Repetido* & *Póliza No Repetido*) ENTONCES *Evaluación 3 Probable Repetido*

Para un mayor orden y visualización, las mismas se pueden acomodar dentro de una matriz de reglas difusas.

Tabla 4.8. Matriz difusa para evaluar CURP y RFC.

	RN	RPP	RP	RR
CN	Ev1N	Ev1PP	Ev1P	Ev1P
CPP	Ev1PP	Ev1PP	Ev1P	Ev1P
CP	Ev1P	Ev1P	Ev1P	Ev1P
R	Ev1P	Ev1P	Ev1P	Ev1R

Donde los renglones de la derecha (CN, CPP, CP y CR) equivalen a los conjuntos del campo CURP, las columnas (RN, RPP, RP y RR) al campo RFC y las casillas donde se intersectan se refiere a la Evaluación 1 mostrada en la Figura 4.16. Agregar un tercer antecedente a esta tabla, resultaría en una matriz cúbica, dificultando el proceso y la complejidad de las reglas.

Pensando en mantener al mínimo la cantidad de lados que pudiera tener la matriz de reglas difusas, se diseñó un camino de comparaciones simplificado donde se utilizarán como máximo 2 conjuntos difusos para las mismas.

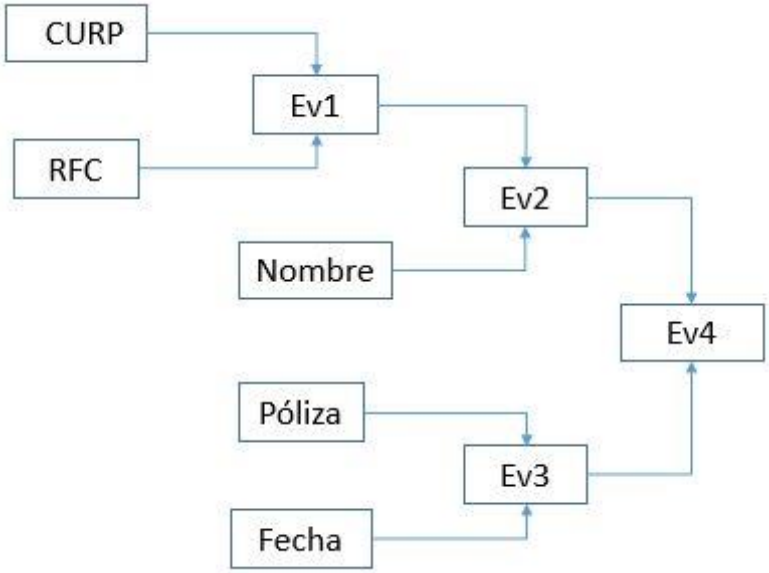


Figura 4.17. Flujo de las comparaciones difusas.

4.5 Codificación

La programación del proyecto se dividió en 2, como ya se ha mencionado con anterioridad, la parte gráfica del sistema y de interacción con los usuarios finales, sin restarle importancia, se maneja por separado (Anexo A). Esto debido al lapso de tiempo que conlleva el actualizar y comparar los datos de los pacientes, es el mismo tiempo en el que las unidades médicas crean o modifican los expedientes clínicos, debido al constante movimiento con el que cuentan dichas instituciones, donde algunas atienden las 24 horas.

Esta separación del sistema, permite que los usuarios corran la interfaz gráfica por su cuenta, mientras se ejecuta el sistema difuso en el servidor ajeno a los movimientos de los usuarios. Como ya se ha comentado, cada uno de los padrones cuenta con diferente tiempo de actualización, siendo el Seguro Popular el que más tiempo lleva, al ser un sistema federal que conlleva un mes en su actualización por las diferentes comprobaciones que realiza, dichas actualizaciones aunadas al constante movimiento en las unidades no permite, al momento, tener un punto final al sistema, donde éste se deje de usar y se empiecen todas las comparaciones desde cero.

4.5.1 Herramientas

La base del sistema descrito es la comparación de los datos personales de los pacientes para encontrar registros pertenecientes al mismo paciente a pesar de las diferencias que puedan existir en los datos. Para su escritura se utilizó el lenguaje de programación Java, un lenguaje muy dinámico multiplataforma que puede ser utilizado en las diferentes máquinas con las que cuentan los Servicios de Salud del Estado de Chihuahua, sin importar el sistema operativo con el que cuenten y se puede conectar a los diferentes gestores de bases de datos. Para su manejo, se optó por el Entorno de Desarrollo Integrado (IDE, por sus siglas en inglés) Netbeans, mientras que para guardar los registros en las bases de datos, se eligió PostgreSQL, un sistema libre para la gestión de bases de datos relacionales basado en el modelo cliente/servidor y que funciona muy bien con grandes cantidades de datos.

4.5.2 Estructura

El sistema difuso se diseñó para trabajar principalmente sobre 2 clases, una clase que lleva el control del flujo de las comparaciones y una para los conjuntos difusos junto al motor difuso. Como se tiene la principal función de comparar personas, además de las clases, se crearon diferentes POJO (acrónimo de Plain Old Java Object) para cada tipo de paciente, diferenciándose cada uno por el padrón al que pertenece cada uno de ellos.

Además, el sistema se delineó para trabajar en segundo plano sin la interacción de un usuario, más que para su inicialización. Una vez iniciado el proceso se corren una serie de procedimientos en automático.

La lista de procedimientos es la siguiente:

- Conectarse a la BD pivote.
- Obtener paciente a comparar.
- Conectarse al padrón siguiente.
- Obtener lista de pacientes similares al que se está comparando.
- Comparar los pacientes por medio de Lógica Difusa.
- Guardar los registros en las bases correspondientes y/o cambios necesarios.

Estos procedimientos se ejecutan cíclicamente hasta terminar los pacientes a comparar en las tablas pivote SP e ICHISAL, este flujo entre las bases de datos se explica más a fondo en el anexo C. Para agilizar las comparaciones futuras, así como para optimizar las comparaciones, el sistema difuso lleva un control de los pacientes que ya se han comparado, así como si son registros duplicados o no, para poder enviar los datos necesarios a las tablas correspondientes.

El sistema ubica los pacientes a comparar de entre todos los registros, destacándolos mediante comparaciones realizadas a códigos fonéticos creados a partir de los campos destinados al nombre propio de los pacientes. Estos códigos se obtuvieron a partir de un algoritmo destinado para ello, el cual es una variante de los algoritmos fonéticos soundex y phonetic spanish desarrollado para este proyecto (Anexo B).

De la lista obtenida, los valores correspondientes a las variables difusas tienen que ser convertidos a valores numéricos para poder ser comparados por medio de la lógica difusa, para ello, se utilizó una variante de la distancia de Levenshtein (Magakian, 2013), que en lugar de traer un valor entero que se incrementa con las diferencias entre los campos, regresa un número entre 0 y 1 que decrece conforme aumentan las discrepancias entre los campos, este algoritmo se explica más a fondo en el anexo D.

Se decantó por esta variante, dado que nos permite ingresar fácilmente valores a los conjuntos difusos mediante la fórmula, A * B = C, donde:

A = Resultado distancia de Levenshtein.
B = Ponderación variable difusa.
C = Valor de entrada al conjunto difuso.

De esta manera aseguramos que el porcentaje de similitud entre los datos, sea el mismo que se ingresa a los conjuntos difusos para su comparación.

Tabla 4.9. Ejemplos de comparaciones con Levenshtein.

Nombre a comparar	Nombre en SP o ICHISAL	Diferencia por Levenshtein	Valor de la variable difusa
ACOSTA FLORES ANABEL	ACOSTA FLORES BRISSEIDA ANABEL	0.666666667	20
FLORES ALVARADO DANIEL ISRAEL	CERVANTES ALVARADO DANIEL ISRAEL	0.75862069	22.75862069
HERANDEZ PEREDA CRISTHOPER	HERANDEZ PEREDA CRISTHOPER LEONARDO	0.75	22.5
MUÑOZ PARRA SOCORRO AYDALIE	MUÑOZ PARRA AYDALIE	0.708333333	21.25

Los valores determinados por la distancia de Levenshtein, son evaluados por las funciones de membresía con el fin de obtener los grados de pertenencia a los conjuntos en cuestión. Los resultados conseguidos para cada conjunto, son evaluados por las reglas difusas para definir el curso de acción. Para las comparaciones difusas, la parte a resaltar es el valor de entrada que se utiliza para las comparaciones Evaluación 2 y Evaluación 4 (mostradas en la Figura 4.19 como Ev2 y Ev4), donde uno (Ev2) o ambos (Ev4) valores son los resultados desfusificados de las evaluaciones anteriores (Ev1 y Ev3 respectivamente), los cuales fueron obtenidos por el método del centroide y vueltos a fusificar por medio de los conjuntos difusos creados para ello.

Las comparaciones definidas como duplicadas son aquellas que obtienen un valor desfusificado mayor a 60 para la Evaluación 4, a las cuales una bandera booleana las clasifica como verdaderas y son agregadas a un arreglo para guardarse posteriormente en la tabla designada para ello. Este arreglo se envía a la base de datos y reinicia cada cierto número de comparaciones con la intención de conservar la memoria de la máquina.

4.5.3 Algoritmos

Una manera de entender más a fondo la estructura del sistema, es presentando el código utilizado para su creación. Como se mencionó anteriormente, son pocas las clases usadas para la parte difusa del sistema, dejando a la mayoría de las mismas de soporte para la realización de sus funciones, entre las que se encuentran:

- Clases para la conexiones a los gestores de bases de datos.
- Clases con los querys para cada uno de los diferentes manejadores.
- Clases de los POJO.
- Clase de comparación de distancia de Levenshtein.
- Clase de conjuntos difusos.

