Universidad Internacional de La Rioja (UNIR)

Escuela Superior de Ingeniería y Tecnología

Máster Universitario en Visual Analytics & Big Data

Modelo de clasificación de las condiciones clínicas que componen la prediabetes

Trabajo Fin de Máster

presentado por: Rivadeneira Fuel, Gabriel Alejandro

Director/a: Gámez, Nadia

Ciudad: Quito, Ecuador

Fecha : Septiembre de 2018

Índice de contenidos

Índice de figuras

Índice de tablas

Agradecimiento

A todas las personas que, mediante sus acciones y gestos, positivos y negativos, permitieron que llegue al lugar donde estoy, brindándole diversión al camino trazado.

A todos aquellos quienes supieron con su ejemplo mostrarme el camino correcto y a quienes supieron con sus experiencias alejarme de lo indebido.

A quienes creyeron en mi forma de ver el mundo y mi forma de interpretarlo.

A mi madre por brindarme el criterio para ignorar todo lo antes dicho y vivir la vida a mi manera.

A mi abuelito por creer ciegamente en mí y ganarse el derecho de ser llamado papá.

A quienes me acompañaron a disfrutar de las experiencias diversas que se vivieron en el camino recorrido, después de ignorarlo todo.

A quienes me ayudaron mientras recorría el camino trazado a recuperar lo que en algún punto se me había perdido.

A quienes hoy me acompañan a disfrutar de lo bueno que se obtuvo al recorrer esta pequeña parte del camino y me ayudan a superar lo malo que se obtuvo de él.

Por último, a quienes me enseñaron a superar los obstáculos que se presenten en el interminable camino trazado diciendo “...No se preocupe, la vida es hermosa...”.

Resumen

Con el paso del tiempo se evidencia el incremento en la cantidad de personas que padecen diabetes mellitus tipo dos a lo largo del mundo. Según la Organización Mundial de la Salud, desde el año 1980 hasta la actualidad la cantidad de personas con esta patología ha crecido en más de 4 veces su valor original, por lo que afirma que, en una sola generación, la diabetes pasó de ser una rareza a una epidemia crónica y progresiva. Hasta el momento, no se ha abordado el tratamiento de la prediabetes como un medio de prevención de la aparición de diabetes, a pesar que las investigaciones realizadas en los últimos años aseguran que una intervención oportuna puede disminuir en aproximadamente un 58% la progresión de prediabetes a diabetes mellitus tipo2. Este trabajo tiene por objeto aplicar un modelo de clasificación que permita identificar de manera temprana potenciales complicaciones que pueden desembocar en el aparecimiento de la diabetes tipo dos, generando así una alternativa para la disminución del gasto público en la remediación de sus síntomas y su posible aplicación como política pública de prevención. Este modelo se desarrolla basado en los resultados obtenidos desde el año 2011 hasta el año 2016 de la encuesta estadounidense NHANES, los cuales miden diversos criterios de salud y nutrición de los encuestados, utilizando un algoritmo de inteligencia artificial que induce la creación de un árbol de clasificación. Esta aproximación metodológica nos permite hallar un árbol de clasificación optimista que presenta un nivel de acierto del 77.8% y posee un diseño amigable con el usuario.

Palabras Clave: Diabetes, Prediabetes, Clasificación, Inteligencia artificial, Big Data, Ciencia de Datos

Abstract

Over time, the increase in the number of people suffering from type two diabetes mellitus throughout the world is becoming increasingly evident. According to the World Health Organization, from 1980 to the present day, the number of people with this pathology has grown more than 4 times its original value, so it states that, in a single generation, diabetes went from being a rarity to a chronic and progressive epidemic. So far, the treatment of prediabetes has not been addressed as a means of preventing the onset of diabetes, although research conducted in recent years ensures that an early intervention can reduce the progression of prediabetes to type two diabetes by approximately 58%. The purpose of this document is to apply a classification model that allows early detection of potential complications that may lead to the onset of type two diabetes, thus generating an alternative for reducing public expenditure on the remediation of its symptoms and its possible application as a public prevention policy. This model is developed based on the C4.5 algorithm of artificial intelligence, based on the results obtained from years 2011 to 2016 of the North American survey NHANES, which measures diverse health and nutrition criteria of the respondents, using an artificial intelligence algorithm that induces the creation of a classification tree. This methodological approach allows us to find a classification tree with optimistic qualities that has a prediction level of 77.8% and has a user-friendly design.

Keywords: Diabetes, Prediabetes, Classification, Artificial Intelligence, NHANES, Big Data, Data Science

1. Introducción

En el año 2016, la Organización Mundial de la Salud (OMS) emitió su primer informe global sobre la diabetes con resultados contrarios a los esperados. En la generación correspondiente al año 1980, la cantidad de personas a nivel mundial que padecían de diabetes se había cuadruplicado.

En una sola generación, la diabetes pasó de ser una rareza a una epidemia, considerada una enfermedad crónica y progresiva que promete una vida de declive lento, doloroso y que avanza hacia una muerte prematura (OMS, 2016).

La diabetes mellitus tipo 2 es una enfermedad crónica que surge cuando el páncreas no logra producir suficiente insulina o cuando el organismo no utiliza eficazmente la insulina que produce. Su efecto más preocupante es que con el tiempo daña gravemente muchos órganos y sistemas, especialmente los nervios y los vasos sanguíneos.

1.1 Motivación y Planteamiento

La Diabetes Mellitus tipo 2 (DM2) es un desorden metabólico degenerativo, cuya prevalencia se ha incrementado constantemente a nivel mundial. Como resultado de su crecimiento la DM2 se ha convertido en una epidemia, de la que se estima tener 626.6 millones de afectados para el año 2045 debido a la falta de cuidado con respecto a sus factores de riesgo y el envejecimiento general de la población (FID, 2017).

Para el año 2017, existían aproximadamente 425 millones de personas con DM2, que se encontraban entre los 20 y 79 años, equivalente al 8.8% de la población mundial, generando un gasto sanitario anual de 727 000 millones de dólares americanos (USD), con proyección de incrementarse a 776 000 millones USD para el año 2045. (FID, 2017). Este enorme volumen e incremento acelerado de personas, diagnosticadas o no, con DM2 a lo largo del tiempo, generan enormes inconvenientes para los prestadores médicos, convirtiéndose en un problema de salud pública con un impacto económico extraordinario, particularmente para los países en vías de desarrollo.

Si consideramos que los valores anteriormente mencionados ignoran completamente la existencia de 352 millones de personas con riesgo de desarrollar DM2 y los agregamos al valor previamente citado, tenemos un 16.1% de personas a nivel mundial que presentan algún tipo de alteración en sus índices de medición de glucosa u otro factor de riesgo que desemboca en DM2, tal como se muestra en la Tabla 1.

Tabla 1. Diabetes de un vistazo (FID, 2017)

De un vistazo	2017	2045
Estimaciones mundiales sobre diabetes		
Prevalencia (20 a 79 años)	8.80%	9.90%
Número de personas con diabetes (20 a 79 años)	425 millones	628.6 millones
Gastos sanitarios totales por diabetes (20 a 79 años)	USD 727 000 millones	USD 776 000 millones
Estimaciones sobre alteración de la tolerancia a la glucosa (ATG)		
Prevalencia mundial (20 a 79 años)	7.30%	8.30%
Número de personas con ATG (20 a 79 años)	352 millones	532 millones

Analizando las complicaciones por las que atraviesan las personas que padecen de DM2, podemos mencionar problemas cardiovasculares, daños en la visión, enfermedades renales, lesiones nerviosas, entre otras, como se resume en la Tabla 2. (FID, 2017)

Tabla 2. Complicaciones generadas por la diabetes (Elaboración propia)

Complicaciones diabéticas	
Enfermedades cardiovasculares (ECV)	
Enfermedad cardiaca coronaria	Problemas con vasos sanguíneos cardíacos
Enfermedad cerebrovascular	Problemas con vasos sanguíneos del cerebro
Enfermedad arterial periférica	Problemas con vasos sanguíneos que riegan brazos y piernas
Enfermedad cardiaca reumática	Complicaciones con músculos o válvulas cardíacas por fiebre reumática
Enfermedad cardiaca congénita	Malformaciones cardíacas presentes en el nacimiento
Enfermedad del ojo diabético (EOD)	
EOD	Lesiones en los capilares de la retina
Enfermedad renal diabética (ERD)	
Enfermedad renal crónica	Pérdida parcial de la funcionalidad de los riñones
Enfermedad renal terminal	Pérdida total de la funcionalidad de los riñones con desenlaces mortales
Lesiones nerviosas	
Pie diabético	Lesiones de los tejidos profundos de las extremidades inferiores
Amputaciones	Separación de una extremidad del cuerpo mediante cirugía
Salud bucodental	
Enfermedad periodontal	Inflamación de las encías y pérdida de piezas dentales
Complicaciones en el embarazo	
Pérdida fetal	Aborto espontáneo
Malformaciones congénitas	Alteraciones anatómicas en el feto
Muerte fetal	Fallecimiento intrauterina de un feto
Muerte perinatal	Fallecimiento de un recién nacido
Pre-eclampsia	Alta presión arterial en el embarazo
Eclampsia	Convulsiones o coma en una mujer embarazada
Muerte materna	Fallecimiento de la madre

Al observar el tremendo impacto socioeconómico y de salud sobre el individuo con DM2, al igual que sobre los sistemas de salud pública de sus respectivos países, es inevitable concluir que es vital el diagnosticar de manera temprana el padecimiento o posible padecimiento de la patología, con la intención de prevenir sus terribles consecuencias.

La prediabetes (PD) es una condición clínica asintomática donde los niveles de glucosa en sangre superan los valores normales, pero se encuentran por debajo de los necesarios para sustentar un diagnóstico de diabetes. (Rosas-Saucedo, 2017)

Contrario a los que podría pensarse, la PD presenta la necesidad de cuidados de salud importantes debido a que una intervención temprana y oportuna puede disminuir aproximadamente en un 58% la progresión de esta condición hacia DM2. (Rosas-Saucedo, 2017)

Este desorden presenta consecuencias tan significativas en la población mundial, que ha sido considerado tanto una pandemia como un problema de salud pública por parte de la Federación Internacional de Diabetes (IDF). En términos generales, se estima que aproximadamente entre el 15% y el 25% de la población mundial sufre de prediabetes, variando su prevalencia dependiendo de la región, género, edad y etnia del grupo humano estudiado. (Rosas-Saucedo, 2017)

Los estudios que mayor impacto han tenido a nivel mundial, relacionados con el diagnóstico de PD, fueron un cálculo de riesgo de contraer tanto DM2 como PD y la generación de un índice de PD. Estos estudios fueron desarrollados en los años 2003 y 2008 respectivamente con variables no invasivas o ignorando factores de riesgo que no eran relevantes para la época, como por ejemplo el índice HOMA que mide, en términos generales, el grado de resistencia a la insulina.

En el presente trabajo de investigación se desea crear un modelo de clasificación que permita identificar de manera temprana potenciales complicaciones que pueden desembocar en el aparecimiento de la DM2. Esto se logra mediante el reconocimiento temprano y clasificación de los factores de riesgo que inciden en la PD, generando así una alternativa para la disminución del gasto público en la remediación de sus síntomas y su posible aplicación como política pública de prevención.

Este trabajo de titulación nace con la intención de generar un aporte significativo desde el punto de vista cuantitativo al creciente estudio y análisis de la prediabetes, proponiendo el desarrollo de un modelo de clasificación de las condiciones clínicas que definen su comportamiento. Este modelo se desarrollará basado en los resultados medidos desde el año 2011 hasta el año 2016 de la encuesta estadounidense NHANES (National Health and

Nutrition Examination Survey), los cuales miden diversos criterios de salud y nutrición de los encuestados, utilizando el algoritmo de clasificación C4.5 de inteligencia artificial (AI), que basado en un set de datos, bien tratado, induce un árbol de clasificación con cortes tanto cuantitativos como categóricos.

Esta aproximación metodológica nos permite hallar un árbol de clasificación con cualidades optimistas, que presenta un nivel de predicción del 77.8% y posee un diseño amigable con el usuario, independientemente de su conocimiento clínico y técnico. Su potencial predictivo, en comparación con los estudios previamente realizados, se incrementa considerablemente, debido a que solo requiere del ingreso de 2 variables provenientes de exámenes de laboratorio y 4 variables fácilmente observables, convirtiendo a la aplicación desarrollada en una herramienta que puede ser utilizada tanto por pacientes que tienen curiosidad de conocer su estado de salud, como por facultativos del área de la salud.

1.2 Estructura del Trabajo

A continuación, se analizará un resumen breve de los contenidos existentes en éste documento, con la finalidad de brindar al lector un hilo conductor que le permita navegar por sus diferentes secciones, acercándose a puntos específicos del desarrollo temático en función de sus intereses y necesidades.

Capítulos existentes en el documento

Para el desarrollo del trabajo de investigación, se han desarrollado los siguientes capítulos:

- Capítulo 1: Introducción

Esta es una visión general de la investigación, donde se presenta al lector el tema propuesto. Aquí se plantea el problema central del trabajo, con respecto a la prediabetes y la encuesta norteamericana NHANES.

En términos generales, se busca responder la siguiente pregunta:

¿Es posible utilizar un algoritmo de clasificación sobre data asociada a la prediabetes de manera satisfactoria?

- Motivación y Planteamiento

En este apartado, se explican las complicaciones existentes debido al padecimiento de DM2 y la necesidad de prevención de su condición clínica previa, conocida como prediabetes y la importancia de su análisis hacia la integración de políticas públicas de prevención.

- Capítulo 2: Contexto y Estado del Arte

En esta sección, se analizan dos puntos fundamentales para definir un contexto que nos permita entender el comportamiento del fenómeno:

- Referencias que tengan una incidencia, directa o indirecta, sobre el tema de investigación.

- Contexto relacionado al tema de investigación.

En términos generales, se busca abordar las siguientes referencias:

- Uso de técnicas de inteligencia artificial (AI) en temas relacionados con la prediabetes.

- Criterios de diagnóstico de la prediabetes.

- Factores de riesgo detectables en exámenes clínicos o variables fácilmente observables asociadas a la prediabetes.

- Trabajos Relacionados

En esta sección, se muestran investigaciones similares tanto sobre DM2 como PD que permiten comprender la motivación técnica para desarrollar este proceso investigativo. Adicionalmente, se analiza una investigación que invita al uso de algoritmos de AI para el análisis de procesos metabólicos de los cuales se hayan recolectado grandes volúmenes de información.

- Implicaciones Sobre el Desarrollo del Trabajo

Aquí, se muestra la conclusión obtenida de las investigaciones relacionadas que guían el desarrollo de la investigación hacia la generación de un árbol de clasificación que permita a los usuarios definir su condición clínica entre las categorías "Normal" y "Prediabetes".

- Capítulo 3: Objetivos y Metodología de Trabajo

En esta sección, se definen los objetivos concretos, general y específicos, que posee el proyecto, mismos que deben ser resueltos o rechazados a lo largo del trabajo, particularmente en las conclusiones.

- Capítulo 4: Desarrollo del Modelo de Clasificación

En esta sección se detallan las herramientas y conceptos que se utilizarán para el desarrollo de la investigación en curso.

- Tecnologías Implicadas

En este apartado, se detalla el software que fue utilizado para todas las etapas de captura, almacenamiento, limpieza, procesamiento, transformación y clasificación de la información junto con la motivación que llevó a su elección como herramientas de trabajo.

- Contenidos Implicados

En esta sección, se explican en detalle los fundamentos teóricos que permiten el correcto desarrollo de la investigación.