Cabe resaltar que aunque la etapa de codificación está dedicada a presentar el código del sistema, éste no se presenta en su totalidad, sino solamente partes que se consideran de importancia para el entendimiento y/o funcionamiento de la aplicación.

4.5.3.1 Clase Comparaciones

Método público

- void puntuaciones()

Método principal del sistema y donde se llevan a cabo el proceso de las comparaciones.

Métodos privados

- void comparaICHISAL()

Método con las operaciones necesarias para cargar y manejar los datos de los pacientes provenientes de ICHISAL que se van a comparar, de igual manera existen métodos similares para SIGHO y SIHO.

- Connection con()

Método dedicado a abrir y cerrar conexiones manteniendo una u dos conexiones a la vez para mantener la estabilidad del sistema.

- void pupa()

Método que se encarga de enviar a una tabla dedicada a las personas que no se encontraron repetidas dentro del proceso de comparaciones.

- void estatal()

Método que envía a una tabla los registros marcados como duplicados para su comparación por un usuario final.

Procedimiento de comparaciones

El método en el que se lleva a cabo el proceso de las comparaciones, se encuentra dentro de la clase a revisar y se denomina puntuaciones, este método es el que se encarga de dictar los pasos a seguir de una manera secuencial, sin embargo, su principal función es la toma de decisiones en cada uno de los pasos del proceso. Dentro de estas decisiones, la primera que se tiene, es el conectarse a la tabla pivote a comparar, una vez que se realizan las conexiones necesarias, se procede a realizar la paginación de los registros y anexarlos a una lista de pacientes, esto se aprecia brevemente en la figura siguiente.

```
while (pagActual <= totalPags) { //= totalPags) {

    String qInicial = queries(pivote);

    //calcular pagina y offset
    if (pagActual == 1) {
        offset = 0;
    } else {
        offset = (pagActual * limit) - 1;
    }

    stPivote = connPivote.prepareStatement(qInicial);
    stPivote.setInt(1, limit);
    stPivote.setInt(2, offset);

    rsPivote = stPivote.executeQuery();

    //Guardar los datos del rsPivote en una lista de objetos tip
    List<PacienteSP> lstPacPivote = new ArrayList<PacienteSP>();
```

Figura 4.18. Paginación de los registros.

Los registros obtenidos se guardan en una lista de tipo paciente del Seguro Popular, sin importar la tabla pivote de la que provengan, lo que nos permite cerrar la conexión a la tabla pivote y tener un mayor control del manejo de la memoria.

El arreglo obtenido es más eficiente para ser comparado con los registros de las demás tablas con las que se cuenta. Una manera de hacer esto es mediante un ciclo que recorra el mismo junto a las tablas en cuestión, como se muestra en la siguiente figura, donde para evitar comparar los pacientes de ICHISAL con ellos mismos, se diseñó la validación que se muestra en la misma imagen.

```
for (PacienteSP pacientePiv : lstPacPivote) {
    startTime = System.currentTimeMillis();
    Pacientes.pacPiv(pacientePiv, pacSP);
    punt.setTotal(punt.getTotal() + 1);

    if (pivote != 3) {
        comparaIchisal(connIchisal);
    }
```

Figura 4.19. Comparación del arreglo de pacientes.

Los métodos de comparación con las otras tablas son de manera obligada sin importar la tabla pivote que se esté comparando, motivo por el cual, no es necesario algún tipo de validación sobre su uso.

Una vez que se terminó de cotejar el arreglo, se verifica si una segunda lista cuenta con registros de los pacientes, de ser así, se ejecuta el método para agregar pacientes a la tabla denominada estatal, la cual contiene los registros a comparar por un usuario final. Esta segunda lista se reinicia con cada iteración del ciclo y contiene sólo los registros que se catalogaron como duplicados.

```
if (lstRelacion.size() > 0) {
    // Insertar en tabla relacion con datos BDLL y SP
    addALH(connLH, lstRelacion);
}
```

Figura 4.20. Envió de pacientes a tabla estatal.

4.5.3.2 Clase MetodosLD

Métodos públicos

- double[] curp (double puntuacion)

Método que regresa los valores de los sistemas difusos del campo CURP dentro de un arreglo de valores tipo double, según la ponderación dada por la distancia de Levenshtein.

- double[] rfc (double puntuacion)

Método que retorna un arreglo tipo double con los resultados a los diferentes conjuntos difusos asociados a la variable difusa RFC.

- double[] cr (PacienteLH paciente, PacienteSP pacSP)

Este proceso recibe por entrada 2 POJO de pacientes, los cuales son la base para iniciar las siguientes tareas de manera secuencial.

  - Obtener las distancias de Levenshtein para los 2 pacientes referentes a los campos CURP y RFC.
  - Invocar los métodos CURP y RFC descritos anteriormente.
  - Evaluar las reglas difusas referentes a los campos mencionados.
  - Diferenciar los destinos para las diferentes reglas y enviar los resultados al proceso de desfusificación.

```
/**
 * Evalua CURP
 *
 * @param puntuacion, puntuacion levenshtein de la curp
 * @return array con los valores de los conjuntos CN,CPP, CP y CR
 */
public double[] curp(double puntuacion) {...72 lines }//curp
```

Figura 4.21. Declaración del método para fusificar la CURP.

Procedimiento de comparaciones difusas

El método CR arriba mencionado, es un ejemplo de los diferentes procedimientos que se diseñaron para la evaluación de las reglas difusas. Dentro del mismo se obtienen los valores necesarios para su evaluación, seguimiento y resultado final. A continuación se describe el flujo del mismo:

1.- El proceso se inicia con la declaración de los arreglos donde se guardarán los valores que se obtengan para cada uno de los conjuntos difusos CURP y RFC, así como las variables que contendrán los resultados de la distancia de Levenshtein de dichos campos.

```
double curp[] = new double[0];
double rfc[] = new double[0];

double puntuacionCURP = 0.0;
double puntuacionRFC = 0.0;
```

Figura 4.22. Declaración de arreglos y variables para los resultados.

2.- Se procede a calcular la distancia de Levenshtein para ambos campos, según la figura siguiente.

```
puntuacionCURP = Levenhstein.getDistance(paciente.getPaCURP().trim(),
        pacSP.getSpCURP().trim()) * punt.getCurp();
```

Figura 4.23. Obtención del valor de Levenhstein para el campo CURP.

3.- Se obtienen los valores para los conjuntos difusos.

```
curp = curp(puntuacionCURP);
```

Figura 4.24. Asignar valores al arreglo CURP.

4.- Se obtienen los resultados de las reglas de control, el cual es el valor mínimo de los 2 valores a comparar.

```
double cNrN = min(curp[0], rfc[0]);
double cNrPP = min(curp[0], rfc[1]);
```

Figura 4.25. Resultados de las reglas de control.

5.- Se agrupan las reglas según el camino a tomar y se obtiene el valor máximo entre ellas.

```
//Reglas que envian a Probable repetido
if (cNrP != 0 || cPPrP != 0 || cPrN != 0 || cPrPP != 0 || cPrP != 0) {

    Set<Double> max = new HashSet<Double>();

    max.add(cNrP);
    max.add(cPPrP);
    max.add(cPrN);
    max.add(cPrPP);
    max.add(cPrP);

    paciente.setResEvlP(Collections.max(max));

} else {
    paciente.setResEvlP(0);
}
```

Figura 4.26. Agrupación de reglas difusas según su salida.

6.- Para terminar el proceso se regresan los valores obtenidos, donde serán asignados al arreglo que maneja los valores de salida de cada evaluación, con el fin de llevar un mayor control de los resultados y continuar con el proceso correspondiente.

```
return new double[]{paciente.getResEvlN(), paciente.getResEvlPC(),
    paciente.getResEvlP(), paciente.getResEvlR(), puntuacionCURP,
    puntuacionRFC};
```

Figura 4.27. Salida del método de Evaluación 1.

4.6 Pruebas

Como medida de verificación del funcionamiento del controlador difuso se realizaron pruebas de corroboración al mismo. Con la finalidad de determinar su comportamiento en un ambiente real y el firme propósito de obtener conclusiones precisas sobre el sistema difuso, las comprobaciones se enfocan principalmente en el rendimiento y en los resultados obtenidos. Estas pruebas se realizaron en el siguiente equipo de cómputo.

Tabla 4.10. Especificaciones del equipo de cómputo.

Disco duro	500 GB.
Procesador	Intel® Core™ i5-4210M (2.6 GHz).
Tarjeta de video	Integrada.
Sistema operativo	Windows 10 Pro 64.
Memoria	8.0GB PC4-17000 DDR4 2133 MHz.
Software	Postgres 9.3.

4.6.1 Evaluación de pruebas

La primera verificación realizada consistió en comprobar la veracidad y eficacia de del algoritmo realizado. Debido a los constantes cambios que surgen las bases de datos en el sector salud, se eligió una cantidad pequeña de registros a comparar. Siendo elegidos sólo los primeros 1,000 registros del Seguro Popular, al igual que los primeros mil para ICHISAL y SIGHO, donde los resultados fueron por demás satisfactorios con un 99.9% de veracidad, mismos que se detallan a fondo en el capítulo denominado Pruebas y Resultados.

4.6.2 Comprobación de resultados por medio de probabilidad

Por medio del teorema de Bayes se pueden calcular probabilidades de un evento que ha sucedido con anterioridad, dígase un paciente que ya se ha atendido previamente. El teorema de causas, como también se le conoce a la teoría de Bayes, se ayuda del conocimiento previo para poder estimar posibles valores basados en la experiencia obtenida de este conocimiento a priori, ayudando a corroborar los resultados previamente obtenidos mediante las pruebas de evaluación.

Para lo cual, los conocimientos a priori (A_i) forman el universo de sucesos conocidos y mutuamente excluyentes entre si $\{A_1, A_2, A_3,..., A_n\}$, y donde la teoría nos dice que de un suceso B se conocen las probabilidades, proporcionando así cierta información relevante sobre las posibles ocurrencias a obtener, dado que las probabilidades de que ocurra B, varían según el suceso A_i que haya ocurrido.

La fórmula de Bayes se expresa de la siguiente manera:

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{P(B)} \qquad \text{(IV.1)}$$

Donde:

P (A_i) son las probabilidades a priori que se conocen.
P (B | A_i) es la probabilidad de que ocurra B dado el suceso A_i.
P (A_i | B) son las probabilidades a posteriori basadas en el suceso B.

La cual se modifica de la siguiente manera para variables que pueden tomar más de 2 valores:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum P(B)\,P(B|A_i)} \qquad \text{(IV.2)}$$

4.7 Implementación

Dado que el sistema difuso para la detección de registros duplicados de pacientes en múltiples fuentes de información forma parte del proyecto total que se está implementando por parte de los Servicios de Salud del Estado de Chihuahua, el sistema se encuentra en modo de espera para ser implementado a nivel estatal.

Se encuentra en espera la terminación total del proyecto por parte de los Servicios de Salud del Estado de Chihuahua, así como de la consecución de todas las bases pertenecientes a los diferentes sistemas, se procede a realizar las pruebas mencionadas en la sección 4.6.1, contando solamente con las bases de datos del Seguro Popular (verificada al mes de mayo 2016), la tabla de ICHISAL al mismo mes, además de los registros pertenecientes al Centro de Salud División del Norte también al mes de mayo, dicha unidad labora con el sistema SIGHO.

V) PRUEBAS Y RESULTADOS

5.1 Resultados de la pruebas de evaluación

La corroboración realizada como primer prueba arrojó resultados muy satisfactorios a lo que se espera del sistema difuso, para el cual se considera como resultado óptimo aquel donde entre los resultados obtenidos los registros catalogados falsamente como duplicados sean mínimos y los catalogados equivocadamente como no duplicados sean nulos. Esto se busca con el fin de minimizar el cansancio al uso del sistema por parte de los usuarios finales (y posibles fallas de uso por el mismo), así como evitar posibles faltas de enlaces entre los registros correspondientes.

Los resultados obtenidos se muestran en la siguiente tabla, señalando una cantidad muy alta de comparaciones (274,396) si se considera que solo se compararon mil pacientes del Seguro Popular contra 2,000 registros de otros padrones. Esta cantidad pudiera ser considerada como una cantidad mínima si se le compara con la cantidad de equiparaciones que se realizarían por el método de fuerza bruta, donde este último método arrojaría un resultado estimado de 1,215,000,000,000 comparaciones, dada la operación de $|SP| \times |ICHISAL| \times |SIGHO|$ como se especifica en la formula mostrada en el Anexo B.

Tabla 5.1. Resultados obtenidos.

Registros Analizados.	3,000
Comparaciones.	274,396
Registros Duplicados.	1,163
Comparaciones Duplicados.	1,228
Falsos Negativos.	1
Falsos Positivos.	0
Verdaderos Positivos.	1,228
Verdaderos Negativos.	273,167

De la misma tabla se obtiene que un 38% de los registros se catalogaron como duplicados, de igual manera se consideró solo al 0.44% de las comparaciones como iguales, las cuales superan en número a la cantidad de registros duplicados, informando de esta manera que uno o más pacientes de los considerados repetidos, fueron vinculados con más de un registro.

Así mismo, los campos de falsos positivos y falsos negativos indican que se logró en gran parte el objetivo planteado en un inicio de disminuir los casos en cuestión, siendo el único registro no marcado como duplicado un caso a analizar. La siguiente figura muestra un ejemplo de un falso negativo.

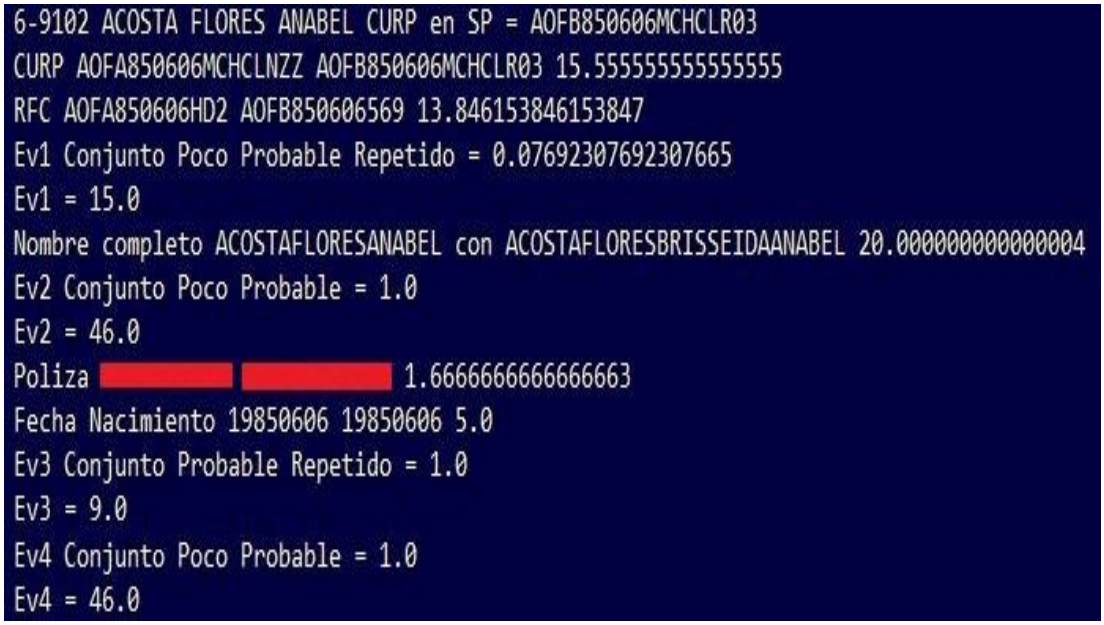


Figura 5.1. Ejemplo de un falso negativo.

En la Figura 5.1 se muestra que una persona de la tabla del SIGHO con nombre Acosta Flores Anabel se comparó con un registro del Seguro Popular que cuenta con el CURP AOFB850606MCHCLR03, a la cual desde la Evaluación 1 la empieza a catalogar como “Poco Probable Repetida”.

La razón de ello es porque los usuarios del sistema SIGHO acostumbran terminar el campo del CURP con la homoclave ZZ para los CURP que desconocen y ZY para cuando se repiten, mientras para el RFC buscan algún generador del mismo, el cual, en esta situación

se generó de forma incorrecta, debido a que como se observa en la Evaluación 2, el dato del nombre propio del paciente se encuentra incompleto para el sistema SIGHO ocasionando que los 3 datos sean equívocos, reduciendo el valor obtenido por la distancia de Levenshtein.

Para la Evaluación 3, donde la póliza por sistema es difícil que se pueda repetir, el sistema ya cataloga a la persona como Probablemente Repetida dada la total similitud en la fecha de nacimiento. Al estimar las Evaluaciones 2 y 3 para obtener el resultado final, el controlado difuso termina por catalogarlo como “Poco Probable Repetido” con un valor final de 46.0, motivo por el cuál no se considera como duplicado.

En el apartado de registros marcados correctamente, las comparaciones realizadas muestran varios casos que demuestran el buen funcionamiento del motor difuso. Uno de ellos es el ejemplo de la imagen siguiente, donde el sistema evalúa a 2 personas que a simple vista dan la apariencia de ser gemelos, situación dada la gran similitud contenida por todos los campos, donde los campos con menor porcentaje de similitud son RFC y Póliza cuentan con una semejanza mayor al 66%, aumentando a 77% y 81% el CURP y nombre respectivamente, mientras la equiparación para la fecha de nacimiento muestra un 100% de afinidad, porcentajes superiores a la media y que en un sistema de evaluación matemática pudieran considerar al paciente como duplicado. La siguiente figura muestra un ejemplo de diferencia entre posibles gemelos.