- Manual Para la Aplicación de la Metodología

En este apartado, se detallan los pasos a seguir para lograr aplicar la metodología diseñada en cualquier población.

- Capítulo 5: Evaluación de la Implementación de la Metodología

En esta sección se realizará una explicación del experimento aplicado a la población, junto con un proceso de validación de la aplicación de la metodología. En este caso en particular, se utilizará un proceso de validación cruzada, donde se destina una parte de la población encuestada como data de aprendizaje y otra como data de prueba. Con el proceso anteriormente mencionado, se consigue obtener diversos criterios de eficiencia en la aplicación del algoritmo.

- Capítulo 6: Conclusiones y Trabajo Futuro

En este apartado se cuentan experiencias sobre el desarrollo del proyecto y se delinea de manera superficial un plan de profundización y extensión del tema de investigación.

- Relevancia y Alcance de la Contribución

En esta sección, se explica el aporte generado a la ciencia por medio del desarrollo de la investigación.

- Conclusiones

En este apartado, se verifica el cumplimiento de los objetivos y se cuentan experiencias relevantes sobre el desarrollo del proceso investigativo.

- Recomendaciones

En esta sección, se brindan sugerencias importantes para la implementación de la metodología en otras circunstancias y/o poblaciones.

- Trabajo futuro

En esta sección, se plantean ideas que pueden permitir el desarrollo futuro de la investigación en distintas líneas de análisis como la regresión logística y generación de índices.

- Referencias

En este apartado, se citan en formato APA todas las investigaciones que aportan, de manera directa o indirecta, al desarrollo del proceso investigativo.

- Anexos

En este apartado, se adjuntan documentos complementarios importantes para la comprensión de la investigación.

2. Contexto y Estado del Arte

En este capítulo, se presentan investigaciones relacionadas con la temática prevista para la investigación que permitan definir un contexto de trabajo, al igual que un análisis relativo a la incidencia de dichas investigaciones sobre el proyecto en desarrollo.

2.1 Trabajos Relacionados

A continuación, se expondrán las ideas centrales de artículos científicos relacionados, de manera directa o indirecta, con la temática de la investigación en curso y cuyos resultados aportan de manera favorable a la evolución del proceso investigativo:

Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling (Cuperlovic-Culf, 2018)

En este documento se describen las potencialidades de los algoritmos de AI en el estudio de procesos metabólicos, basados en el uso de datos recopilados a través del diseño de experimentos minuciosamente planificados. Estos experimentos se diseñan con la intención de poder realizar análisis profundos que definan patrones de comportamiento de sus variables fundamentales y que permitan predecir resultados de comportamientos complejos futuros que integren múltiples variables interrelacionadas.

Cabe aclarar que la diferencia entre Inteligencia Artificial y Aprendizaje de Máquina (ML) radica en el alcance de cada rama y el hecho de que la AI corresponde a una categoría más amplia de análisis. Esto puede observarse con mayor detalle y simplicidad en la Figura 1, misma que se muestra a continuación:

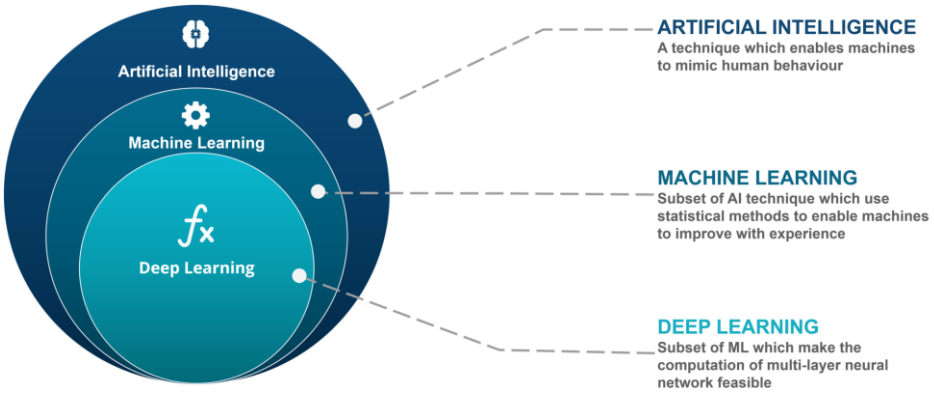


Figura 1. Inteligencia Artificial vs Aprendizaje de Máquina vs Aprendizaje Profundo

Fuente: (Edureka, 2018)

Adicionalmente, este artículo nos invita a utilizar procedimientos y algoritmos de AI para el análisis de fenómenos relacionados al metabolismo y la salud, tales como los árboles de clasificación, máquinas de soporte vectorial, lógica difusa, entre otros. (Cuperlovic-Culf, 2018)

Predicting Metabolic Syndrome Using Decision Tree and Support Vector Machine Methods (Karimi-Alavijeh, 2016)

Este artículo documenta un estudio, de aprendizaje supervisado, sobre el Síndrome Metabólico con la ayuda de un árbol de decisión y una máquina de soporte vectorial, comparando sus resultados y definiendo las ventajas existentes con el uso de cada uno de los métodos.

Al finalizar el análisis, concluye que para este caso en particular la máquina de soporte vectorial presenta una mayor precisión en el proceso de clasificación, pero también se reconoce la claridad y simplicidad de un árbol inducido por un algoritmo de inteligencia artificial. Al no existir diferencias significativas en la comparación de sus índices de calidad, se deja planteada de manera implícita la posibilidad de utilizar uno u otro método. (Karimi-Alavijeh, 2016)

Los resultados obtenidos al comparar los índices definidos por los investigadores pueden observarse en la Figura 2.

Figura 2. Sensibilidad, especificidad y precisión de los métodos de aprendizaje aplicados en los datos balanceados del estudio de la cohorte de Isfahan

Fuente: (Karimi-Alavijeh, 2016)

The Diabetes Risk Score (Lindström, 2003)

Este artículo busca diseñar una herramienta práctica, no invasiva, para la predicción del riesgo de contraer diabetes tipo2.

Para este estudio se seleccionó una muestra de personas que se encontrasen entre los 35 y 64 años de edad, que no hubieren utilizado ningún medicamento para el control de la diabetes y sobre los cuales se pudiera realizar un estudio longitudinal de 10 años de duración.

Las variables utilizadas en este estudio fueron las siguientes:

- Edad,
- Índice de masa corporal,
- Circunferencia abdominal,
- Uso de medicamentos contra la hipertensión,
- Uso de medicamentos para controlar los niveles de glucosa en sangre,
- Consumo de frutas y vegetales,

Al finalizar el estudio, se pudo concluir que el “Diabetes Risk Score” es una herramienta simple, rápida, de bajo costo, no invasiva y confiable para identificar a individuos con un alto riesgo de contraer diabetes tipo2. (Lindström, 2003)

La herramienta utilizada para la evaluación del riesgo de padecer de DM2, puede observarse en la Figura 3 o en tamaño completo en el Anexo A.

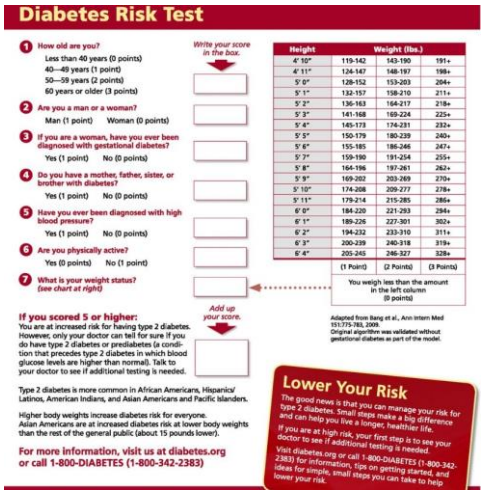


Figura 3. ADA – Test de riego de diabetes

Fuente: (ADA, 2017)

The Diabetes Risk Calculator (Heikes, 2008)

Para el desarrollo de esta investigación se utilizan los resultados de la encuesta NHANES desde el año 1999 hasta el año 2004. El objetivo de este artículo es el de desarrollar una herramienta simple, no invasiva, que permita a la población estadounidense, calcular la probabilidad de tener diabetes o prediabetes no diagnosticadas.

Para el desarrollo de este proyecto, se utiliza una regresión logística y un árbol de clasificación; siendo preferido el árbol por su mayor nivel de precisión en el proceso de clasificación y su simplicidad de uso, mismo que puede observarse en el Anexo B.

Las variables utilizadas para el desarrollo del artículo son las siguientes:

- Edad,
- Circunferencia abdominal,
- Diabetes gestacional,
- Peso,
- Grupo étnico,
- Hipertensión,
- Historia familiar de diabetes,
- Actividad física.

Adicionalmente, se reconoce a la Diabetes Risk Calculator como la única herramienta técnicamente validada, no invasiva, que permite detectar la prediabetes y diabetes no diagnosticadas. (Heikes, 2008)

A Simple Score Model to Assess Prediabetes Risk Status Based on the Medical Examination Data (Ouyang, 2016)

Este artículo plantea la generación de un índice de riesgo que permita detectar a los pacientes con alto riesgo de convertir su estado de PD en DM2, basado en los distintos factores de riesgo seleccionados.

Para el desarrollo de esta investigación se utiliza la data levantada por el ¬ hospital universitario del “Harbin Institute of Technology” de China. Adicionalmente, se utiliza una regresión logística binaria que permite definir la participación de los diversos factores de riesgo seleccionados, hallando su peso ponderado.

Al finalizar la investigación se determina que el índice desarrollado es de fácil uso e interpretación al detectar el riesgo de convertir estados de PD a DM2 y que junto con un programa de intervención bien desarrollado puede detener o demorar la conversión negativa de una patología en otra. (Ouyang, 2016)

2.2 Implicaciones Sobre el Desarrollo del Trabajo

Considerando las ideas fundamentales planteadas anteriormente, tenemos las siguientes premisas para el desarrollo de la investigación:

- Es válido e importante realizar análisis diversos, utilizando algoritmos de AI en temas asociados al metabolismo.
- Existen métodos no invasivos que permiten tener niveles aceptables de clasificación en temas relacionados con el metabolismo.
- Para tener mayores niveles de eficiencia en la clasificación de patologías asociadas al metabolismo, es importante incorporar ciertos exámenes clínicos.
- Un árbol de clasificación permite identificar de manera adecuada patologías, no diagnosticadas, tales como la diabetes y prediabetes.
- Es posible cuantificar la probabilidad y el riesgo de padecer tanto de PD como de DM2.
- Se realizan mayor cantidad de estudios asociados al análisis de la DM2 que a la PD.

Realizando un análisis comparativo de los artículos científicos presentados anteriormente, es posible generar una tabla que considere los factores relevantes mencionados en cada uno de ellos, en contraste con la propuesta de investigación actual, tal como se puede observar en la Tabla 3.

Tabla 3. Análisis comparativo de trabajos relacionados (Elaboración propia)

Análisis comparativo					
Referencia	**Inteligencia Artificial**	**Variables Observables**	**Variables Laboratorio**	**Prediabetes**	**Diabetes**
Cuperlovic-Culf, 2018	x	-	-	-	-
Karimi-Alavijeh, 2016	x	x	x	-	-
Lindström, 2003	-	x	-	-	x
Heikes, 2008	x	x	-	x	x
Ouyang, 2016	-	x	x	x	-
Investigación actual	**x**	**x**	**x**	**x**	**-**

Como se puede observar en la Tabla 3, ninguna de las investigaciones consideradas lleva a cabo un análisis que satisfaga todos los criterios seleccionados en el análisis comparativo, mismos que definen un contexto favorable para un estudio, en AI, con alto poder predictivo en el campo de la PD. Por otro lado, la investigación planteada en el presente trabajo no solo cubre los criterios mencionados, sino que se enfoca únicamente en la prediabetes, lo que le brinda especificidad en su análisis, al no evaluar factores asociados a otra patología, relativamente similar, como es la DM2.

Consecuentemente, se realizará un árbol de clasificación de prediabetes, con información correspondiente a los años 2010 al 2016. El árbol mencionado utilizará tanto variables fácilmente observables, no invasivas, como también mediciones de laboratorio. El uso de factores de riesgo con características invasivas y no invasivas permitirá incrementar la precisión en la clasificación y tener una herramienta diseñada con información actualizada e índices como el “Homeostatic Model Assessment (HOMA)”.

3. Objetivos y Metodología de Trabajo

A continuación, se delimitan los objetivos que guiarán este proceso investigativo, tanto a nivel general como específico. De igual manera, se delinea la metodología de trabajo junto con un diagrama de infraestructura que permitirá comprender de forma gerencial el flujo de información desde que es tomada de la fuente, hasta la última etapa de procesamiento y visualización.

3.1 Objetivo General

Generar un árbol de clasificación de prediabetes que permita, a través de variables fácilmente observables y exámenes clínicos definidos, ubicarse de manera rápida y sencilla en un nivel específico de diagnóstico.

3.2 Objetivos Específicos

- Identificar las condiciones clínicas que inciden en el diagnóstico de la prediabetes.
- Permitir a los pacientes que cuenten con exámenes clínicos definidos, ubicarse en un punto específico del modelo de clasificación de la prediabetes.
- Brindar un criterio básico de diagnóstico que motive a la población a buscar asistencia de salud preventiva dependiendo del diagnóstico de prediabetes provisto y su probabilidad.
- Diseñar un manual detallado de cómo aplicar la metodología en otras poblaciones.

3.3 Metodología de Trabajo

A continuación, se detallan los pasos a seguir para el desarrollo de este tema de investigación. Esto se realiza con la intención de comprender el flujo secuencial de actividades que se requieren para replicar esta metodología con otra población, independientemente de los factores técnicos que inciden en su evolución, esto también puede observarse de forma resumida en el diagrama de infraestructura de la Figura 4.

- Captura y almacenamiento de la data.

En este proceso se busca capturar la información directamente desde la fuente y almacenarla de manera automática en una base de datos, evitando la manipulación humana, garantizando así la integridad y fidelidad de la data.

Para este proceso se utiliza una integración del software R Project con SQL Server, con la intención de descargar la información directamente desde la página web que contiene las tablas de la encuesta NHANES e ingresarla en una base de datos estructurada, respectivamente.

- Limpieza y procesamiento de la data.

En este proceso se limpia la data de categorías no homologadas o mal codificadas, también se analizan temas de completitud, precisión y consistencia.

Para este proceso se diseñan vistas en SQL Server que corrijan y codifiquen la información existente en las tablas de la encuesta NHANES.

- Transformación de la data

En este proceso se recodifica y estructura la información con el formato requerido por el software que brinda la posibilidad de generar los modelos de clasificación.

Para este proceso se utilizan vistas que realizan cruces de información e integren la data contenida en las tablas, para lo cual se utilizó un ERD con forma de estrella.

- Aplicación del modelo de clasificación

En este proceso se parametriza el algoritmo que induce el árbol de clasificación y se verifican sus resultados.

Para este proceso se realiza una integración entre SQL Server y Weka, con la intención de alimentar el algoritmo que induce el árbol de clasificación de manera automática.