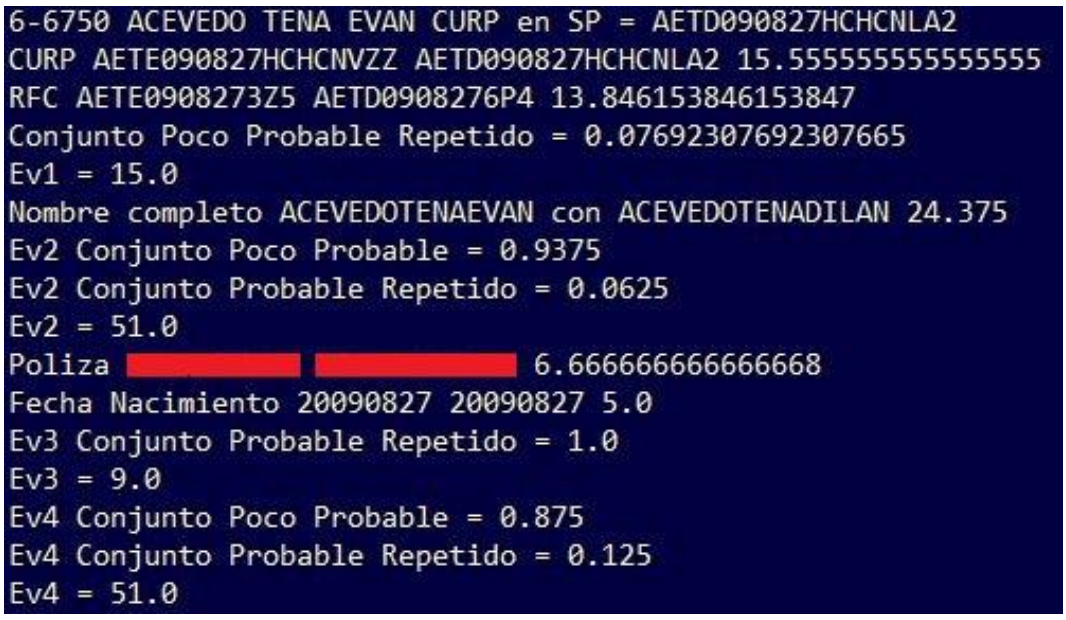


Figura 5.2 Ejemplo de diferencia entre posibles gemelos.

En la figura anterior se nota a mayor detalle la parte difusa del sistema, en la cual una variable forma parte de más de un conjunto difuso a la vez, en este caso la variable perteneciente al nombre es parte de los conjuntos difusos “Poco Probable” y “Probablemente Repetido” con valores diferentes para cada uno de ellos. Caso similar a cuando se calcula el resultado final, donde después de cotejar los resultados de las evaluaciones 2 y 3, se vuelve a presentar la pertenencia a los mismos dos conjuntos pero con un valor diferente, que al ser desfusificado da un resultado final de 51, considerando correctamente al registro como no duplicado.

Para el ejemplo siguiente (Figura 5.3.) se tiene una situación donde los valores son más semejantes entre sí con excepción de la póliza y fecha de nacimiento, mismos que cuentan con una similitud menor al 50%, donde el controlador difuso evalúa dicha situación como “No Repetido” y “Probable Repetido”.

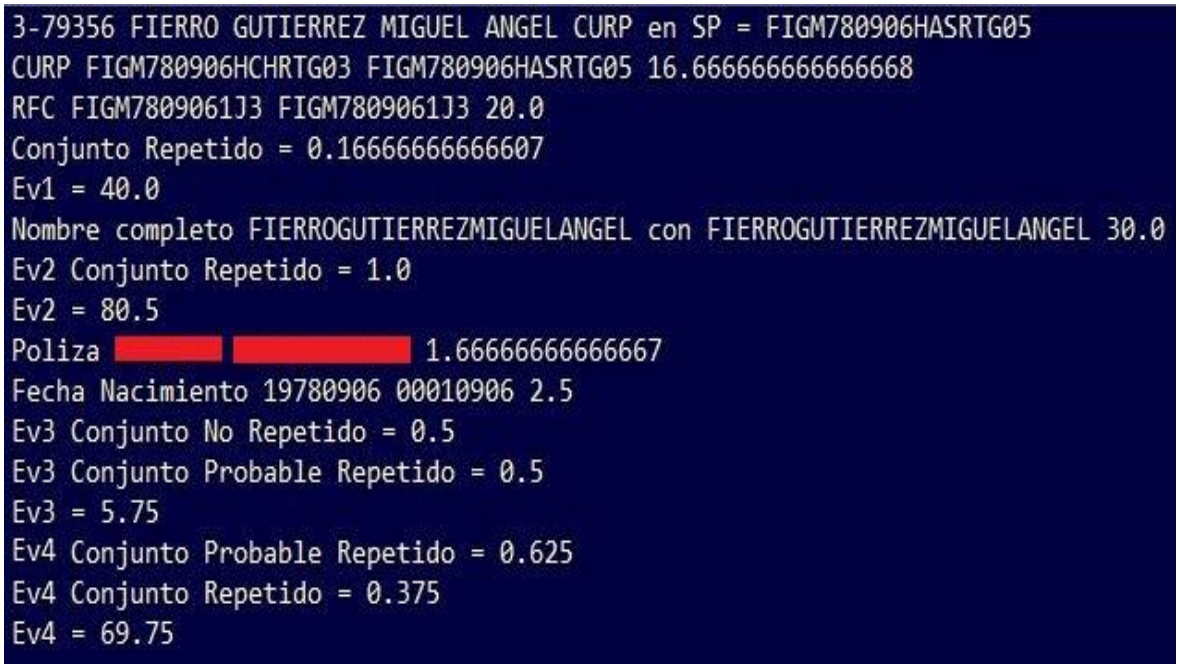


Figura 5.3. Ejemplo de caso duplicado.

Mientras el resto de las variables cuenta con un porcentaje mayor al 83% para el CURP, llegando a coincidir totalmente los datos pertenecientes al RFC y nombre de la persona, situaciones donde el sistema cataloga al registro como “Repetido”, el cual al ser

comparado con los resultados de la Evaluación 3, da un valor de pertenencia mayor al conjunto de “Probablemente Repetido”, identificando de manera correcta al registro como duplicado.

Un segundo ejemplo de un paciente marcado como duplicado es el que se muestra en la Figura 5.4, donde las situaciones de comparación resultan un poco diferentes al caso anterior.

Figura 5.4. Ejemplo de caso duplicado sin CURP.

Para esta ocurrencia, el valor del CURP se desconoce para el padrón de ICHISAL, mientras que el RFC cuenta con incongruencias que reducen su similitud a un 76%, pese a lo cual el controlador considera al registro como probable duplicado. Al realizar la Evaluación 2, el sistema detecta un 94% de igualad para el nombre, registrando dicho valor como perteneciente a dos conjuntos difusos diferentes, los cuales son “Probable Repetido” y “Repetido”. La Evaluación 3 posiciona al registro dentro del mismo conjunto de “Probable Repetido”, esto por la alta homogeneidad mostrada en la fecha de nacimiento, a pesar de la baja paridad con la que cuenta la póliza de la persona.

A diferencia del caso mostrado en la Figura 5.3, donde el paciente es catalogado como dentro del conjunto “Repetido” en dos de las tres evaluaciones, el valor final de salida señala al registro como perteneciente a dos conjuntos diferentes de salida, mostrando un valor final de 69.75 para la última evaluación. Mientras para el evento mostrado en la Figura 5.4 el valor desfusificado muestra un valor final de 80.5 a pesar de haber sido considerado como “Probablemente Repetido” en las tres evaluaciones realizadas y solo en una ocasión como “Repetido”, siendo este último el resultado final de la comparación.

Esta discrepancia en los valores de salida se debe al proceso de desfusificación, el cual obtiene el valor promedio de los conjuntos de salida, dicho lo anterior, la Figura 5.5 presenta a los conjuntos difusos pertenecientes a la evaluación de salida final. En esta imagen los valores del conjunto “Repetido”, usado en el caso anterior, son mayores a los conjuntos “Probable Repetido” y “Repetido” mostrados en la Figura 5.3.

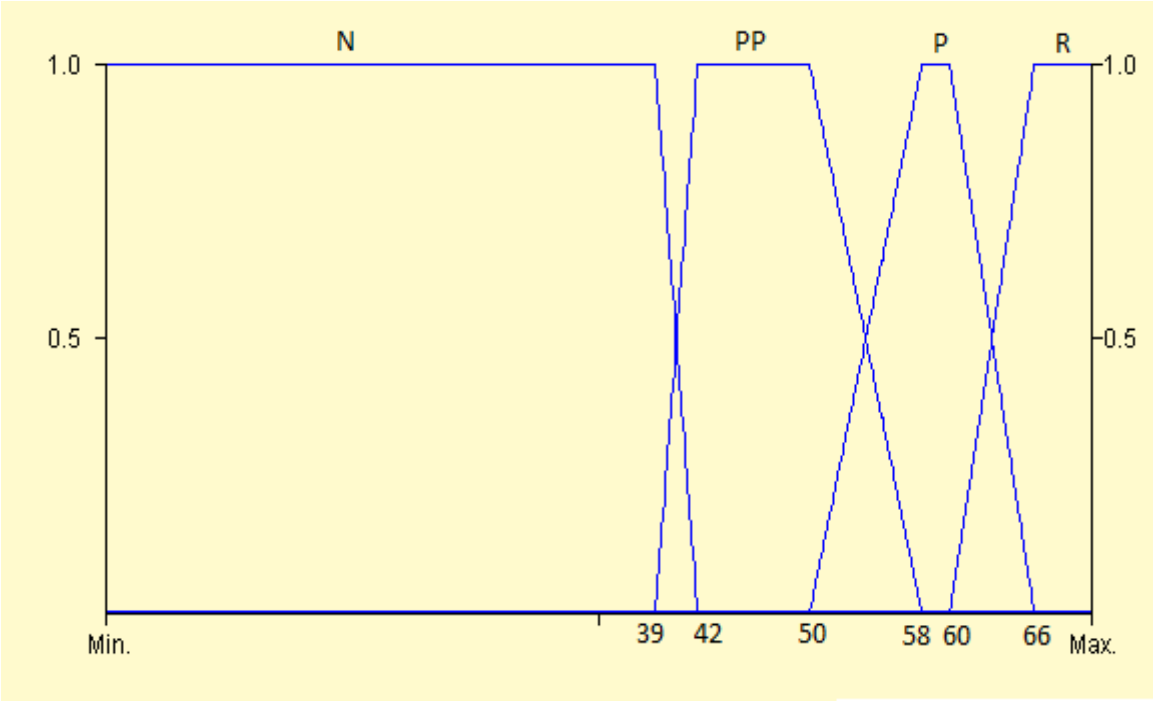


Figura 5.5. Conjuntos de la Evaluación 4.