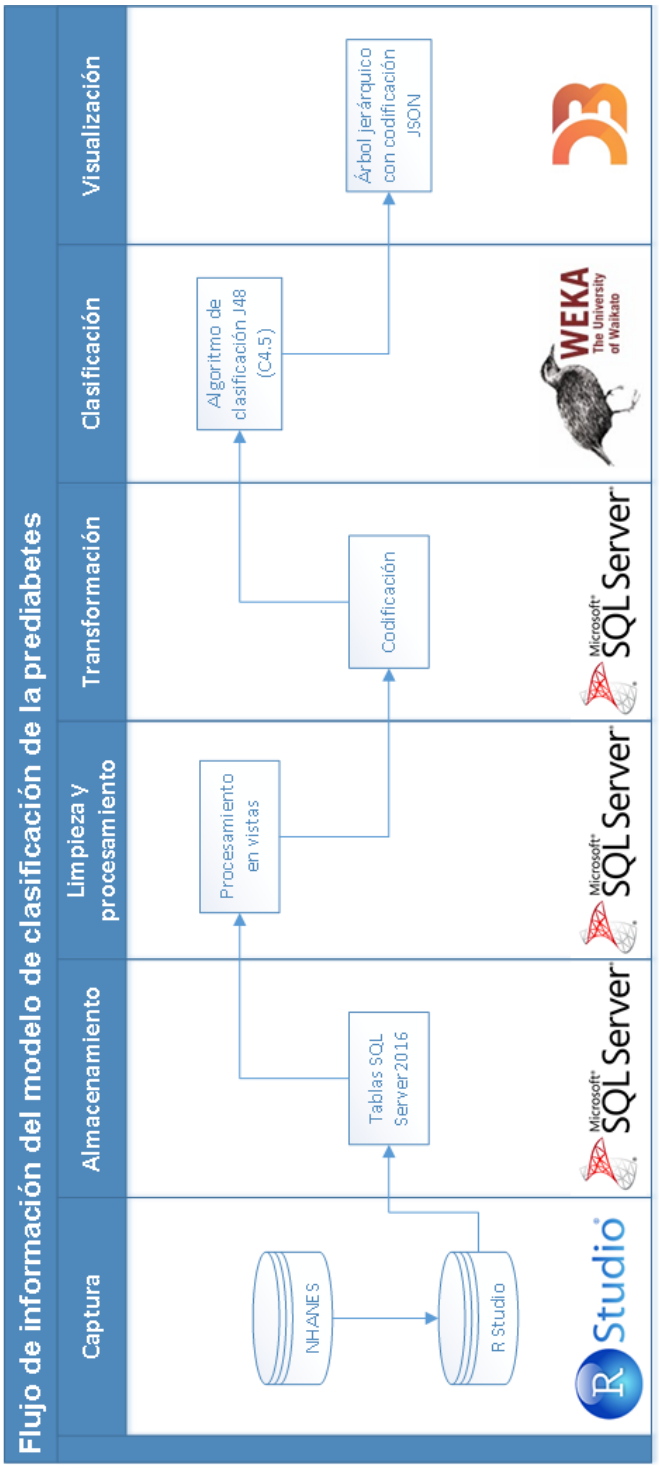


Figura 4. Diagrama de Infraestructura y flujo de información

Fuente: Elaboración propia.

4. Desarrollo del Modelo de Clasificación

A continuación, se detalla el software que fue utilizado para las etapas de captura, almacenamiento, limpieza, procesamiento, transformación y clasificación de la información. De igual manera, se explican a profundidad los fundamentos teóricos que permiten el correcto desarrollo de la investigación y se desarrolla un manual de usuario en el que se detallan los pasos seguidos para conseguir aplicar la metodología diseñada en cualquier población.

4.1 Tecnologías Implicadas

Para el desarrollo técnico de esta investigación se utilizan ciertas herramientas de software para las distintas etapas de su evolución, cuyas descripciones y justificación de uso se muestran a continuación:

- R Project 3.5.0

R[1] es un sistema de código abierto que permite ser ejecutado tanto sobre Windows, Mac OSX y Linux. Puede ser utilizado tanto como entorno de software o como lenguaje de programación, brindando grandes capacidades para realizar análisis matemático y estadístico. Nace como una implementación gratuita y de código libre del lenguaje y sistema S, que permite a los usuarios ejecutar una interfaz de programación, desarrollar gráficos complejos, utilizar un depurador, obtener acceso a ciertas funciones del sistema y ejecutar programas almacenados en archivos de comandos.

Se eligió por encima de un software tradicional de ETL (Extract, Transform and Load), debido a la versatilidad que brinda al trabajar con archivos de distintas extensiones, gracias al uso del paquete “foreign”. Adicionalmente, permite realizar una integración con SQL Server mediante el uso del paquete “RODBC”, que gestiona una comunicación bidireccional fluida entre ambos sistemas.

[1] https://www.r-project.org/

- RStudio 1.1.453

RStudio [2]es un entorno de desarrollo integrado (IDE) para R Project. Incluye una consola, un editor de comandos que admite la ejecución directa de código, así como herramientas para el trazado, historial de ejecución, depuración y la administración del espacio de trabajo. Se encuentra disponible en ediciones de código abierto y comercial, que se pueden ejecutar sobre Windows, Mac OSX y Linux.

Se eligió debido a la facilidad de uso que brinda sobre la herramienta R Project al ser una interfaz de usuario muy amigable y completa, que cubre los requerimientos de usuarios más comunes.

- Microsoft SQL Server Management Studio 2016

SQL Server Management Studio (SSMS[3]) es un entorno integrado para administrar cualquier infraestructura SQL, desde SQL Server hasta Azure SQL Database. Proporciona herramientas para configurar, monitorear y administrar instancias de SQL. Se recomienda el uso de SSMS para implementar, monitorear y actualizar los datos de los componentes utilizados por sus aplicaciones, así como crear consultas y archivos de comandos. Adicionalmente, se puede utilizar SSMS para consultar, diseñar y administrar bases de datos y almacenes de datos, sea que estén, en un computador local o en la nube.

Se seleccionó como herramienta de consulta de bases de datos debido su versatilidad y robustez al trabajar con grandes volúmenes de datos en bases estructuradas. También es importante mencionar que el seleccionar esta herramienta nos permite hacer uso del soporte técnico brindado por parte de la empresa Microsoft al igual que utilizar los abundantes aportes desarrollados en línea para esta herramienta.

[2] https://www.rstudio.com/
[3] https://www.microsoft.com/es-es/sql-server

- Weka 3.8.2

Weka [4](Waikato Environment for Knowledge Analysis) es una aplicación de código abierto, escrita en Java, desarrollada por la Universidad de Waikato con la intención de realizar estudios de AI.

Weka es una colección de algoritmos de AI que contiene herramientas para la preparación, clasificación, regresión, agrupamiento, reglas de asociación y visualización de la información cargada y/o procesada.

Se seleccionó como herramienta de AI debido a que su diseño brinda una interfaz simple, amigable, que guarda grandes potencialidades técnicas y presenta resultados muy detallados. Adicionalmente, la Universidad de Waikato cuenta con un registro video gráfico muy extenso y completo que cubre muchas de las necesidades de análisis de AI.

- HTML 5

HTML [5](Hypertext Markup Language) es el lenguaje estándar para crear páginas y aplicaciones web, que junto con las CSS (Cascading Style Sheets) y JavaScript forman la piedra angular sobre la cual se sustenta el diseño web, tal como se puede observar en la Figura 5.

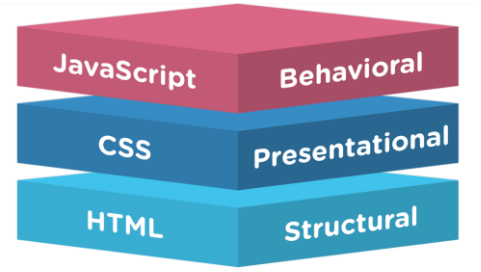


Figura 5. Relación entre HTML, JavaScript, CSS

Fuente: (Grimmett, 2017)

Los elementos HTML son los componentes básicos de las páginas web. El lenguaje HTML proporciona un medio para crear documentos estructurados al denotar semántica estructural para textos como encabezados, párrafos, listas, enlaces, citas y otros elementos.

HTML 5 es la última versión del lenguaje HTML. Esta versión, fue publicada en octubre de 2014, cuenta con nuevos elementos, atributos y comportamientos en relación a versiones

[4] https://www.cs.waikato.ac.nz/ml/weka/
[5] https://developer.mozilla.org/es/docs/HTML/HTML5

anteriores. También dispone de un conjunto más amplio de tecnologías que permite a los sitios Web y a las aplicaciones ser más diversas y de gran alcance.

Se seleccionó con la intención de desarrollar una aplicación web que permita desarrollar una interfaz de comunicación con el usuario final que se encuentre interesado en evaluar su condición de prediabetes mediante la utilización del proceso de clasificación desarrollado en este proyecto de investigación.

- JavaScript

JavaScript [6](JS) es un lenguaje de programación interpretado, orientado a objetos, que permite realizar actividades complejas en una página web, como mostrar actualizaciones de contenido, interactuar con mapas, animaciones, gráficas, etc. JavaScript es el lenguaje que define el comportamiento de la página web y su interacción con el usuario.

Se eligió para el desarrollo de la investigación debido a que se requería generar interacción avanzada con el usuario final, específicamente para generar el árbol de clasificación en la aplicación web diseñada con el uso del lenguaje HTML.

- CSS

CSS [7](Cascading Style Sheets) es un lenguaje de hojas de estilo creado para controlar la apariencia HTML y XHTML. Permite separar los contenidos de la presentación de páginas web complejas, mejorando la accesibilidad del documento y simplificando su mantenimiento.

Se eligió con la intención de mejorar la apariencia estética de la aplicación web a través de la creación de clases que brinden vistosidad a la herramienta.

- D3 v5

D3 [8](Data Driven Documents) es una librería de diseñada en JavaScript para producir, a partir de datos locales o almacenados en línea, visualizaciones dinámicas e interactivas para navegadores web, permitiendo tener control completo sobre el resultado visual terminado.

[6] https://www.javascript.com/
[7] https://developer.mozilla.org/es/docs/Web/CSS
[8] https://d3js.org/

Se eligió para el desarrollo de la aplicación web debido a que brinda enormes potencialidades de diseño web para la interacción con el usuario. Es mediante la librería D3 de JavaScript que se genera la visualización final del árbol de clasificación.

- Brackets 1.13

Brackets [9]es un software gratuito y de código abierto enfocado en la edición de código fuente para desarrollo web. Brackets es una aplicación multiplataforma, escrita en JavaScript, HTML y CSS, que se puede ejecutar sobre Windows, Mac OSX y Linux.

El objetivo principal de Brackets es su funcionalidad de edición y ejecución de código HTML, CSS y JS en vivo, motivo por el cual se eligió como elemento principal para el desarrollo de la aplicación web.

4.2 Contenidos Implicados

En este apartado se exponen investigaciones que explican los motivos por los cuales los factores de riesgo considerados son relevantes en el estudio de la PD, justificando así su uso para el proceso de clasificación con el algoritmo C4.5 de AI.

4.2.1 Diagnóstico de la Prediabetes

La prediabetes es un trastorno metabólico en el que el nivel de la glucosa en sangre se encuentra por encima de lo considerado normal, pero no lo suficientemente alto como para que sea considerado como DM2. (Rosas-Saucedo, 2017). Según la literatura clínica tradicional, existen criterios claros que determinan el diagnóstico de PD, mismos que se validan con las pruebas de laboratorio que se observan en la Tabla 4.

[9] http://brackets.io/

Tabla 4. Diagnóstico Prediabetes (ADA, 2015)

Diagnóstico Prediabetes		
Hemoglobina glucosilada - A1C		
Mide el nivel promedio de glucosa en la sangre durante los últimos 2 o 3 meses	5.7 - 6.4	%
Glucosa plasmática en ayunas		
Generalmente se realiza a primera hora en la mañana, antes del desayuno, y mide el nivel de glucosa en la sangre cuando el paciente está en ayunas	100 - 125	mg/dl
Prueba de tolerancia a la glucosa oral		
Mide el nivel de glucosa en la sangre antes de ingerir una solución con 75 gramos de glucosa y 2 horas después de tomarla	140 - 200	mg/dl

Es importante recalcar que en ciertas ocasiones los médicos prefieren una u otra prueba dependiendo de la historia clínica del paciente, por lo que estas pruebas no son mutuamente excluyentes, por el contrario, son complementarias y pueden enriquecer el contexto para el diagnóstico clínico del especialista.

4.2.2 Factores de Riesgo que Inciden en la Prediabetes

Si bien, el diagnóstico de PD se encuentra claramente definido en la literatura, uno de los objetivos primordiales de este proyecto de investigación consiste en hallar los factores de riesgo que potencialmente pueden derivar en el padecimiento de la patología y definir su interacción a través de un proceso de clasificación basado en el algoritmo C4.5 de AI. Para la generación del árbol de clasificación de prediabetes, que se genera con el algoritmo anteriormente mencionado, se utilizaron 10 factores de riesgo, mismos que sin ningún orden jerárquico específico se listan a continuación:

- Género,
- Edad,
- Diámetro sagital abdominal,
- Insulina,
- Triglicéridos,
- Historia familiar de diabetes,

- Actividad física diaria,
- Horas de sueño diarias,
- Alanina aminotransferasa (ALT),
- Índice HOMA.

Estos factores de riesgo fueron seleccionados en función de su innovación en diagnóstico y relevancia sobre la patología de PD al igual que DM2. Cada uno de los factores de riesgo se encuentra respaldado en los siguientes documentos oficiales de entidades reconocidas como referentes en su área de experticia o en investigaciones científicas relevantes:

Consenso de Prediabetes

En este documento desarrollado por la Asociación Latinoamericana de Diabetes (ALAD) se define la posición que asume dicha institución con respecto a la PD y DM2. Aquí se detallan de forma explícita las definiciones tomadas sobre PD y DM2, su repercusión a nivel socioeconómico, diagnóstico, algoritmos de detección, factores de riesgo, tratamiento, entre otros. Los principales factores de riesgo o manifestaciones clínicas mencionadas en el documento pueden observarse en la Tabla 5.

Tabla 5. Manifestaciones clínicas y de laboratorio que orientan al diagnóstico de prediabetes (Friege, 2014)

Manifestaciones clínicas y de laboratorio que orientan al diagnóstico de prediabetes
Mediciones clínicas y de laboratorio
Índice de masa corporal - IMC
Triglicéridos
HDL
Hipertensión arterial
A1C
Diabetes gestacional
Hijos macrosómicos
Síndrome de ovarios poliquísticos
Enfermedad cardiovascular
Acantosis nigricans
Otras mediciones
Género
Edad
Circunferencia abdominal
Bajo peso al nacer

Adicionalmente, se define un sistema de puntajes sobre variables adicionales que se analizan tanto sobre la población adulta como pediátrica, tal como se observa en la Tabla 6.

Tabla 6. Sistema de puntuación por tipo de población (Friege, 2014)

Sistema de puntuación por tipo de población	
Población pediátrica	
IMC	0 - 4
Historia Familiar con DM2	0 - 2
Signos de resistencia a la insulina	2 - 4
Población adulta	
IMC	1 - 3
Edad	1 - 3
Cintura abdominal	3 - 4
Hipertensión	2
Historia de glucosa elevada	5
Sedentarismo	1
Dieta pobre	1

Sagital Abdominal Diameter, but not Waist Circumference is Strongly Associated with Glycemia, Triacilglycerols and HDL-C Levels in Overweight Adults.

Esta investigación realiza un análisis en adultos de aproximadamente 54 años y un índice de masa corporal (IMC) de 30.5, buscando hallar la relación existente entre variables como el diámetro sagital abdominal, circunferencia de la cintura, ácido úrico, colesterol LDL, entre otras, con la hiperglucemia y otros factores asociados al síndrome metabólico.

Al final de la investigación se determina que la mayor correlación con la hiperglucemia se genera con el diámetro sagital abdominal y no con la circunferencia de la cintura u otros factores, brindando un aporte significativo a los análisis relacionados el metabolismo al mejorar sus estimaciones. (Pimentel, 2011)

Fasting Tests of Insulin Secretion and Sensitivity Predict Future Prediabetes in Japanese With Normal Glucose Tolerance. (Onishi, 2010)

El propósito de este tema de investigación es el de definir si la medición de insulina basal y el índice HOMA pueden ser considerados como factores de riesgo independientes para el diagnóstico de prediabetes. Su análisis se realiza entre personas con un valor de glucosa plasmática en ayunas de 75 mg/dl por un período de entre 5 y 6 años, realizando diversos análisis a lo largo del plazo establecido.

Una vez finalizada la investigación, se determina que la baja secreción de insulina y baja la sensibilidad a la insulina son factores de riesgo independientes en el análisis de la prediabetes, por lo que deben ser tratados como dos variables distintas. (Onishi, 2010)

Increased Liver Markers are Associated With Higher Risk of Type 2 Diabetes. (Ko, 2015)

Este proceso investigativo, busca hallar la asociación existente entre las pruebas de función hepática, transaminasas, con el riesgo de padecer DM2 y con las alteraciones de las mediciones de glucosa en ayunas. Para este análisis se utiliza la aspartato aminotransferasa (AST), la alanina aminotransferasa (ALT) y la gamma-glutamil transpeptidasa o GGT. Para este estudio se seleccionan pacientes que tengan más de 30 años y se utiliza una regresión logística para definir la asociación entre las variables.

Al finalizar la investigación se concluye que las transaminasas ALT, GGT y AST/ALT son independientes, constituyendo factores de riesgo de padecer DM2 e implicando alteraciones en las mediciones de glucosa en ayunas. (Ko, 2015)

Factores de Riesgo Para el Síndrome Metabólico en una Población con Apnea del Sueño; Evaluación en un Grupo de Pacientes de Granada y Provincia; Estudio GRANADA. (Valenza, 2012)

En este estudio se busca hallar la relación causal de las alteraciones del sueño sobre los problemas metabólicos. Para este estudio se analizaron 1016 casos de pacientes que visitaron el Hospital Universitario San Cecilio de Granada por sospecha de apnea de sueño.

Al finalizar el estudio se encontró una correlación significativa entre la apnea del sueño y la saturación de oxígeno nocturna con diferentes alteraciones metabólicas, permitiendo definir que los sujetos con apnea de sueño poseen significativamente más riesgo de desarrollar síndrome metabólico. (Valenza, 2012)

Al analizar las investigaciones científicas previamente mostradas, se obtienen tanto criterios de diagnóstico de PD como factores de riesgo claramente identificados para esta patología. Adicionalmente, se añaden dos variables poco convencionales que inciden en patologías evolutivas de la PD como son la DM2 y el Síndrome Metabólico (SM).

Una vez realizado el análisis de estos documentos, que respaldan la selección de variables a utilizarse para la generación del árbol de clasificación inducido mediante el uso del algoritmo C4.5 de AI, se satisface el primer objetivo específico planteado para este proyecto de investigación, mismo que se muestra a continuación:

“Identificar las condiciones clínicas que inciden en el diagnóstico de la prediabetes.”

4.2.3 Árboles de Clasificación Inducidos con el Algoritmo C4.5

El algoritmo de clasificación C4.5 nace como una mejora del algoritmo ID3, con su principal diferencia marcada por el uso de variables cuantitativas y no solo categóricas o nominales, al igual que por un método de selección de atributos mejorado. Tanto en el algoritmo ID3, como en el algoritmo C4.5 los análisis inician por el cálculo de la entropía, mismo que se muestra en la siguiente ecuación.

$$Entropía\ (E) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

El algoritmo C4.5 genera un árbol de clasificación a partir de la data, mediante particiones realizadas de forma recursiva con una estrategia en profundidad y según un método de selección de atributos basado en la **proporción de ganancia**, mismo que brinda mejores resultados que la medida de **ganancia de información**, cuya ecuación se muestra a continuación:

$$Ganancia\ (E, A) = \ Entropía(E) - \sum_{v \in V_a} \frac{|E_v|}{E} \times Entropía(E_v)$$

La proporción de ganancia busca compensar el hecho de que un atributo pueda tener muchos casos asociados, dividiendo la **ganancia de información** por la medida denominada **información de la división**, tal como se muestra en la siguiente ecuación:

$$Información\ de\ la\ División\ (E, A) = -\sum_{i=v_i}^{v_n} \frac{|E_i|}{|E|} \times \log_2 \frac{|E_i|}{|E|}$$

Siendo E_i , E_{i+1} , … E_n las diferentes particiones de ejemplos que resultan de dividir el conjunto E de ejemplos, teniendo en cuenta los valores v_i … v_n que toma el atributo, respectivamente.

Así la **proporción de ganancia** se calcula como la relación entre **la ganancia de información** y **la información de la división**, según se muestra en la siguiente ecuación:

$$Proporción\ de\ Ganacia\ (E, A) = \frac{Ganancia\ de\ información\ (E, A)}{Información\ de\ la\ División\ (E, A)}$$

Al evaluar este cociente, se consigue penalizar los atributos que poseen muchos valores que se distribuyen uniformemente entre los ejemplos en grupos de igual tamaño. (Quinlan, 1993)

La Figura 6 muestra un ejemplo simple del tipo de decisiones y estructura de un árbol inducido con el algoritmo C4.5 con variables cuantitativas de prueba.

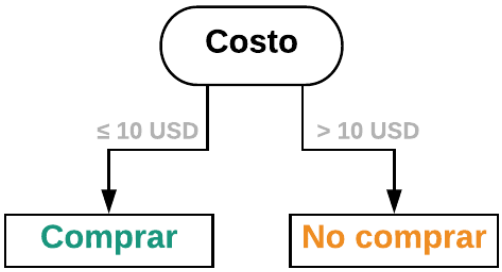


Figura 6. Árbol inducido con el algoritmo C4.5

Fuente: Elaboración propia

4.2.4 Creación de un Modelo Entidad-Relación

El diagrama entidad-relación, también conocido como modelo entidad-relación (ERD) es un diagrama de flujo que ilustra cómo las entidades interactúan entre sí dentro de un sistema de información. Los ERD se desarrollan con la intención de facilitar el diseño de las bases de datos, permitiendo la generación de un esquema que representa la lógica general de un sistema de información, sea este normalizado o no. Los ERD son modelos semánticos que representa el significado de los datos, empleando conjuntos de entidades, relaciones y atributos. (Date, 2001)

Convencionalmente y con fines didácticos se habla de ERD de tipo estrella y copo de nieve, refiriéndose a la profundidad que poseerán las tablas o entidades de dimensiones que se almacenarán en la base de datos. En ambos casos se ubica una tabla central o tabla de hechos que posee una clave primaria (PK) para generar las relaciones con las tablas de dimensiones. La Figura 7 muestra las diferencias gráficas, en formas, que existen entre los ERD de tipo estrella y copo de nieve.

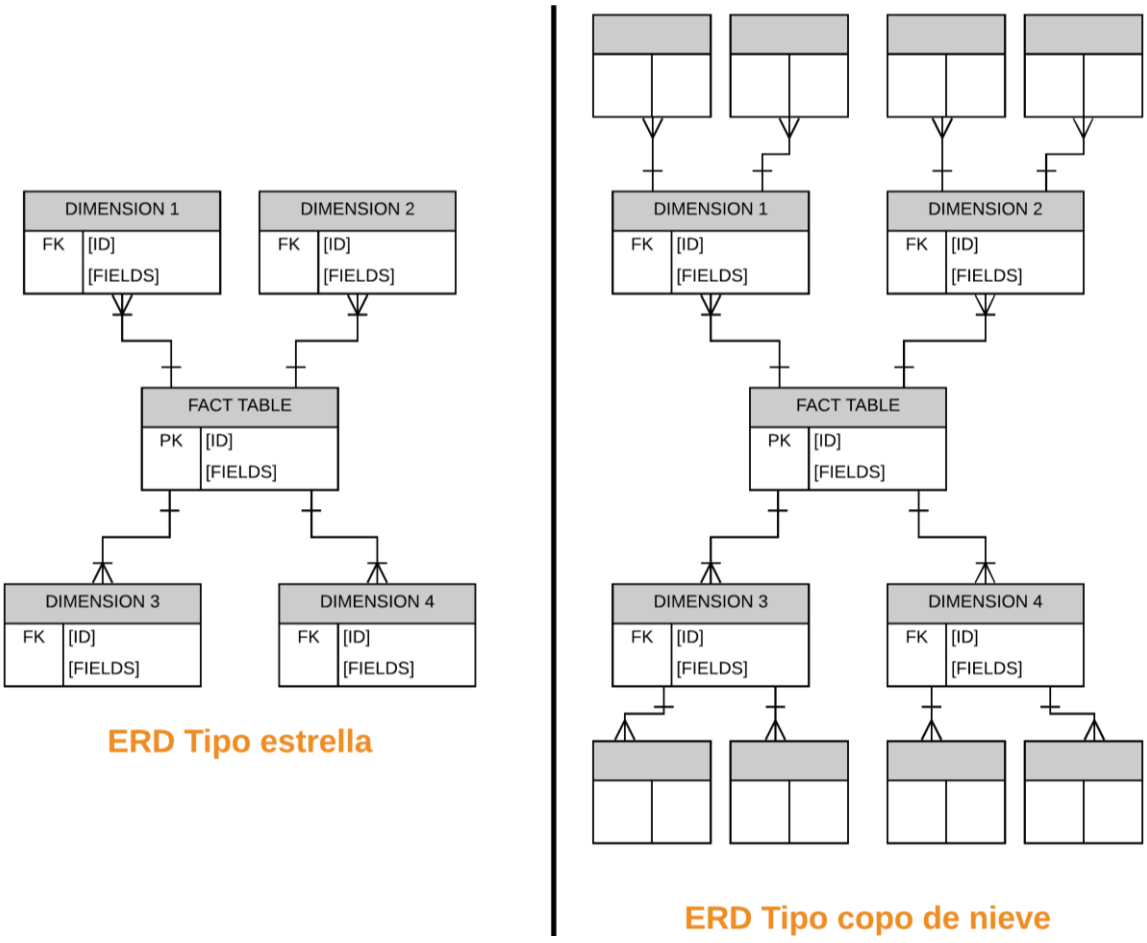


Figura 7. Tipos de ERD

Fuente: Elaboración propia

4.3 Manual Para la Aplicación de la Metodología

A continuación, se presenta un manual de usuario que detalla la forma de llevar a cabo la metodología junto con cada uno de los procedimientos seguidos para la correcta ejecución del proceso de creación del árbol de clasificación con el algoritmo C4.5 de AI, al igual que para el desarrollo de la herramienta web que se presenta para la manipulación e ingreso de información por parte de los usuarios.

4.3.1 Instalación de Paquetes de Trabajo

Para realizar eficientemente los macro procesos de captura, almacenamiento y procesamiento de información que se realizan entre RStudio y SQL Server es necesario utilizar los paquetes “foreign” y “RODBC” de R Project, que permiten descargar los documentos desde la página web del NHANES en formato del software STATA y trabajar con sentencias SQL dentro del entorno de RStudio respectivamente, facilitando el correcto funcionamiento del flujo de trabajo, tal como se muestra en la Figura 8.

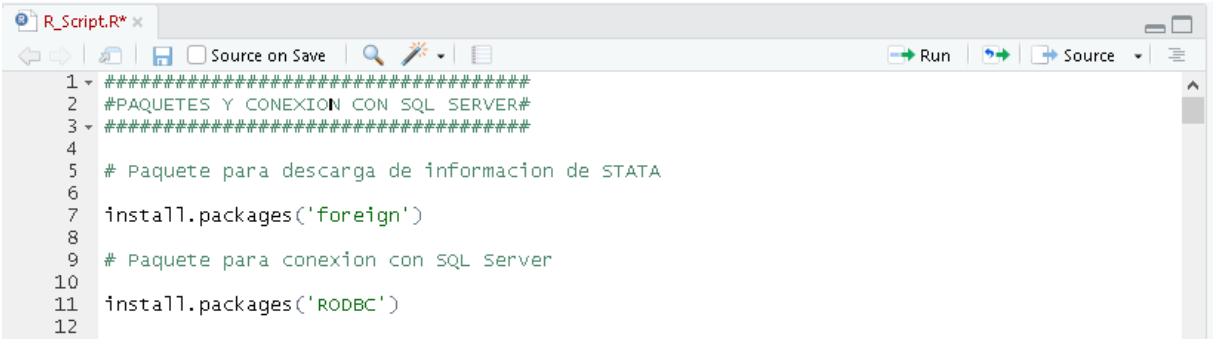


Figura 8. Instalación de paquetes de R Project

Fuente: Elaboración propia

4.3.2 Integración de R con SQL Server

Después de haber instalado los paquetes de trabajo, se realiza la integración de RStudio con SQL Server a través del uso del paquete RODBC, generando un driver de conexión que permita la fluida comunicación entre ambos entornos de trabajo, tal como se observa en la Figura 9.

```
13  library(RODBC)
14
15  myConex <- odbcDriverConnect(connection = "Driver={SQL Server Native Client 11.0};
16                               server=localhost; database=NHANES; trusted_connection=yes; ")
```

Figura 9. Integración de R con SQL Server

Fuente: Elaboración propia

4.3.3 Captura de Información con R Studio

El proceso de captura de información se diseña con la intención de tener un sistema de almacenamiento principal en SQL Server y un sistema de respaldo en ficheros planos (*.csv), con la intención de recuperar las tablas de trabajo en caso de que existieran fallos en el servidor local.

Para explicar el proceso de captura de información, tomaremos como ejemplo la descarga y almacenamiento de la tabla DEMO_I.xpt desde la página web del NHANES[10], tal como se muestra en la Figura 10.

```
17
18  ##########################
19  #TABLAS - ANOS 2015-2016#
20  ##########################
21
22  # Descarga DEMO_I_2015_2016
23
24  library(foreign)
25
26  download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DEMO_I.XPT",
27                "DEMO_I_2015_2016.XPT",mode="wb")
28
29  myData <- read.xport("DEMO_I_2015_2016.XPT")
```

Figura 10. Captura de información con R Studio

Fuente: Elaboración propia

[10] https://wwwn.cdc.gov

Es importante mencionar que este proceso se aplica para todas las tablas, entidades, que se muestran en la Tabla 7, correspondientes a todos los años de análisis, junto con su descripción.