5.2 Probabilidades de duplicidad

La gran cantidad de registros contenidos en cada una de las tablas, dificulta el poder llevar un control exacto de cuántos registros cuentan con duplicados dentro de una misma tabla, más aún con las demás tablas existentes dentro de los sistemas usados por los Servicios de Salud.

Una herramienta de gran utilidad es el Teorema de Bayes la cual nos ayuda a calcular la probabilidad de que exista una duplicidad pero sin llegar a determinar realmente si el registro pertenece o no a la sección de duplicados.

Para llegar a estos valores se tiene que obtener primeramente el total de los registros contenidos en cada uno de los diferentes padrones, por lo que se tiene que:

Tabla 5.2. Total de registros por padrón.

Padrón	Registros
Seguro Popular	1,411,764
ICHISAL	123,738
SIGHO	17,976

De los cuales se sabe que los registros pertenecientes al Seguro Popular son únicos (Anexo C), dando un porcentaje de duplicidad de 0% para cualquiera de ellos. Estos datos se conocen debido a que el Seguro Popular reduce la cantidad total de registros de la población estatal en un aproximado 33%, con lo que logra eliminar los registros que se encuentran duplicados.

Pensando que la cantidad de registros duplicados es semejante para el resto de los padrones se seleccionaron al azar 30% del total de sus registros, esto para comprobar manualmente cuál sería el porcentaje de registros duplicados para cada uno de ellos. Los datos que se obtuvieron se muestran en la siguiente tabla:

Tabla 5.3. Total registros duplicados de ICHISAL y SIGHO.

Padrón	Total de Pacientes	30%	Duplicados del 30%	No duplicados del 30%
ICHISAL	123,738	37,121	616	36,505
SIGHO	17,976	5,392	543	4,849

Con base en esos resultados se obtuvieron los siguientes porcentajes:

Tabla 5.4. Porcentajes de duplicados y no duplicados por padrón.

Clave	Descripción	Porcentaje
A	De que el paciente pertenezca al SP.	90.88%
B	De que el paciente pertenezca al SP y sea duplicado.	0%
C	De que el paciente pertenezca al SP y no sea duplicado.	100%
D	De que el paciente pertenezca al ICHISAL	7.96%
E	De que el paciente pertenezca al ICHISAL y sea duplicado.	1.66%
F	De que el paciente pertenezca al ICHISAL y no sea duplicado.	98.34%
G	De que el paciente pertenezca al SIGHO.	1.16%
H	De que el paciente pertenezca al SIGHO y sea duplicado.	10.07%
I	De que el paciente pertenezca al SIGHO y no sea duplicado.	89.93%

Con estos porcentajes se pueden calcular las probabilidades de que un registro de un paciente esté duplicado en cualquiera de los padrones conocidos, como se muestra en la siguiente tabla, en la que se usan las claves de la tabla anterior para simplificar las descripciones de las probabilidades.

Tabla 5.5. Probabilidades de que el paciente pertenezca a un padrón y sea repetido.

Probabilidad de	Sustituyendo valores	Resultado
$P_A \cap P_B$	(0.9088) (0)	0
$P_D \cap P_E$	(0.0796) (0.0166)	0.0013
$P_G \cap P_H$	(0.0116) (0.1007)	0.0012

Las probabilidades de que el paciente sea duplicado sin importar el padrón al que pertenezca es la sumatoria de las probabilidades anteriores, por lo que se tiene lo siguiente:

$$P(sea\ \ duplicado) = (P_A \cap P_B) + (P_D \cap P_E) + (P_G \cap P_H) \qquad \text{(V.I)}$$

$$P(sea\ \ duplicado) = (0) + (0.0013) + (0.0012)$$

$$P(sea\ \ duplicado) = 0.0025$$

Con esta probabilidad de que un registro sea duplicado podemos obtener las probabilidades condicionales mencionadas en el Teorema de Bayes, por ejemplo, la probabilidad de que un registro pertenezca al ICHISAL y sea duplicado

$$P(sea\ \ duplicado\ y\ del\ ICHISAL) = \frac{(P_D \cap P_E)}{P(sea\ duplicado)} \qquad \text{(V.2)}$$

$$P(sea\ \ duplicado\ y\ del\ ICHISAL) = \frac{(0.0796) * (0.0166)}{0.0025}$$

$$P(sea\ \ duplicado\ y\ del\ ICHISAL) = 0.5285$$

5.3 Pruebas controladas

Para esta prueba final se seleccionó con anticipación una cantidad de pacientes, de los cuales una cantidad mínima está duplicada en una o más bases de datos, esto con el fin de tener un conocimiento mayor sobre los puntos clave con los cuales trabaja el controlador difuso diseñado.

El INEGI menciona que para mayo de 2015 el nombre más en utilizado en México es Hernández Hernández José, motivo por el cual se utilizó dicho nombre como el paciente seleccionado para su búsqueda en la tablas seleccionadas. Se conoce que dicho dato se encuentra registrado en la tabla del Seguro Popular 36 ocasiones y 4 veces en la tabla del ICHISAL. De estos registros se tiene que el dato Hernández Hernández José Antonio está registrado 3 veces para el Seguro Popular y 2 en el ICHISAL, coincidiendo uno de ellos como duplicado.

Con base en el parámetro seleccionado de nombre propio se espera obtener el resultado máximo posible donde los falsos positivos y negativos dan un valor igual a 0, los valores obtenidos por esta prueba, fueron los siguientes:

Tabla 5.6. Resultados prueba controlada.

Registros Analizados.	36
Comparaciones.	412
Registros duplicados.	1
Comparaciones duplicados.	1
Falsos Negativos.	0
Falsos Positivos	0
Verdaderos Positivos	1
Verdaderos Negativos	411

La siguiente figura muestra la única comparación obtenida como duplicada:

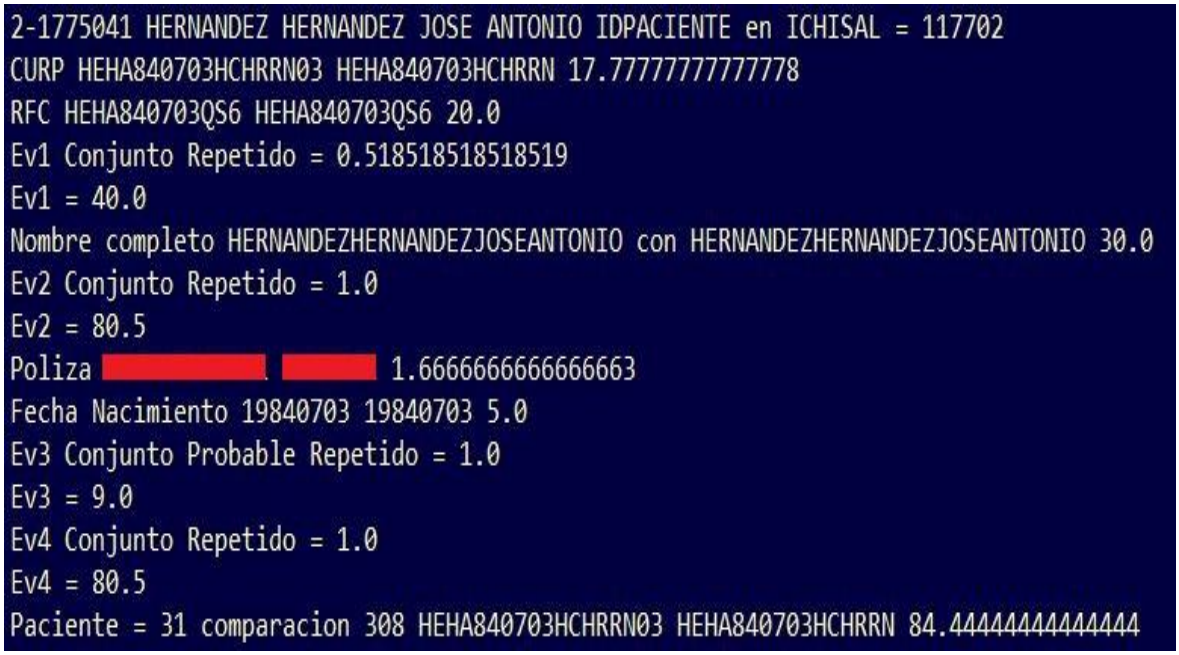


Figura 5.6. Hernández Hernández José duplicado.

De la cual podemos observar que el único dato con información diferente es el CURP, mismo que se encuentra incompleto, obteniendo un porcentaje cercano al 90%, motivo por el cual el sistema desde un comienzo lo registra como duplicado con un valor final de salida igual a 80.5.

Para los casos donde se compara el mismo nombre pero de diferentes personas, podemos observar en las Figuras 5.7 y 5.8 como los campos CURP y RFC tienen un gran peso en la salida final de las comparaciones, donde los valores de salida van de acorde a los resultados obtenidos en la Evaluación 1.