Tabla 7. Entidades NHANES (Elaboración propia)

Tablas NHANES				
Nombre	**2011-2012 (G)**	**2013-2014 (H)**	**2015-2016 (I)**	**Descripción**
BIOPRO	x	x	x	Perfil bioquímico
BMX	x	x	x	Medidas corporales
DEMO	x	x	x	Variables demográficas
GHB	x	x	x	Hemoglobina glucosilada - A1C
GLU	x	x	x	Glucosa en ayunas
INS		x	x	Insulina
MCQ	x	x	x	Cuestionario de salud general
OGTT	x	x	x	Prueba de tolerancia a la glucosa
PAQ	x	x	x	Actividad física
SLQ	x	x	x	Desórdenes del sueño
TRIGLY	x	x		Colesterol LDL y triglicéridos

4.3.4 Almacenamiento de la Información en SQL Server

Una vez que la data ha sido almacenada en RStudio en la variable myData, procedemos a enviar esta información a nuestro sistema de almacenamiento en un servidor local en SQL Server y al sistema de respaldo redundante, plano, que se encuentra en un directorio local en la dirección "C:/data/NHANES/", tal como puede observarse en la Figura 11.

```
31 write.csv(myData, file = "C:/data/NHANES/DEMO_I_2015_2016.csv")
32
33 library(RODBC)
34
35 sqlSave(myConex,myData,tablename = "DEMO_I_2015_2016")
36
```

Figura 11. Almacenamiento de la información en SQL Server

Fuente: Elaboración propia

Adicionalmente, se realiza un ajuste a la estructura de las tablas almacenadas en SQL Server con la intención de diferenciar la información por cada uno de los años que se desean analizar. Esta modificación puede observarse en la Figura 12.

```
sqlQuery(myConex,"
         USE [NHANES]

         ALTER TABLE [dbo].[DEMO_I_2015_2016]
         ADD Years varchar(9)
         ")

sqlQuery(myConex,"
         USE [NHANES]

         UPDATE [dbo].[DEMO_I_2015_2016]
         SET Years='2015-2016'
         ")

sqlQuery(myConex,"
         USE [NHANES]

         ALTER TABLE [dbo].[DEMO_I_2015_2016]
         ADD ID Varchar(20)
         ")

sqlQuery(myConex,"
         USE [NHANES]

         UPDATE [dbo].[DEMO_I_2015_2016]
         SET ID=Years+'-'+CONVERT(VARCHAR,SEQN)
         ")
```

Figura 12. Ajuste de estructura de tablas en SQL Server

Fuente: Elaboración propia

4.3.5 Generación del Modelo Entidad-Relación

Para poder generar el Modelo Entidad-Relación, es indispensable garantizar que todas las tablas se encuentren almacenadas en la base de datos NHANES del servidor local. Una vez realizada la verificación del almacenamiento de todas las tablas requeridas, se diseñan vistas que integren dichas tablas a través del identificador único ID, que cumple las funciones de clave primaria (PK), tal como se observa en la Figura 13 y con mayor detalle en el Anexo C.

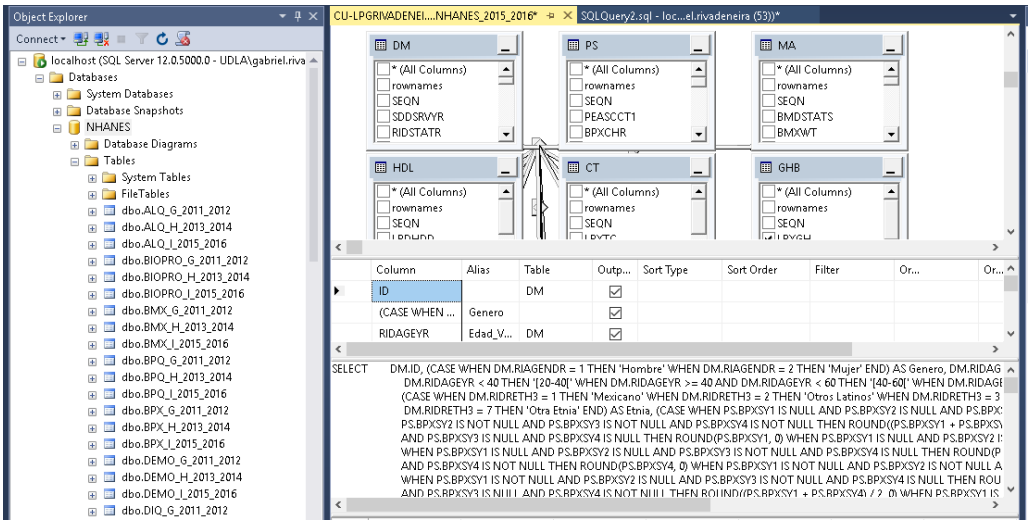


Figura 13. Diseño de vistas en SQL

Fuente: Elaboración propia

Para este proceso en particular se diseñará un modelo tipo estrella, como se observa en la Figura 14, para los años 2015-2016. Este proceso debe replicarse para todos los años de análisis.

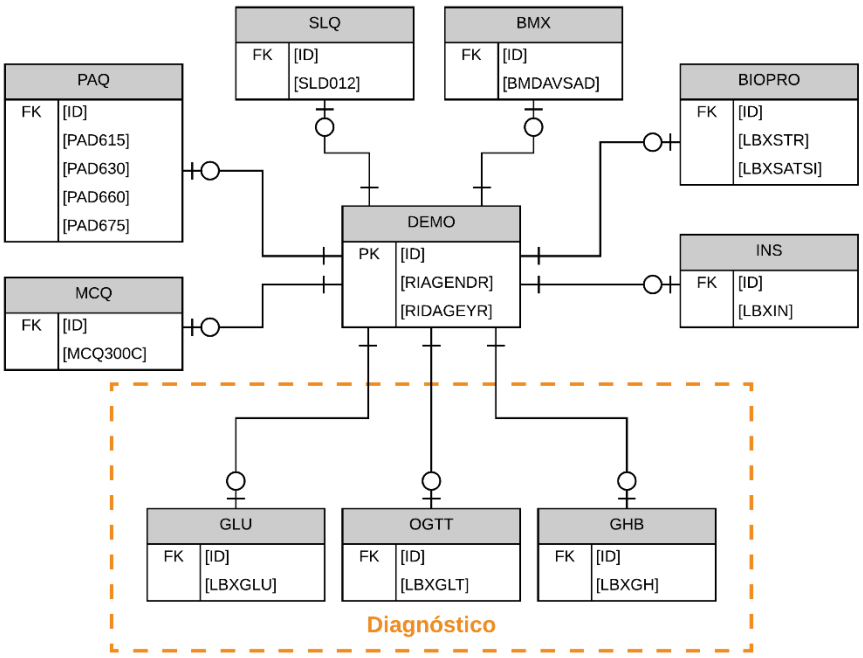


Figura 14. Modelo Entidad-Relación del proceso investigativo

Fuente: Elaboración propia

Es importante aclarar que una vez integradas las tablas, solo se deben utilizar las que no pertenecen al grupo de criterios de diagnóstico para evitar redundancia con el proceso de clasificación implementado en Weka.

4.3.6 Limpieza, Procesamiento y Codificación de la Información con SQL Server

Es primordial mencionar que el proceso de limpieza de la información almacenada en SQL Server se realizará sobre las vistas diseñadas para todos los años, evitando la manipulación de la información contenida en las tablas y así garantizar la integridad de la data descargada desde la fuente.

En el proceso de limpieza se buscará solventar problemas de completitud, precisión, consistencia e interpretabilidad de la data, bajo criterios de codificación específicos que alimentarán posteriormente el algoritmo de clasificación C4.5, implementado bajo el nombre J48 en Weka. Este proceso de filtrado y codificación, puede observarse en la Figura 15.

```
USE NHANES

SELECT
      [ID]
     ,[Genero]
     ,[Edad_Valor]
     ,[Edad]
     ,[DiamSagitalAbd_Valor]
     ,[Insulina_Valor]
     ,[Trigliceridos_Valor]
     ,[HistoriaFamiliar]
     ,[ActividadFisica]
     ,[HorasSueno_Valor]
     ,[AlanineALT_Valor]
     ,[HOMAIr_Valor]
     ,[Diagnostico]

FROM [dbo].[Vw_NHANES_2011_2016_UNION]

WHERE
[Edad_Valor]>=20 AND
[Edad_Valor]<=65 AND
[Genero] IS NOT NULL AND
[Etnia] IS NOT NULL AND
[DiamSagitalAbd_Valor] IS NOT NULL AND
[Insulina_Valor] IS NOT NULL AND
[Trigliceridos_Valor] IS NOT NULL AND
[HistoriaFamiliar] IS NOT NULL AND
[ActividadFisica] IS NOT NULL AND
[HorasSueno_Valor] IS NOT NULL AND
[HorasSueno_Valor]<=12 AND
[AlanineALT_Valor] IS NOT NULL AND
[HOMAIr_Valor] IS NOT NULL AND
[Diagnostico] IN ('Normal','Prediabetes')
```

Figura 15. Filtrado y codificación de la información

Fuente: Elaboración propia

Los procesos de unión de información tabular anual, que será analizada por el algoritmo de clasificación implementado en Weka, pueden observarse en el Anexo D, mientras que los procesos de limpieza y filtrado plurianual se pueden observar en el Anexo E.

4.3.7 Integración de Weka 3.8 con SQL Server 2016

Para generar la integración entre Weka con SQL Server, debe crearse una variable de entorno que permita utilizar un driver JDBC que habilite la comunicación entre ambos sistemas. Este proceso puede observarse en las figuras subsecuentes.

En principio, debe descargarse e instalarse un driver JDBC. En este caso en particular, se utilizó el "Microsoft JDBC Driver for SQL Server 6.0", disponible en la página de Microsoft. El proceso a seguir para la creación de la variable de entorno CLASSPATH, requerida para la integración entre Weka y SQL Server se muestra en la Figura 16.

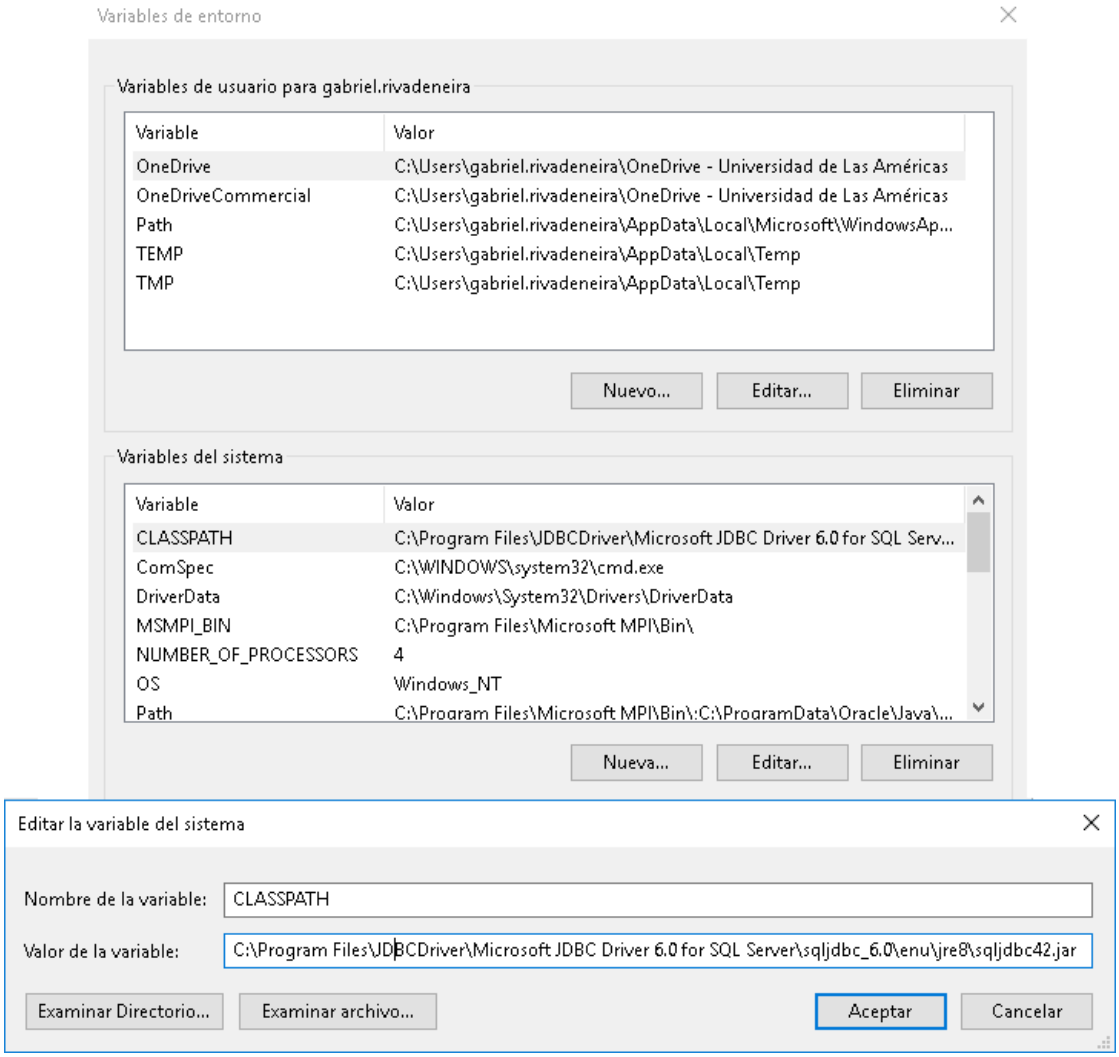


Figura 16. Creación de la variable de entorno CLASSPATH

Fuente: Elaboración propia

Posteriormente, se realiza una modificación sobre el archivo Weka.jar, ubicado en la carpeta de instalación de Weka. Al ingresar al contenido del archivo anteriormente mencionado, se realiza un cambio sobre el archivo de conexión DatabaseUtils.props, ubicado en la dirección \\weka\experiments. Las modificaciones generadas para la configuración del documento de configuración, se muestran en la Figura 17.

```
# JDBC driver (comma-separated list)
jdbcDriver=com.microsoft.sqlserver.jdbc.SQLServerDriver

# database URL
jdbc:sqlserver://localhost:1433;databasename=NHANES

# specific data types
# string, getString() = 0;    --> nominal
# boolean, getBoolean() = 1;  --> nominal
# double, getDouble() = 2;    --> numeric
# byte, getByte() = 3;        --> numeric
# short, getByte()= 4;        --> numeric
# int, getInteger() = 5;      --> numeric
# long, getLong() = 6;        --> numeric
# float, getFloat() = 7;      --> numeric
# date, getDate() = 8;        --> date
# text, getString() = 9;      --> string
# time, getTime() = 10;       --> date
# timestamp, getTime() = 11;  --> date

# map SQL Server data type to WEKA data type
# default mappings
varchar=0
float=2
tinyint=3
int=5
# values added manually
string=0
bigint=6
nvarchar=9
decimal=2
bit=1
```

Figura 17. Configuración del archivo de conexión de Weka

Fuente: Elaboración propia

Una vez realizadas las modificaciones descritas sobre el archivo DatabaseUtils.props, Weka queda configurado para conectarse únicamente con el servidor local y la base de datos NHANES. Este proceso puede replicarse con otros servidores, cambiando los parámetros de la conexión a través del driver JDBC.

4.3.8 Consulta SQL Desde Weka

Con el proceso de integración completo, Weka puede conectarse con SQL Server y operar con las vistas y tablas existentes en la base de datos NHANES. En este caso, el proceso de conexión con SQL Server, se realiza específicamente con la vista Vw_NHANES_2011_2016WEKA_J48, que se encuentra previamente configurada para trabajar con Weka.

Para conectarse con la base de datos NHANES de SQL Server, se ingresa en la sección Open DB de Weka y se realiza una consulta SQL a la tabla o vista deseada, como se observa en la Figura 18.