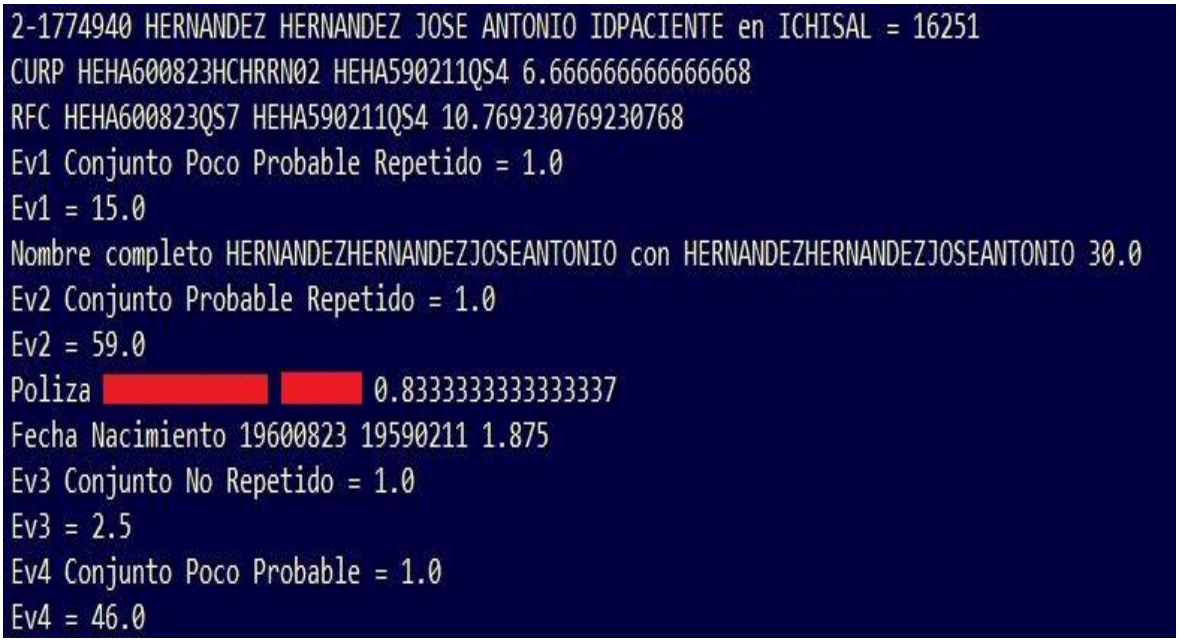


Figura 5.7. Mismo nombre no duplicado.

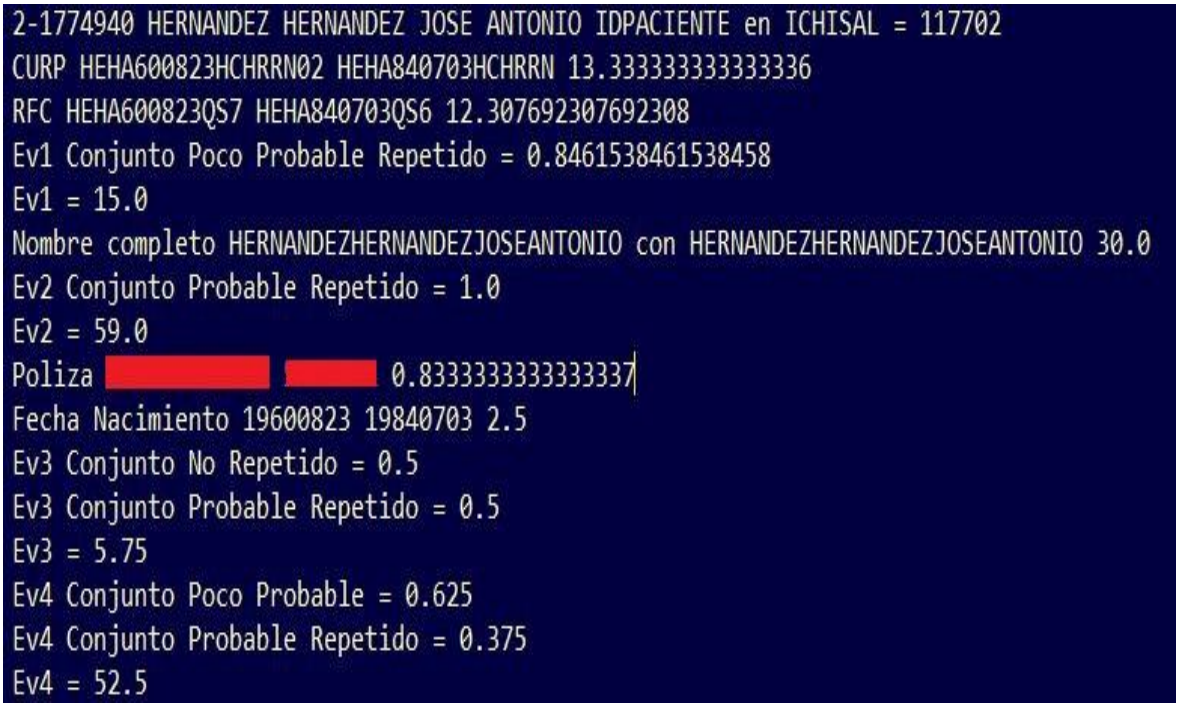


Figura 5.8. Segundo ejemplo de mismo nombre no duplicado.

VI) CONCLUSIONES

Para este proyecto se utilizó la Lógica Difusa para encontrar los registros duplicados pertenecientes a los pacientes de los Servicios de Salud del Estado de Chihuahua, que se encuentran distribuidos en las múltiples bases de datos correspondientes a los diferentes padrones y sistemas con los que cuentan los Servicios de Salud.

La solución desarrollada trabaja sobre datos persistentes al paso del tiempo, siendo capaz de encontrar satisfactoriamente registros duplicados aun cuando se cuenten con datos incompletos o faltantes como se mostró en el Capítulo V, con lo que se puede concluir que el sistema cumple con el objetivo general del proyecto.

Los resultados obtenidos superan a los esperados, donde entre las metas se tenía el obtener el mínimo de falsos positivos en las comparaciones, así como el obtener cero falsos negativos en las mismas; metas que se lograron parcialmente al conseguir 0 falsos positivos y 1 falso negativo.

Siendo el único caso no detectado correctamente (Figura 5.1) un caso excepcional debido al gran porcentaje de incongruencias con las que cuentan los datos del paciente como es el nombre incompleto, además de un CURP y RFC generados con base a dicho dato erróneo, donde en otras situaciones semejantes de nombre incompleto pero con CURP o RFC bien ingresado, el sistema es capaz de identificar a la persona como duplicada correctamente.

Así mismo, durante el desarrollo del proyecto se hacen observaciones en cuanto a posibles mejoras a futuro, como lo es una segunda y/o tercera prueba de datos para verificar el buen funcionamiento del mismo y con ello ajustar los valores de los conjuntos difusos en caso de ser necesario.

ANEXO A Funcionamiento de la parte gráfica del sistema.

Uno de los principales requerimientos solicitados durante las entrevistas con el personal de Servicios de Salud del Estado de Chihuahua, fue la creación de una pantalla donde los registros que se catalogaran como duplicados, pudieran ser evaluados por el personal de los Servicios de Salud, de manera que funcione de comprobación a los resultados que se marcaron como exitosos en la etapa anterior.

Decisión que se tomó primordialmente por el hecho de trabajar sobre los expedientes clínicos de los pacientes, en donde una relación incompleta de los mismos o con adiciones a ellos puede dirigir al especialista hacia un diagnóstico o tratamiento inadecuado, lo que pudiera generar mayores problemas a la salud de la persona y con ello, posibles problemas al doctor y/o institución por una posible mal práctica médica.

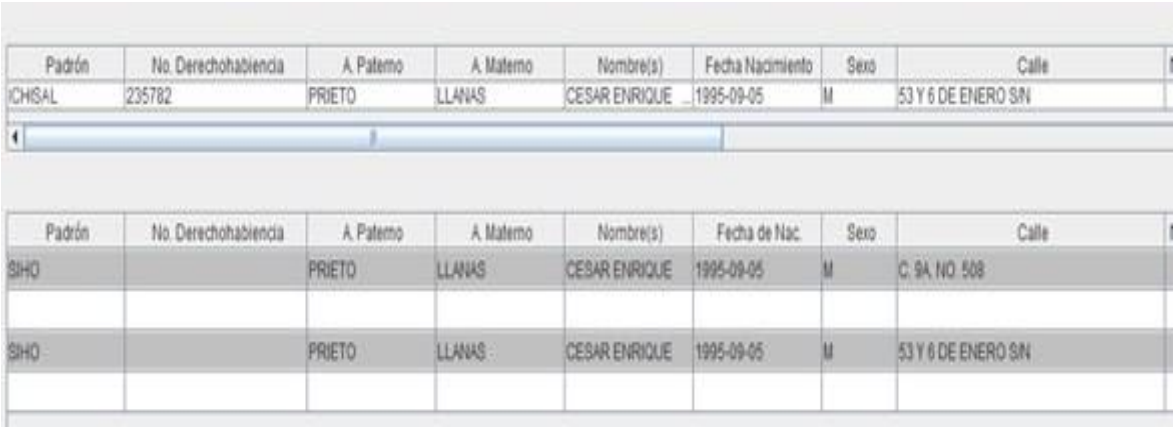
Para esta parte del sistema se pidió se tomaran en cuenta el poco conocimiento de los usuarios con respecto a los sistemas computacionales, que además de ello, cuentan con una gran carga de trabajo, motivos por los cuales el sistema debe ser de fácil uso e intuitivo, caso contrario, el personal no lo usará de la manera adecuada o dejaría de usarlo, siendo estos los factores principales para el éxito del sistema desde el punto de usabilidad.

La primera pantalla muestra el acceso al sistema, siendo este por usuario y contraseña registrados para su uso.

ANEXO A Funcionamiento de la parte gráfica del sistema.

Figura A.1. Pantalla de acceso al sistema.

Una vez que se ingresa al sistema, nos lleva directamente a la zona de comparación de pacientes, donde en la parte superior se muestran los datos del paciente que fue marcado como duplicado en una de las tablas pivote (Seguro Popular e ICHISAL), mientras que en la parte inferior se muestran los registros con los que se marcó la duplicidad.



Padrón	No. Derechohabiencia	A. Paterno	A. Materno	Nombre(s)	Fecha Nacimiento	Sexo	Calle
ICHISAL	235782	PRIETO	LLANAS	CESAR ENRIQUE	1995-09-05	M	53 Y 6 DE ENERO S/N

Padrón	No. Derechohabiencia	A. Paterno	A. Materno	Nombre(s)	Fecha de Nac.	Sexo	Calle
SIHO		PRIETO	LLANAS	CESAR ENRIQUE	1995-09-05	M	C. 9A. NO. 508
SIHO		PRIETO	LLANAS	CESAR ENRIQUE	1995-09-05	M	53 Y 6 DE ENERO S/N

Figura A.2. Resultados de pacientes duplicados.

Al final de cada registro duplicado se encuentran las opciones que podemos realizar con dichos pacientes.

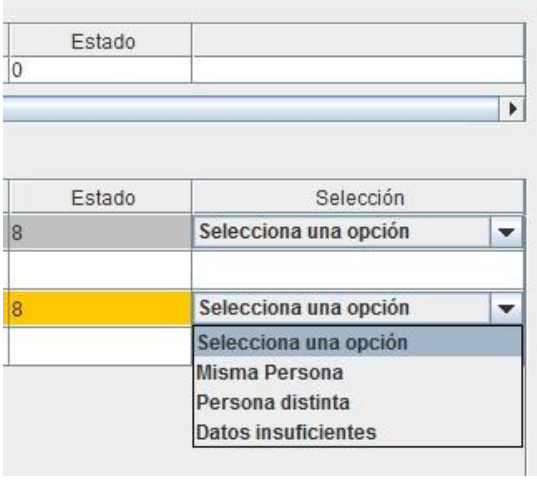


Figura A.3. Opciones para catalogar un registro catalogado como duplicado.

Estas opciones son:

- Misma Persona.- Cuando estamos seguros que el paciente está duplicado.
- Persona Distinta.- Cuando los datos demuestran que se trata de pacientes diferentes.
- Datos Insuficientes.- Cuando los datos mostrados no muestran la información necesaria para decidir entre las dos primeras opciones.

Cuando más de un registro se considera duplicado, el sistema nos preguntará cuál de los pacientes será catalogado como el registro ancla o principal con el cual se relacionarán los demás registros.

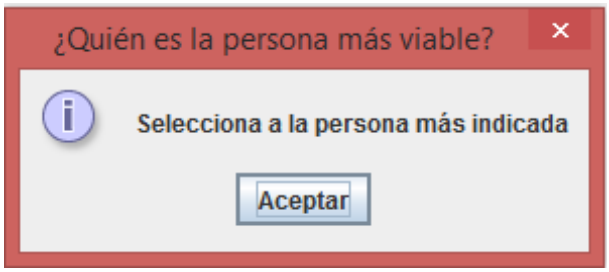


Figura A.4. Recordatorio para seleccionar un registro como principal.

En este caso, el botón de selección se cambia por un radio button, para asegurar que sólo seleccionemos uno y solo un registro.

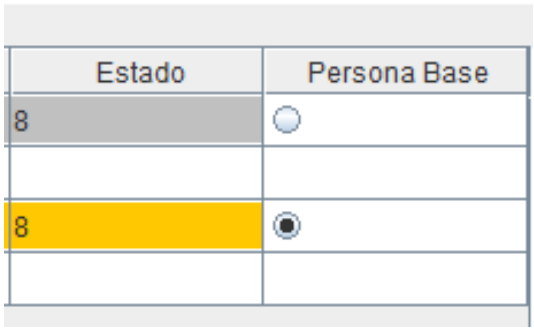


Figura A.5. Selección de registro principal.

ANEXO B Método de codificación fonética.

La cantidad de registros totales en todas las tablas supera los 2 millones y continúa aumentando día con día. Realizar una serie de equiparaciones del tipo todos contra todos, para encontrar los registros similares, es una tarea titánica que no sólo verificaría registros sin semejanza alguna, también conllevaría una gran pérdida de tiempo debido al número de comparaciones a realizar, el cual estaría dado por |A| x |B| x |C| donde:

A = Número de registros en la base Pivote (Seguro Popular o ICHISAL).

B = Número de registros en la tabla concentrada.

C = Número de registros en la tabla de ICHISAL para cuando la base concentrada es diferente de ICHISAL.

Teniendo de base los cálculos realizados por el Seguro Popular, donde el porcentaje de registros duplicados es de alrededor del 30%, la mayoría de los registros a comparar deben ser únicos, ocasionando como ya se mencionó, una gran pérdida de tiempo debido al gran porcentaje de comparaciones que habría entre registros sin semejanza alguna.

Para reducir el número de comparaciones se realiza un proceso de indexación, con la finalidad de agrupar los registros que realmente se parezcan entre sí. Una de las maneras de hacer esto, es mediante una comparación fonética entre los campos, la cual se basa en las similitudes con las que se pronuncian las palabras, en este caso, los nombres propios.

Mediante la codificación fonética podemos generar un código asociado al sonido que produce una palabra al ser pronunciada, como lo es el caso de la codificación por Soundex. Debido a que el algoritmo de codificación Soundex está diseñado para el idioma inglés, que causa ruido en los datos a utilizar, motivo por el cual se buscó una opción diseñada al español, encontrando en el Phonetic Spanish una opción muy viable para nuestro sistema.

El Phonetic Spanish se basa en estudios realizados donde se demuestra que el 90% de los errores en la escritura de lengua española son errores simples o más bien dicho, un solo error por palabra, el cual puede ser causado por sustituciones fonéticas, como b por v a

la hora de escribir las palabras. Estos errores, junto a los errores múltiples y demás problemas ortográficos, pueden llegar al 63% del total de los errores en la ortografía española.

Estos métodos de codificación consisten en intercambiar una letra por un valor númerico, mientras Soundex utiliza la letra inicial y 3 dígitos para su representación, el Phonetic Spanish descarta dicho método y convierte todas las letras. Los valores usados por cada algoritmo se muestran en las Tablas B.1 y B.2:

Tabla B.1. Codificación Soundex.

Digito	Caracteres
1	B,F,P,V
2	C,G,J,K,Q,S,X,Z
3	D, T
4	L
5	M,N
6	R

Tabla B.2. Codificación Phonetic Spanish.

Digito	Caracteres
0	P
1	B, V
2	F, H
3	D, T
4	C, S, X, Z
5	Y, L, LL
6	M, N, Ñ
7	Q, K
8	G, J
9	R, RR

Una de las diferencias principales se debe a que en el lenguaje inglés, la letras A, E, I, O, U, H, W e Y no cuentan con peso fonético para diferenciarse y son descartadas totalmente de la codificación.

Con base en esos valores el Phonetic Spanish genera los siguientes códigos (Tabla B.3) para las diferentes variaciones de un mismo apellido.

Tabla B.3. Códigos del Phonetic Spanish.

Apellido	Código
Armendáriz	966394
AArmendariz	96694
Amendariz	66364
Aremndariz	966394
Armandariz	966394
Armendrariz	966394
Armendriz	966394
Armendaiz	966394

Pese al buen rango de variaciones que se agrupan en el mismo código, se diseñó una variante del Phonetic Spanish para su uso durante el proyecto, esto debido a las posibles diferencias en nombres entre Colombia (país originario del Phonetic Spanish) y México, donde la población del estado de Chihuahua, al colindar con Estados Unidos acostumbra poner nombres americanos a sus hijos, lo que pudiera ser causa de algún problema en dicho algoritmo.

El algoritmo diseñado mezcla las reglas del Soundex con los valores de codificación del Phonetic Spanish, de tal manera que las reglas generales de nuestro algoritmo quedaron de la siguiente manera:

- De igual forma que el Soundex se utiliza la letra inicial de la palabra.
- Se codifican el resto de los caracteres con base en los valores del Phonetic Spanish mostrados en la Tabla B.2.
- Si existen dígitos iguales adjuntos, se deja uno de ellos, eliminando el resto.
- Se borran caracteres innecesarios, para ello los marcamos como X.
- Si el código resultante es menor a 3 dígitos, se agregan 0’s a la derecha para completar el código, en caso contrario, se eliminan los dígitos a la derecha del tercero.

Implementando este cifrado a los campos del nombre y apellidos podemos agrupar aquellos casos semejantes en un solo valor para su comparación, dejando de lados todos los registros innecesarios para nuestro proceso de comparación.

Como ejemplo de ello, en la Tabla B.4 se pueden ver las diferentes formas en que ha sido ingresado el apellido Armendáriz y los códigos generados para dichos casos.

Tabla B.4. Códigos generados para las variantes del apellido Armendáriz.

Apellido	Código
Armendáriz	A966
AArmendariz	A966
Amendariz	A663
Aremndariz	A963
Armandariz	A966
Armendrariz	A966
Armendriz	A966
Armendaiz	A966

Como se puede apreciar, no todas las variaciones van a quedar con el mismo código, incluso podemos llegar a comparar apellidos que realmente sean diferentes al evaluado, pero sí aseguramos que una gran parte de las variaciones sean comparadas.