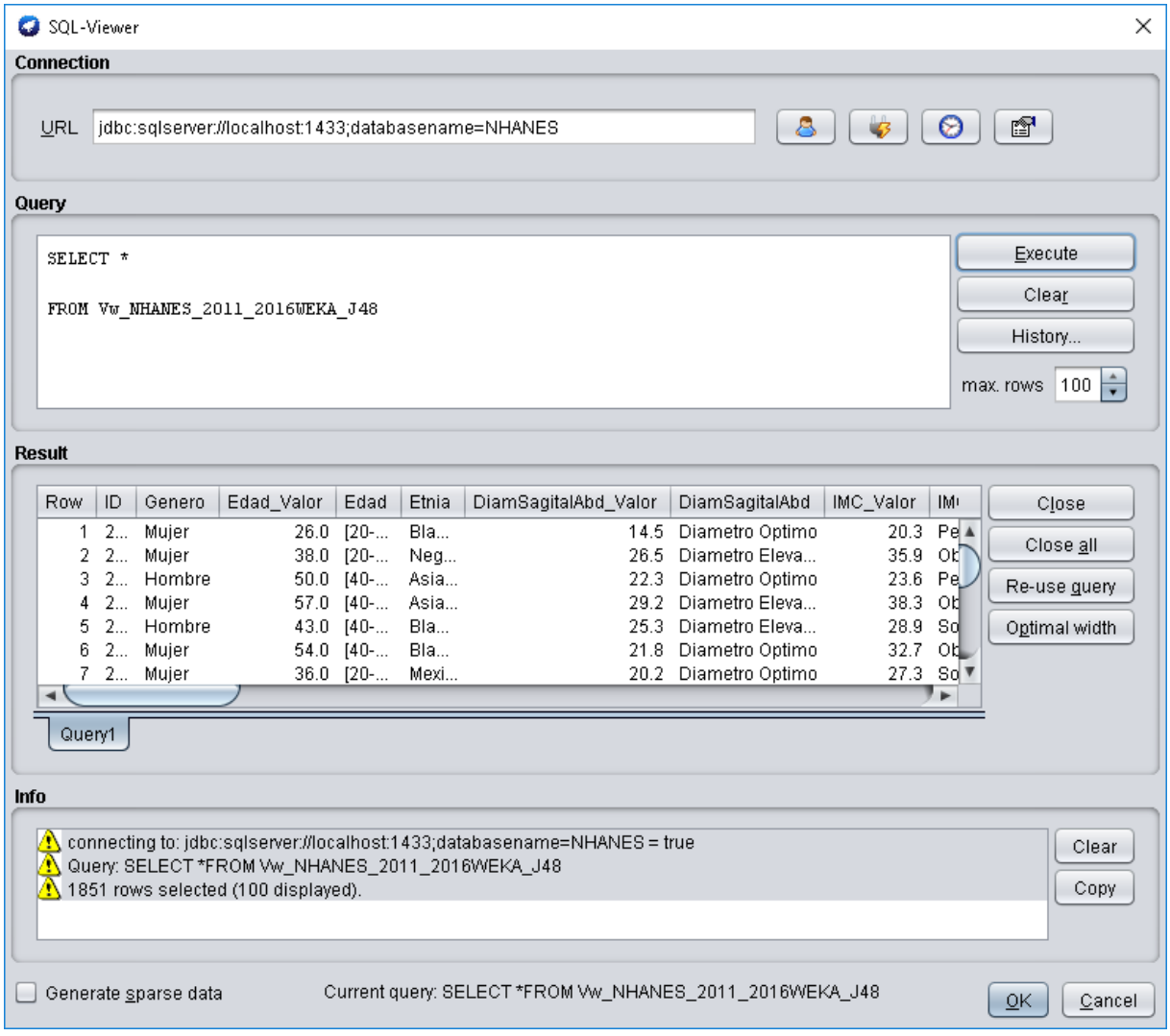


Figura 18. Consulta SQL desde Weka

Fuente: Elaboración propia

4.3.9 Diseño del Árbol de Clasificación con el Algoritmo C4.5 en Weka

Una vez ejecutada las consulta SQL desde Weka, se pueden acceder a los algoritmos de AI implementados en la aplicación. El algoritmo C4.5 se encuentra en la pestaña Classify, seleccionando el algoritmo J48, que es la implementación de C4.5 realizada por Weka en Java.

Posteriormente, se selecciona la configuración deseada para la ejecución del algoritmo.C4.5, buscando el mejor equilibrio entre la clasificación con datos de entrenamiento y con datos de prueba, evitando así el sobre entrenamiento. En términos generales, se buscó que el algoritmo genere una poda post entrenamiento del árbol, que no cree clases con menos de 10 objetos y que maneje un factor de confianza de 0.01, como se puede evidenciar en la Figura 19.

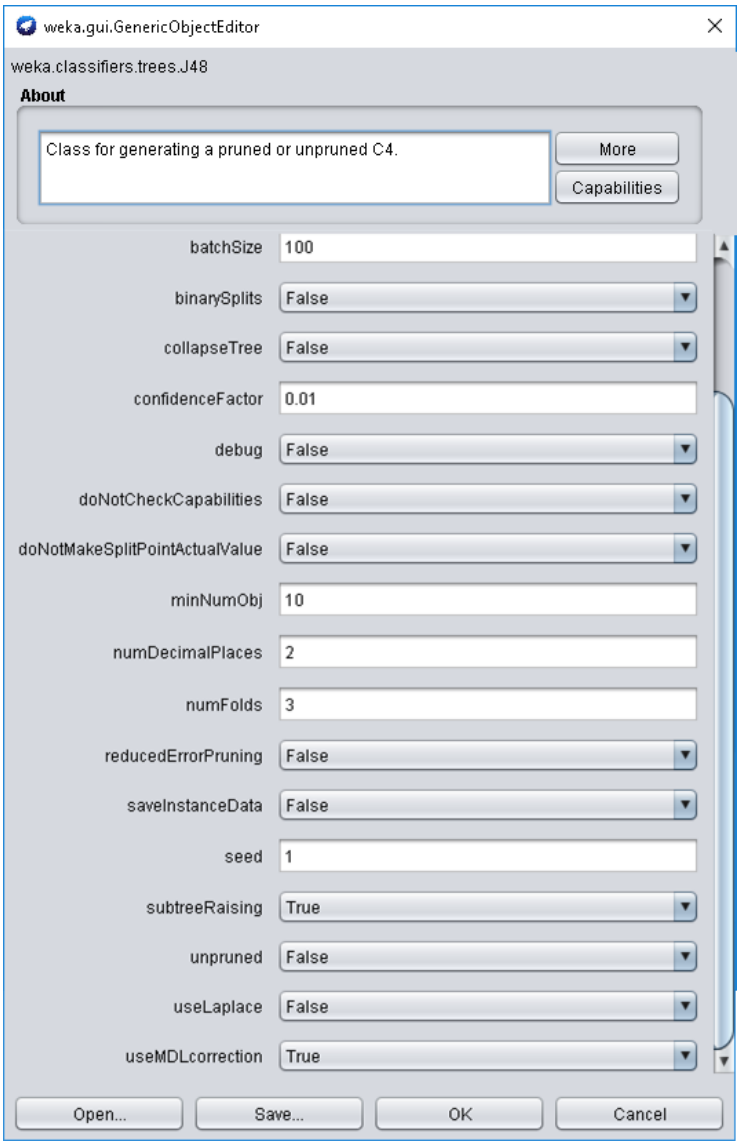
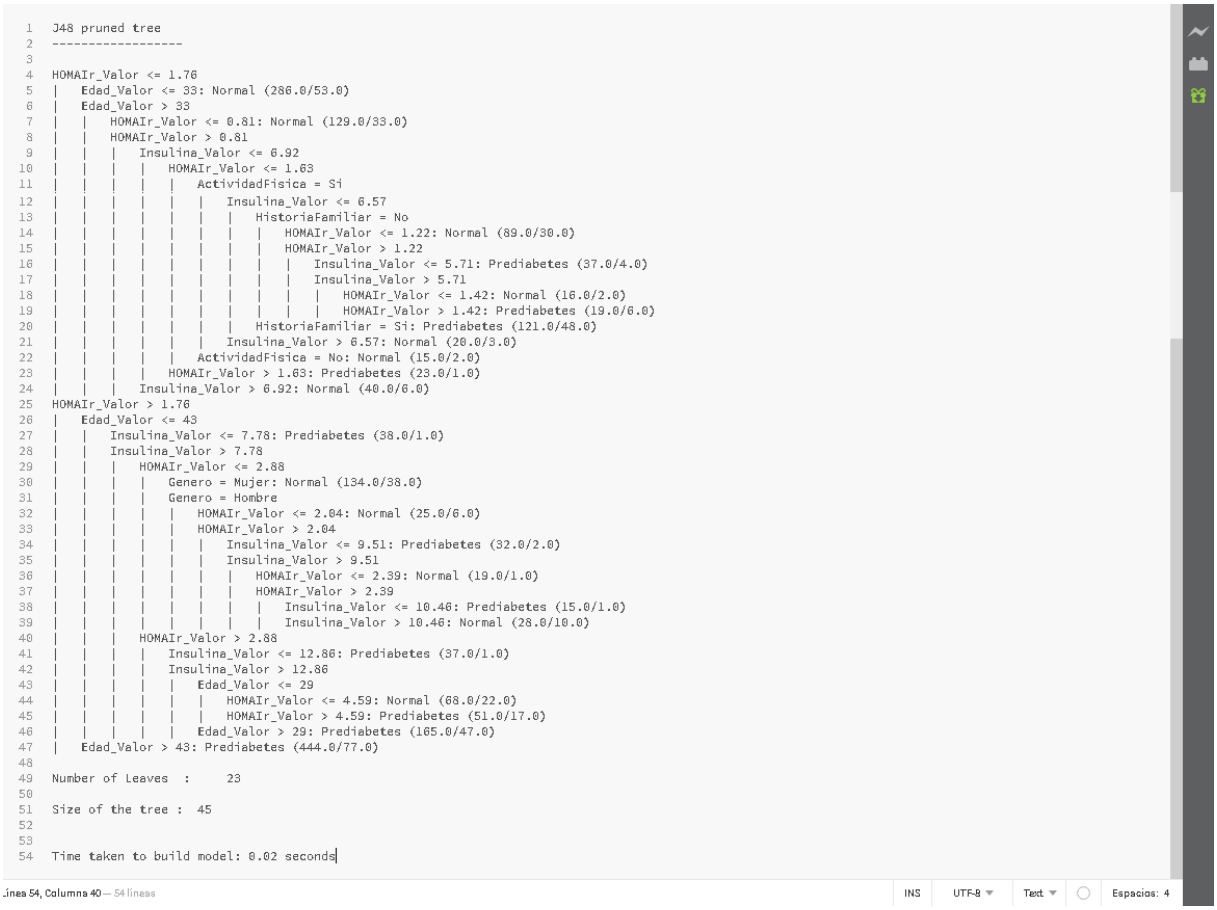


Figura 19. Configuración del algoritmo C4.5 (J48) en Weka

Fuente: Elaboración propia

Al concluir la parametrización del algoritmo se procede a la ejecución del mismo, obteniendo un árbol de clasificación que puede ser desplegado tanto en forma de texto, como en una interfaz gráfica. La Figura 20 muestra la caracterización del gráfico a manera de texto, mientras que, en el Anexo F puede observarse la representación del árbol en modo gráfico.



```
J48 pruned tree
------------------

HOMAIr_Valor <= 1.76
|   Edad_Valor <= 33: Normal (286.0/53.0)
|   Edad_Valor > 33
|   |   HOMAIr_Valor <= 0.81: Normal (129.0/33.0)
|   |   HOMAIr_Valor > 0.81
|   |   |   Insulina_Valor <= 6.92
|   |   |   |   HOMAIr_Valor <= 1.63
|   |   |   |   |   ActividadFisica = Si
|   |   |   |   |   |   Insulina_Valor <= 6.57
|   |   |   |   |   |   |   HistoriaFamiliar = No
|   |   |   |   |   |   |   |   HOMAIr_Valor <= 1.22: Normal (89.0/30.0)
|   |   |   |   |   |   |   |   HOMAIr_Valor > 1.22
|   |   |   |   |   |   |   |   |   Insulina_Valor <= 5.71: Prediabetes (37.0/4.0)
|   |   |   |   |   |   |   |   |   Insulina_Valor > 5.71
|   |   |   |   |   |   |   |   |   |   HOMAIr_Valor <= 1.42: Normal (16.0/2.0)
|   |   |   |   |   |   |   |   |   |   HOMAIr_Valor > 1.42: Prediabetes (19.0/6.0)
|   |   |   |   |   |   |   |   HistoriaFamiliar = Si: Prediabetes (121.0/48.0)
|   |   |   |   |   |   |   Insulina_Valor > 6.57: Normal (20.0/3.0)
|   |   |   |   |   ActividadFisica = No: Normal (15.0/2.0)
|   |   |   |   HOMAIr_Valor > 1.63: Prediabetes (23.0/1.0)
|   |   |   Insulina_Valor > 6.92: Normal (40.0/6.0)
HOMAIr_Valor > 1.76
|   Edad_Valor <= 43
|   |   Insulina_Valor <= 7.78: Prediabetes (38.0/1.0)
|   |   Insulina_Valor > 7.78
|   |   |   HOMAIr_Valor <= 2.88
|   |   |   |   Genero = Mujer: Normal (134.0/38.0)
|   |   |   |   Genero = Hombre
|   |   |   |   |   HOMAIr_Valor <= 2.04: Normal (25.0/6.0)
|   |   |   |   |   HOMAIr_Valor > 2.04
|   |   |   |   |   |   Insulina_Valor <= 9.51: Prediabetes (32.0/2.0)
|   |   |   |   |   |   Insulina_Valor > 9.51
|   |   |   |   |   |   |   HOMAIr_Valor <= 2.39: Normal (19.0/1.0)
|   |   |   |   |   |   |   HOMAIr_Valor > 2.39
|   |   |   |   |   |   |   |   Insulina_Valor <= 10.46: Prediabetes (15.0/1.0)
|   |   |   |   |   |   |   |   Insulina_Valor > 10.46: Normal (28.0/10.0)
|   |   |   HOMAIr_Valor > 2.88
|   |   |   |   Insulina_Valor <= 12.86: Prediabetes (37.0/1.0)
|   |   |   |   Insulina_Valor > 12.86
|   |   |   |   |   Edad_Valor <= 29
|   |   |   |   |   |   HOMAIr_Valor <= 4.59: Normal (68.0/22.0)
|   |   |   |   |   |   HOMAIr_Valor > 4.59: Prediabetes (51.0/17.0)
|   |   |   |   |   Edad_Valor > 29: Prediabetes (165.0/47.0)
|   Edad_Valor > 43: Prediabetes (444.0/77.0)

Number of Leaves  :     23

Size of the tree :  45


Time taken to build model: 0.02 seconds
```

Línea 54, Columna 40 — 54 líneas INS UTF-8 Text Espacios: 4

Figura 20. Árbol de clasificación en modo texto de C4.5 (J48)

Fuente: Elaboración propia

4.3.10 Salida del Algoritmo C4.5

Posterior a la ejecución del algoritmo, se presenta una salida con un resumen de la información cargada, junto con un análisis de la precisión del árbol generado, tanto con datos de entrenamiento como con datos de prueba en el proceso de validación cruzada. La salida del algoritmo, con el árbol generado presenta un 70.45% de instancias bien clasificadas, junto con métricas adicionales de precisión en la clasificación, como se muestra en la Figura 21.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -O -C 0.01 -M 10
Relation:     QueryResult-weka.filters.unsupervised.attribute.Remove-R1,4-5,7-13,15-19,21,25,27-29,31
Instances:    1851
Attributes:   11
              Genero
              Edad_Valor
              DiamSagitalAbd_Valor
              Insulina_Valor
              Trigliceridos_Valor
              HistoriaFamiliar
              ActividadFisica
              HorasSueno_Valor
              AlanineALT_Valor
              HOMAIr_Valor
              Diagnostico
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Number of Leaves  :     23

Size of the tree :  45


Time taken to build model: 0.02 seconds

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

Correctly Classified Instances        1304               70.4484 %
Incorrectly Classified Instances       547               29.5516 %
Kappa statistic                          0.4062
Mean absolute error                      0.3834
Root mean squared error                  0.453
Relative absolute error                 76.9793 %
Root relative squared error             90.7695 %
Total Number of Instances             1851