Si comparamos los resultados comparados con los del Phonetic Spanish, se tienen las siguientes diferencias entre los códigos generados (ver Tabla B.5).

Tabla B.5. Comparación entre códigos.

Apellido	Phonetic Spanish	Código generado
Armendáriz	966394	A966
AArmendariz	96694	A966
Amendariz	66364	A663
Aremndariz	966394	A963
Armandariz	966394	A966
Armendrariz	966394	A966
Armendriz	966394	A966
Armendaiz	966394	A966

Donde ambas codificaciones obtienen el mismo porcentaje de diferencias agrupadas con el mismo código, variando en el 50% de los casos que no obtienen el mismo código.

ANEXO C Flujo de conexiones a las diferentes Bases de datos.

El sistema difuso comparará los registros contenidos en todas las tablas existentes. Para obtener un control de estas comparaciones se optó por tener una tabla pivote, con la cual se van a estar comparando las demás tablas. La tabla elegida fue la del Seguro Popular, tabla que originalmente cuenta con 1,800,000 registros, los cuales se reducen un 33% después de que el Seguro Popular aplica una serie de procesos para evitar duplicados. Procesos como: revisar CURP mal escritas, errores en los nombres y registros duplicados, son realizados mes con mes para mantener limpia su tabla, la cual se reduce a un aproximado de 1, 215,000 registros.

Para un mayor control, la tabla de ICHISAL se eligió como una segunda tabla pivote de comparación, quedando el esquema de comparación entre tablas de la siguiente manera.

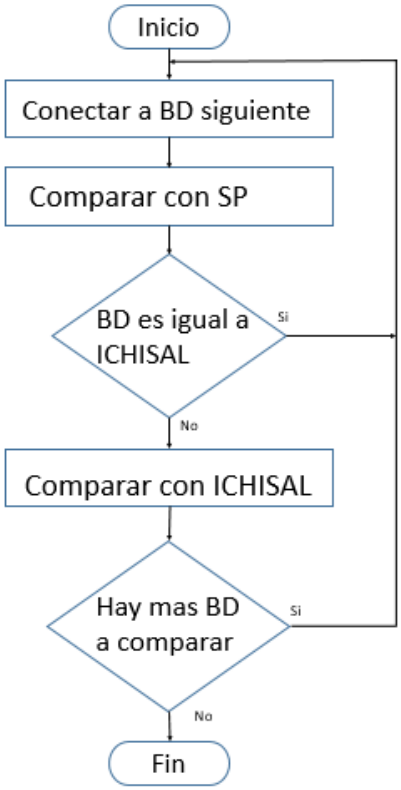


Figura C.1. Flujo de conexiones a las diferentes tablas.

Con esto se redujo la cantidad de tablas a comparar, con el fin de disminuir la cantidad de conexiones y aumentar la confianza en las mismas, además de mejorar la velocidad de comparación e integridad de los datos. Esto se logró mediante la creación de un catálogo de las tablas, agrupadas por sistemas de salud. El catálogo quedó como se muestra en la tabla siguiente.

Tabla C.1 Catálogo de las Bases de Datos.

Identificador	Nombre	Descripción
1	SP	Seguro Popular.
2	ICHISAL	Instituto Chihuahuense de la salud.
3	SIHO	Sistema Hospitalario.
4	SIGHO	Sistema Integral de Gestión Hospitalaria.

ANEXO D Distancia de Levenshtein.

El algoritmo de distancia de Levenshtein, es un algoritmo para medir la similitud entre dos cadenas, el cual regresa un valor entero que se incrementa por cada cambio que sufre una cadena para convertirse en la otra. Entre estas operaciones se consideran:

- Borrado de un carácter.
- Inserción de un carácter.
- Substitución de un carácter por otro.

Como mencionamos este algoritmo regresa el valor de cuántos pasos tiene que hacer una cadena para ser igual a la otra, siendo 0 cuando son totalmente iguales, pero para saber qué tan semejantes son esas cadenas es necesario modificar el algoritmo para que regrese un valor real entre 0 y 1, donde 1 representa la igualdad total. En la tabla siguiente se muestran los valores de distancia y similitud por Levenshtein.

Tabla D.1. Valores de Levenshtein para el apellido Armendáriz.

Apellido Original	Apellido ingresado	# de cambios	Similitud
Armendáriz	Armendáriz	0	1
Armendáriz	AArmendariz	1	0.909090909
Armendáriz	Amendariz	1	0.9
Armendáriz	Aremndariz	2	0.8
Armendáriz	Armandariz	1	0.9
Armendáriz	Armendrariz	1	0.909090909
Armendáriz	Armendriz	1	0.9
Armendáriz	Armendaiz	1	0.9

REFERENCIAS

AHIMA (2010). *"Fundamentals for Building a Master Patient Index/Enterprise Master Patient Index (Updated)."* Journal of AHIMA (Actualizado en Septiembre 2010).

Amón, I., Jiménez, C., (2010), "*61SPA. Funciones de Similitud sobre Cadenas de Texto: Una Comparación Basada en la Naturaleza de los Datos*". CONF-IRM 2010 Proceedings. Paper 58. http://aisel.aisnet.org/confirm2010/58

Barros, L., Bassanezi, R. and Lodwick, W. (2017). *A first course in fuzzy logic, fuzzy dynamical systems, and biomathematics*. Berlín, Alemania.: Springer.

Benito, T., Durán, M. (2009) *"Lógica Borrosa. Universidad".* Universidad Carlos III. Recuperado de http://www.it.uc3m.es/jvillena/irc/practicas/08-09/10.pdf

Do Carmo, A., & Motz, R. (2001, Abril). *Propuesta para integrar bases de datos que contienen información de la Web. Workshop Iberoamericano de Ingeniería de Requisitos y Ambientes de Software (IDEAS'2001), Costa Rica*.

Gutiérrez, J. (2008). *Sistemas Expertos Basados en Reglas*.

Kisko, S. (2005). *Fuzzy Logic and its Practical use in Mass Transit Systems*. Skisko.blogspot.mx. Recuperado de: http://skisko.blogspot.mx/2005/06/fuzzy-logic-and-its-practical-use-in.html el 14 de abril. 2016.

Instituto Nacional Electoral. (2015). *Estadísticas Lista Nominal y Padrón Electoral*, [en línea] Recuperado de http://www.ine.mx/archivos3/portal/historico/contenido/Estadisticas_Lista_Nominal_y_Padron_Electoral/

Magakian, M. (2013). *Detecting string similarity in Java and C# - doduck*. [en línea] Recuperado de: http://doduck.com/string-similarity-java-csharp/ el 20 de mayo 2016].

Martínez, C., Cimarelli, R., Fazio, T. & Fuentes, G. (Septiembre 2015). *Desarrollo de un Índice Maestro de Pacientes Utilizando Estándares y Software Open Source*. 6to Congreso Argentino de Informática y Salud. Congreso llevado a cabo dentro de las 44° JAIIO en Rosario, Arg.

Mendel, J., Hagras, H., Tan, W., Melek, W. and Ying, H. (2014). *Introduction to type-2 fuzzy logic control*. 1ra ed. Hoboken, NJ: Wiley, pp.1-2.

Manual del Expediente Clínico Electrónico (2011). Dirección General de Información en Salud. Secretaría de Salud. México.

Morales, G. (2002). "*Introducción a la lógica difusa"*. Centro de Investigación y Estudios Avanzados. México. Recuperado de http://delta.cs.cinvestav.mx/~gmorales/ldifll/ldifll.pdf

Oracle. (2010) *Sun Master Patient Index,* Recuperado de https://docs.oracle.com/cd/E19509-01/820-3377/dsgn_mpi-about_c/index.html

Ordaz, S. (2011). *Detección de Leucemia Linfoblástica Aguda usando lógica difusa y redes neuronales*. Tesis para Maestría en Ciencias en Ingeniería Electrónica. Instituto Politécnico Nacional.

Ortiz, C. (2015). *Sistema evaluador de calidad de datos en MySQL*. Tesis de Ingeniero en Computación. Universidad Nacional Autónoma de México.

Pollack, A. (1989). *Fuzzy Computer Theory: How to Mimic the Mind?. [en línea] Nytimes.com.* Recuperado de: http://www.nytimes.com/1989/04/02/us/fuzzy-computer-theory-how-to-mimic-the-mind.html el 11 febrero 2016.

REFERENCIAS

Pressman, R. S. (2010). *Ingeniería del software. Un enfoque práctico*. México: McGraw-Hill.

Sánchez, R., (2004) *"Lógica Difusa".* Universidad Carlos III. Recuperado de https://www.it.uc3m.es/jvillena/irc/descarga.htm?url=practicas/estudios/Logica_Difusa.pdf

Soberats, C. (2011). *Sistema de detección de personas duplicadas en un Sistema de Gestión de Información Médica*. Tesis de Ingeniero Informático. Universidad de las Ciencias Informáticas.

Starczewski, J. (2013). *Advanced Concepts in Fuzzy Logic and Systems with Membership Uncertainty*. Primera Edición. Berlín, Alemania: Springer.

Sujansky, W. Jones, L. (2004). *Patient Data Matching Software: A Buyer's Guide for the Budget Conscious*.

Yasunobu, S., Miyamoto, S. and Ihara, H. (2002). *A Fuzzy Control for Train Automatic Stop Control. Transactions of the Society of Instrument and Control Engineers*, E-2(1), pp.1-9.

Zadeh, L. (1965).*"Fuzzy sets"*. Information and Control (8): 338–353.