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.679    0.273    0.687      0.679    0.683      0.406    0.746     0.691     Normal
                 0.727    0.321    0.719      0.727    0.723      0.406    0.746     0.749     Prediabetes
Weighted Avg.    0.704    0.299    0.704      0.704    0.704      0.406    0.746     0.722

=== Confusion Matrix ===

   a   b   <-- classified as
 589 279 |   a = Normal
 268 715 |   b = Prediabetes

```

Figura 21. Salida del algoritmo C4.5 de Weka

Fuente: Elaboración propia

4.3.11 Visualización Interactiva en D3

Una vez terminado el proceso de ejecución del algoritmo C4.5 se obtiene el árbol de clasificación definitivo. Para poder poner este árbol a disposición del usuario es importante implementar una visualización interactiva del modelo en una herramienta web desarrollada en Brackets, para lo cual se utiliza la librería D3 de JavaScript.

La visualización desarrollada, muestra el árbol de clasificación de D3 junto con una serie de campos de texto para poder ingresar los valores de las variables utilizadas en el modelo de clasificación y recibir, tras la ejecución, la clasificación de la nueva instancia junto con su probabilidad de acierto. Esto puede observarse a breves rasgos en la Figura 22, al igual que su código HTML, CSS y JS en los Anexos G y H respectivamente.

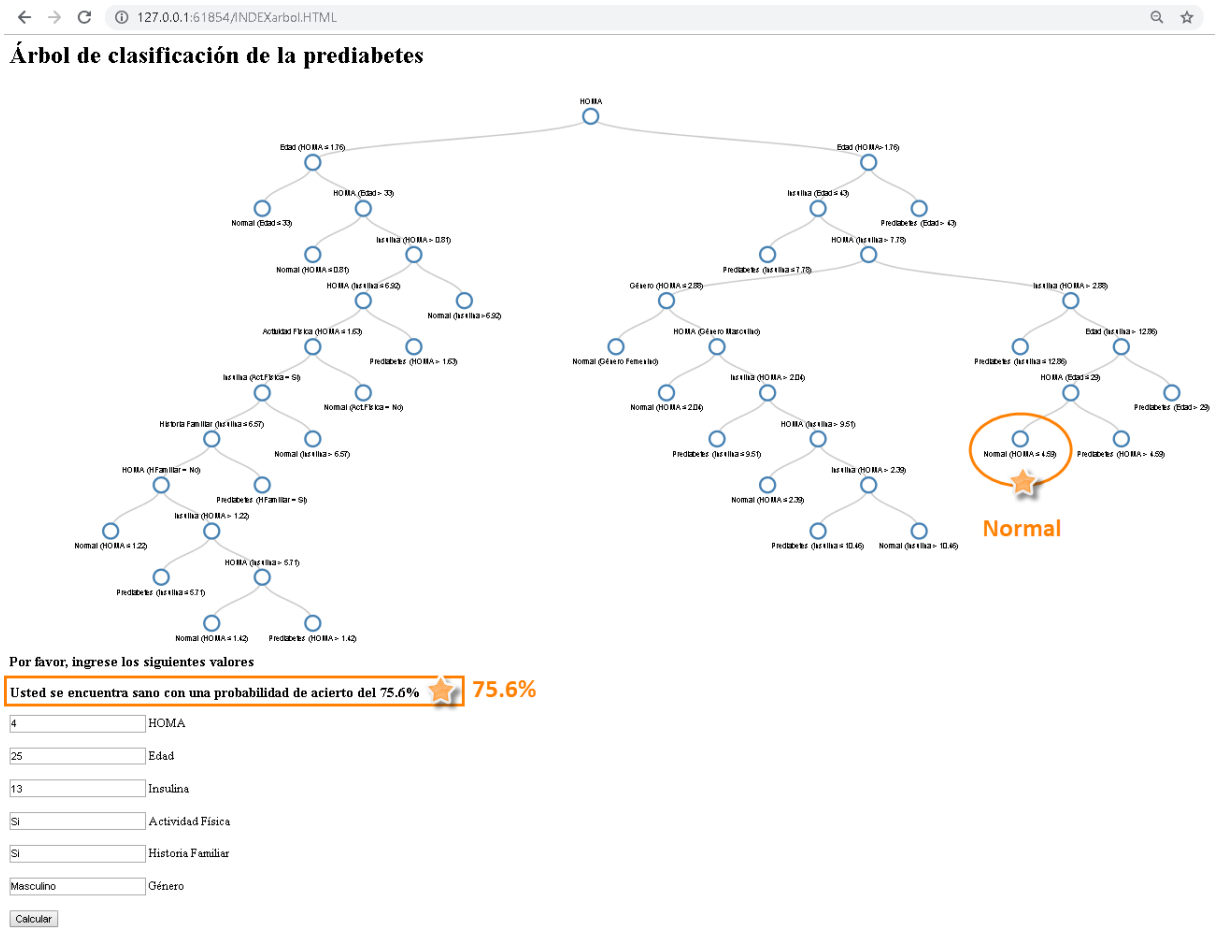


Figura 22. Visualización Interactiva utilizando D3 y HTML

Fuente: Elaboración propia

Una vez generado, en la aplicación web mediante el uso de la librería D3, el árbol inducido a través de la implementación del algoritmo C4.5 de IA en Weka, es posible ingresar nuevas instancias para su clasificación. Estas instancias corresponden a personas que cuentan con sus exámenes clínicos específicos y los resultados de sus variables observables.

Las posibles salidas de la aplicación web brindan al usuario información relevante en dos posibles escenarios. El primer escenario presenta un resultado favorable donde el usuario se encuentra sano y el segundo le informa que padece de PD, por lo que debe buscar asistencia médica preventiva y realizarse un chequeo general. En ambos casos se provee una probabilidad de acierto del proceso de clasificación.

Estos resultados satisfacen el segundo y tercer objetivo específico planteado para este proyecto de investigación, mismos que se muestran a continuación:

“Permitir a los pacientes que cuenten con exámenes clínicos definidos, ubicarse en un punto específico del modelo de clasificación de la prediabetes.”

“Brindar un criterio básico de diagnóstico que motive a la población a buscar asistencia de salud preventiva dependiendo del diagnóstico de prediabetes provisto y su probabilidad.”

Terminado el manual con el diseño de la aplicación web y operativamente funcionando, se cumple con el cuarto y último objetivo específico del presente proyecto de investigación, mismo que se presenta a continuación:

“Diseñar un manual detallado de cómo aplicar la metodología en otras poblaciones.”

Cubiertos todos los objetivos específicos, habiendo generado el árbol de clasificación mediante la implementación del algoritmo C4.5 de IA y habiendo evaluado nuevas instancias de forma ágil y eficiente, se cumple con el objetivo general de la investigación, mismo que se muestra a continuación:

“Generar un árbol de clasificación de prediabetes que permita, a través de variables fácilmente observables y exámenes clínicos definidos, ubicarse de manera rápida y sencilla en un nivel específico de diagnóstico.”

5. Evaluación de la Metodología Propuesta

En este capítulo se describirá, en términos generales, el experimento realizado para la ejecución del presente proyecto de investigación, junto con sus métricas de precisión y calidad.

El proyecto se inicia diseñando un Data Warehouse (DWH) que se almacena en un servidor local y que contiene todas las tablas requeridas para el desarrollo del estudio. Para cumplir con este requerimiento, se realiza un proceso de ETL que alimenta el repositorio de información con todas las entidades provenientes de la encuesta estadounidense NHANES. En este caso en particular se han seleccionado las encuestas bianuales correspondientes a los años 2011 hasta el año 2016, mismas que representan toda la data existente desde el año 2011 a la actualidad.

La extracción de información desde la página web de la encuesta NHANES se realiza mediante la utilización del software R Project, a través de su interfaz R Studio. Aquí se utilizan los paquetes “foreign” y “RODBC”, mismos que permiten descargar los archivos almacenados en formato del software STATA, convertirlos en dataframes de R e insertarlos en la base de datos NHANES generada en SQL Server.

Posterior a este proceso, se crean vistas que realizan una limpieza y codificación de las variables existentes en las tablas bianuales, sin tocar la data original almacenada en el DWH creado en el servidor local. En estas vistas se realizan operaciones de integración entre las tablas de dimensiones con la tabla de hechos para cada encuesta bianual, estructurando de esta forma un ERD con forma de estrella.

Una vez terminada la creación del ERD, se procede a agregar sobre las tablas de hechos una Clave Primaria (PK) que permita integrar todas las tablas bianuales en una sola vista que alimenta el algoritmo C4.5 de AI que induce la creación del árbol de clasificación requerido.

Concluida la generación de la vista plurianual de datos, se realiza un proceso de integración de SQL Server con Weka. Este proceso de integración permite que Weka lea de forma directa todas las tablas y vistas almacenadas en el DWH que se encuentra en el servidor local.

Al conectar Weka a la vista plurianual de información, se procede a seleccionar las variables a analizarse y la clase de contraste que se utilizará para alimentar el algoritmo de aprendizaje supervisado que induce el árbol de clasificación creado con el algoritmo C4.5.

Terminado el proceso de carga de información, selección de variables y clase de salida, se procede a parametrizar el algoritmo C4.5 para evitar problemas de sobre entrenamiento en el momento de la generación e inducción del árbol de clasificación.

Al terminar de parametrizar el algoritmo C4.5 se lo ejecuta y se obtiene como resultado un árbol de clasificación junto con los parámetros de calidad y precisión propios del proceso de clasificación.

Para la generación de este árbol de clasificación, se cargaron 1851 instancias con 11 atributos, correspondientes a los 10 factores de riesgo elegidos para el análisis y el diagnóstico de la patología, mismo que permite utilizar cualquier algoritmo de aprendizaje supervisado.

Adicionalmente, se utiliza un método de validación cruzada de 10 iteraciones, que permite dividir la data cargada en 10 grupos que se alternarán, de uno en uno, cumpliendo la función de data de data de prueba, permitiendo así extraer medidas de precisión confiables.

Como se puede observar en la Figura 23, se tienen 411 instancias mal clasificadas en el proceso de entrenamiento y 547 en el proceso de validación cruzada, brindándole al árbol una cualidad optimista con un nivel de predicción del 77.8% y del 70.45% de aciertos, respectivamente.

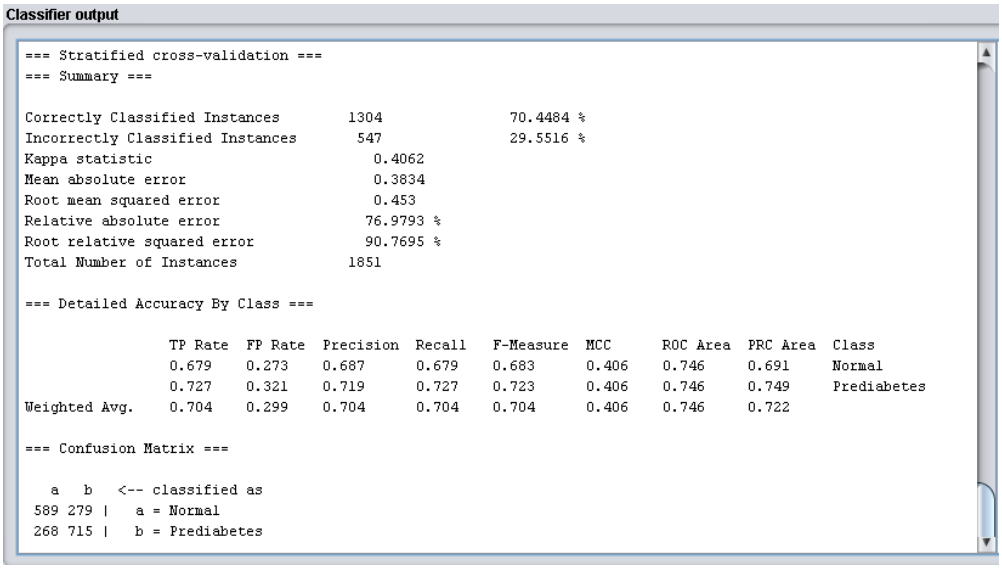
La sensibilidad del árbol de clasificación generado, TPrate, es del 72.7%. Para lograr obtener niveles de sensibilidad y precisión elevados, evitando el sobreajuste debido al entrenamiento, se eligió un proceso de poda posterior a la creación del árbol junto con una cantidad mínima de instancias por hoja de 10 elementos.

Con estos insumos se procede a generar un archivo de extensión *.json que permite crear un documento con cualidades jerárquicas (padre-hijo), mismo que satisface completamente la estructura de un árbol de clasificación. Con el archivo jerárquico generado, es posible diseñar una aplicación web que grafique con grandes cualidades estéticas el árbol de clasificación, mediante la utilización de la librería D3 de JavaScript.

Una vez concluida con la graficación del árbol de clasificación, se crean puntos de inserción de información (input box) que permitan al usuario ingresar su información tanto de variables observables como de los resultados de sus exámenes clínicos.

Adicionalmente, se agrega un botón de cálculo que devuelve al usuario una explicación de su ubicación en la estructura del árbol, su resultado en el proceso de clasificación y la probabilidad de acierto.

Para la evaluación de la metodología se analizan las medidas de precisión y calidad del árbol de clasificación generado con el algoritmo C4.5, mismos que se encuentran en la salida de Weka que se observa en la Figura 23



Classifier output

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

Correctly Classified Instances        1304               70.4484 %
Incorrectly Classified Instances       547               29.5516 %
Kappa statistic                          0.4062
Mean absolute error                      0.3834
Root mean squared error                  0.453
Relative absolute error                 76.9793 %
Root relative squared error             90.7695 %
Total Number of Instances             1851

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.679    0.273    0.687      0.679    0.683      0.406    0.746     0.691     Normal
                 0.727    0.321    0.719      0.727    0.723      0.406    0.746     0.749     Prediabetes
Weighted Avg.    0.704    0.299    0.704      0.704    0.704      0.406    0.746     0.722

=== Confusion Matrix ===

   a   b   <-- classified as
 589 279 |   a = Normal
 268 715 |   b = Prediabetes
```

Figura 23. Medidas de calidad y precisión de la clasificación

Fuente: Elaboración propia

6. Conclusiones y Trabajo Futuro

A continuación, se define la relevancia y alcance de la contribución. De igual manera se cuentan las conclusiones, producto de la experiencia adquirida durante el desarrollo investigativo y las recomendaciones descritas para evitar que futuros lectores incurran en errores y malas prácticas por las que se transitó a lo largo de la evolución de este proyecto.

Adicionalmente, se delinea una proyección de trabajo futuro para la implementación en la práctica de la metodología de investigación desarrollada con la ayuda de el algoritmo de clasificación y aprendizaje supervisado C4.5 de AI.

6.1 Relevancia y Alcance de la Contribución

El presente proyecto de investigación nace con la finalidad de prevenir el riesgo de padecimiento de prediabetes, reduciendo así la posibilidad de aparición de la diabetes mellitus tipo 2 que es una enfermedad crónica y degenerativa que surge cuando el páncreas no logra producir suficiente insulina o cuando el organismo no utiliza eficazmente la insulina que produce.

Para cumplir con este objetivo, se propone el desarrollo de un modelo de clasificación de las condiciones clínicas que definen su comportamiento, debido a que una intervención temprana y oportuna puede disminuir aproximadamente en un 58% la progresión de esta condición hacia la diabetes mellitus tipo 2.

Dentro del desarrollo del proyecto, se diseña un árbol de clasificación de prediabetes, que permite al usuario ubicarse en un punto específico de diagnóstico con los resultados de sus exámenes clínicos específicos y variables fácilmente observables.

Es posible acceder al árbol de clasificación a través de una aplicación web, brindando al usuario un criterio básico de diagnóstico que lo motive a buscar asistencia de salud preventiva dependiendo del diagnóstico de prediabetes provisto y su probabilidad.

6.2 Conclusiones

A continuación, se presentan las conclusiones obtenidas al finalizar este proceso investigativo:

- Es posible generar un árbol de clasificación de prediabetes que permita a los usuarios, a través de variables fácilmente observables y exámenes clínicos definidos, ubicarse de manera rápida y sencilla en un nivel específico de diagnóstico con una probabilidad de acierto.

- Es perfectamente viable el identificar las condiciones clínicas, no convencionales, que inciden en el diagnóstico de la prediabetes, realizando una investigación profunda y consciente de la literatura clínica-científica existente.

- Es posible realizar una aplicación web amigable que permita a los usuarios que cuenten con exámenes clínicos definidos y mediciones fácilmente observables, ubicarse en un punto específico del modelo de clasificación de la prediabetes.

- Es factible brindar un criterio básico de diagnóstico, basado en data histórica recolectada, que motive a la población a buscar asistencia de salud preventiva dependiendo del diagnóstico de prediabetes provisto y su probabilidad.

- Es posible realizar un estudio clasificatorio de los factores de riesgo que inciden en la prediabetes a partir de datos históricos levantados de forma trasversal.

- Con la ayuda de un buen diseño de experimentos, es viable el implementar, en otras poblaciones, una encuesta sobre la cual se pueda aplicar la metodología de clasificación desarrollada para el estudio clasificatorio de la prediabetes.

- Es importante identificar las variables que puedan levantarse tanto por simple inspección, por exámenes de laboratorio como por derivación de otras variables simples antes de iniciar el proceso diseño de una metodología que implemente un algoritmo de AI.

- Para evitar inconvenientes debidos a la manipulación humana de data provista por una fuente oficial, es necesario manejar sistemas integrados de información que permitan procesos de ETL (Extracción, transformación y carga) fluidos y eficientes.

- Independientemente del alto nivel de integración y automatización que se pueda tener en los sistemas de información, siempre se necesita un ser humano en algún punto del proceso que defina criterios importantes de operación y análisis de información.

- Una vez concluidos los procesos de diseño y calibración de metodologías en las que se aplican algoritmos de AI, es importante someter esos resultados a procesos de validación con datos de prueba, que permitan identificar conflictos e inconsistencias debido al sobreajuste de los modelos.

- Posterior a la validación de las metodologías que integran algoritmos de AI en temas relacionados con salud, es crucial someter los resultados a procesos de diseño de experiencia de usuario, que permitan generar herramientas sencillas y funcionales que sean amigables con todos los potenciales usuarios.

6.3 Recomendaciones

- Si se desea implementar un algoritmo de aprendizaje supervisado, es importante validar si existe el diagnóstico requerido o las variables que se utilizan para el diagnóstico, ya que sin esta información los algoritmos mencionados son inútiles.

- Antes de iniciar un proceso investigativo que involucre algoritmos de AI, es importante realizar una revisión preliminar de literatura que ayude a delimitar los alcances y la viabilidad del proyecto.

- Independientemente de cuan confiable pueda ser la fuente de información, es indispensable realizar un análisis descriptivo de la data, al igual que una revisión de parámetros de calidad en términos de precisión, consistencia e interpretabilidad.

- En todo proceso investigativo es importante contar con un experto en la materia que pueda evitar el incurrir en errores elementales, propios del área investigada.

- Es importante evaluar el costo de utilizar una distribución de software comercialmente integrado contra el costo de buscar generar integraciones entre programas de nuestro dominio o software libre.

- Es importante incluir en el cronograma de trabajo una etapa de diseño enfocado en la experiencia del usuario, ya que esto facilita los procesos de difusión y comercialización de la aplicación resultante de la implementación del modelo.

- Es fundamental elegir las variables de trabajo con un sustento teórico fuerte, evitando ambigüedades que puedan desencadenar en implicaciones discriminatorias.

- Debemos conocer una amplia gama de modelos que se ajusten a la naturaleza de los datos, eligiendo el que mejor facilite el cumplimiento de los objetivos.

- Una vez ejecutado el algoritmo de AI, es importante calibrar su funcionamiento, buscando la mejor configuración de sus parámetros.

6.4 Trabajo Futuro

Para el desarrollo y profundización de una investigación completa sobre prediabetes, es importante el poder integrar los elementos previamente diseñados, mejorarlos y juntarlos con herramientas innovadoras de AI.

En este caso en particular, las intenciones de trabajo futuro se encuentran enfocadas en el desarrollo de un programa de salud pública en Ecuador, que permita integrar herramientas no invasivas como el “Diabetes Risk Calculator” con el diseño de un índice único de prediabetes, una regresión logística sobre los factores de riesgo de padecer prediabetes, un análisis de series de tiempo, el árbol de clasificación, objeto de este proyecto de investigación e investigaciones que aporten de manera significativa con recomendaciones nutricionales para personas con PD y DM2.

Con la ayuda de este paquete de herramientas, es posible generar una intervención profunda en temas de PD y DM2 a distintos niveles de la población nacional; en ciertos casos con métodos no invasivos y en otros suplementando las observaciones con exámenes clínicos.

El objetivo del desarrollo de este plan de salud pública es el de revertir la tendencia alarmante de crecimiento de la cantidad de personas que padecen de PD como de DM2, brindando oportunidades de empleo a personas capacitadas e interesadas en este tema tan importante en el área de la salud.

Referencias

(ADA, 2015) American Diabetes Association. (2015). El diagnóstico de la diabetes e información sobre la prediabetes. Recuperado el 10 de diciembre de 2018 de http://www.diabetes.org/es/informacion-basica-de-la-diabetes/diagnstico.html.

(ADA, 2017) American Diabetes Association. (2017). Classification and Diagnosis of Diabetes. Diabetes Care, (40)1. Recuperado el 01 de noviembre de 2018 de http://care.diabetesjournals.org/content/40/Supplement_1/S11. DOI:10.2337/dc17-S005

(Cuperlovic-Culf, 2018) Cuperlovic-Culf, M. (2018). Machine Learning Methods for Analysis of Metabolic Data and Metabolic Pathway Modeling. Metabolities, (08)4, 1-12. Recuperado el 02 de noviembre de 2018 de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5875994/. DOI:10.3390/metabo8010004

(Date, 2001) Date, C. (2001). Introducción a los sistemas de bases de datos. Pearson Education, (07)1. ISBN:968-444-419-2

(Edureka, 2018) Atul. (2018). AI vs Machine Learning vs Deep Learning. Recuperado el 18 de diciembre de 2018 de https://www.edureka.co/blog/ai-vs-machine-learning-vs-deep-learning/.

(FID, 2017) Federación Internacional de la Diabetes FID. (2017). Atlas de la diabetes de la FID. Panorama mundial, (08)1, 41. Recuperado el 01 de noviembre de 2018 de http://www.diabetesatlas.org/IDF_Diabetes_Atlas_8e_interactive_ES.

(Friege, 2014) Friege, F., Lara-Esqueda, A., Suverza A., Campuzano, R., Vanegas, E., Vidrio, M., Cañete, F., Hernández-Yero, A., Zúñiga-González, S., Romero, A., Gruber, E., Zúñiga-Guajardo, S., Lyra, R., Islas, S., García, R., Lara-Esqueda, A., Sampaio, R., González-Chávez, A., Vélez, J., Hernández, L. (2014). Consenso de Prediabetes. Documento de posición de la Asociación Latinoamericana de Diabetes (ALAD). Revista de la ALAD, (17)4, 149-150. Recuperado el 03 de noviembre de 2018 de http://www.revistaalad.com/files/0904_ConsPred.pdf.

(Grimmett, 2017) Grimmett, C. (2017). HTML and CSS Basics for WordPress. Recuperado el 17 de diciembre de 2018 de http://www.cagrimmett.com/development/2017/04/24/tools-for-learning-css.html.

(Heikes, 2008) Heikes, K., Eddy, D., Arondekar, B. & Schlessinger L. (2008). Diabetes Risk Calculator - A simple tool for detecting undiagnosed diabetes and pre-diabetes. Diabetes Care, (31)5. Recuperado el 03 de noviembre de 2018 de http://care.diabetesjournals.org/content/31/5/1040#aff-1. DOI:10.2337/dc07-1150

(Karimi-Alavijeh, 2016) Karimi-Alavijeh, F., Jalili, S. & Sadeghi, M. (2016). Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA Atheroscler, (08)3. Recuperado el 02 de noviembre de 2018 de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5055373/. PMCID:PMC5055373

(Ko, 2015) Ko, SH., Baeg, MK., Han, KD., Ko, SH. & Ahn, YB. (2015). Increased liver markers are associated with higher risk of type 2 diabetes. World Journal of Gastroenterology, (21)24. Recuperado el 17 de noviembre de 2018 de https://www.ncbi.nlm.nih.gov/pubmed/26139993. DOI:10.3748/wjg.v21.i24.7478

(Lindström, 2003) Lindström, J. & Tuomilehto, J. (2003). The Diabetes Risk Score - A practical tool to predict type 2 diabetes risk. Diabetes Care, (26)3. Recuperado el 03 de noviembre de 2018 de http://care.diabetesjournals.org/content/26/3/725. DOI:10.2337/ diacare.26.3.725

(OMS, 2016) Organización Mundial de la Salud. (2016). Informe mundial sobre la diabetes. La carga mundial de la diabetes, (01)1. Recuperado el 10 de diciembre de 2018 de http://www.who.int/iris/handle/10665/254649.

(Onishi, 2010) Onishi, Y., Hayashi, T., Kogawa-Sato, K., Ogihara, T., Kuzuya, N., Anai, M., Tsukuda, K., Boyko, E., Fujimoto, W. & Kikuchi, M. (2010). Fasting tests of insulin secretion and sensitivity predict future prediabetes in Japanese with normal glucose tolerance. Journal of Diabetes Investigation, (01)5. Recuperado el 17 de noviembre de 2018 de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4020720/. DOI:10.1111/j.2040-1124.2010.00041.x

(Ouyang, 2016) Ouyang, P., Guo, X., Shen, Y., Lu, N. & Ma, C. (2016). A Simple Score Model to Assess Prediabetes Risk Status Based on the Medical Examination Data, (40)5. Recuperado el 15 de diciembre de 2018 de https://www.ncbi.nlm.nih.gov/pubmed/27184300. DOI:10.1016/j.jcjd.2016.02.013

(Pimentel, 2011) Pimentel, G., Moreto, F., Takahashi, M.n Portero-Mclellan, C. & Burini, R. (2011). Sagital abdominal diameter, but not waist circumference is strongly associated with glycemia, triacilglycerols and HDL-Clevels in overweight adults. Nutrición Hospitalaria, (26)5. Recuperado el 16 de noviembre de 2018 de http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112011000500031. DOI:10.3305/nh.2011.26.5.5241

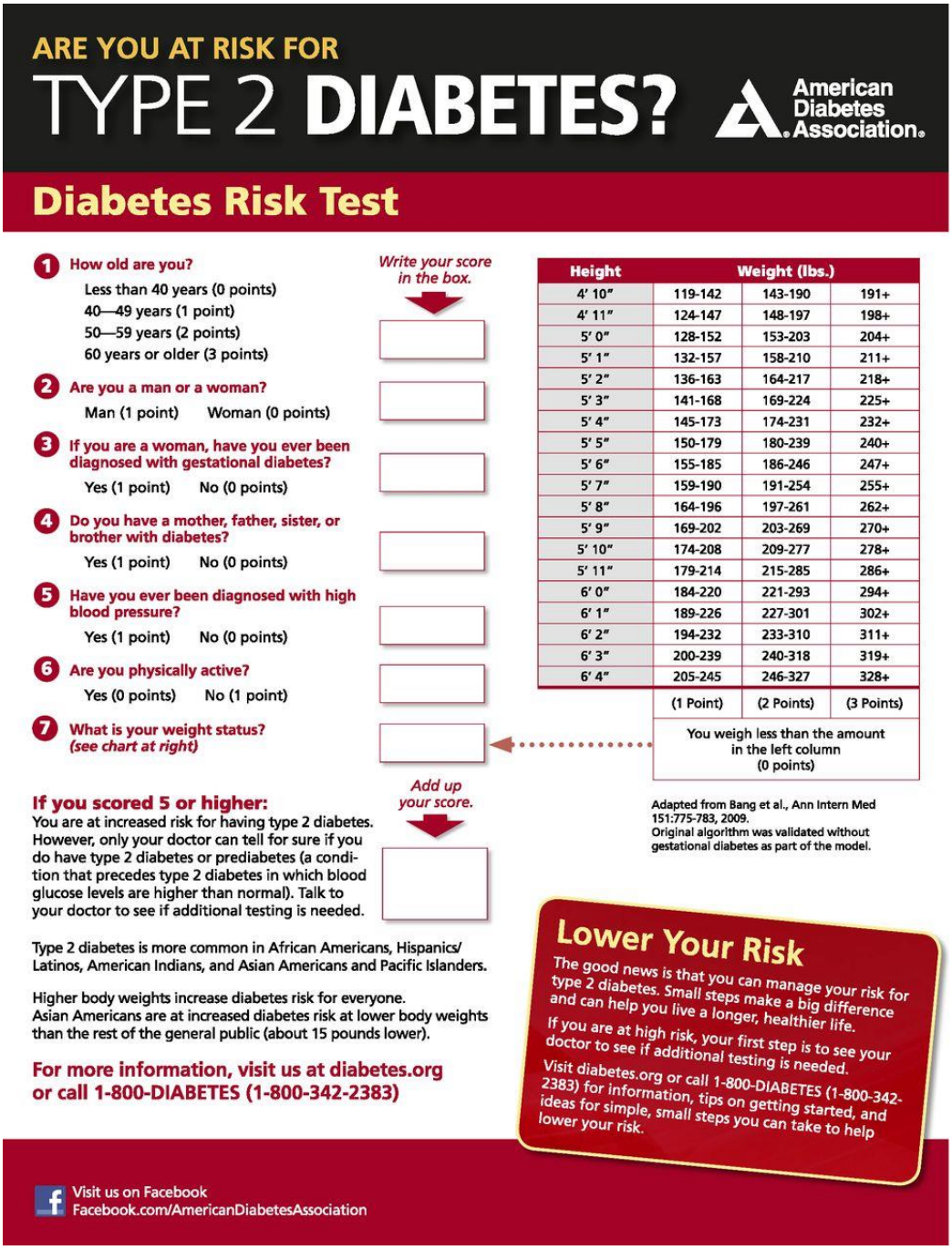
(Quinlan, 1993) Quinlan, J. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc, (01)1. ISBN-13:978-1558602380

(Rosas-Saucedo, 2017) Rosas-Saucedo, J., Caballero, E., Brito-Córdova, G., García, H., Costa, J., Lyra, R. & Rosas-Guzman, J. (2017). Consenso de Prediabetes. Documento de posición de la Asociación Latinoamericana de Diabetes (ALAD). Revista de la ALAD, (07)4, 186-187. Recuperado el 01 de noviembre de 2018 de http://www.alad-americalatina.org/wp-content/uploads/2018/03/alad_v7_n4_184-202.pdf. DOI:10.24875/ALAD.17000307

(Valenza, 2012) Valenza, M., Martín, L., González, E., Aguilar, C., Botella, M. Muñoz, T. & Valenza, T. (2012). Factores de riesgo para el síndrome metabólico en una población con apnea del sueño; evaluación en un grupo de pacientes de Granada y provincia; estudio GRANADA. Nutrición Hospitalaria, (27)4. Recuperado el 19 de noviembre de 2018 de http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112012000400042. DOI:10.3305/nh.2012.27.4.5825

Anexos

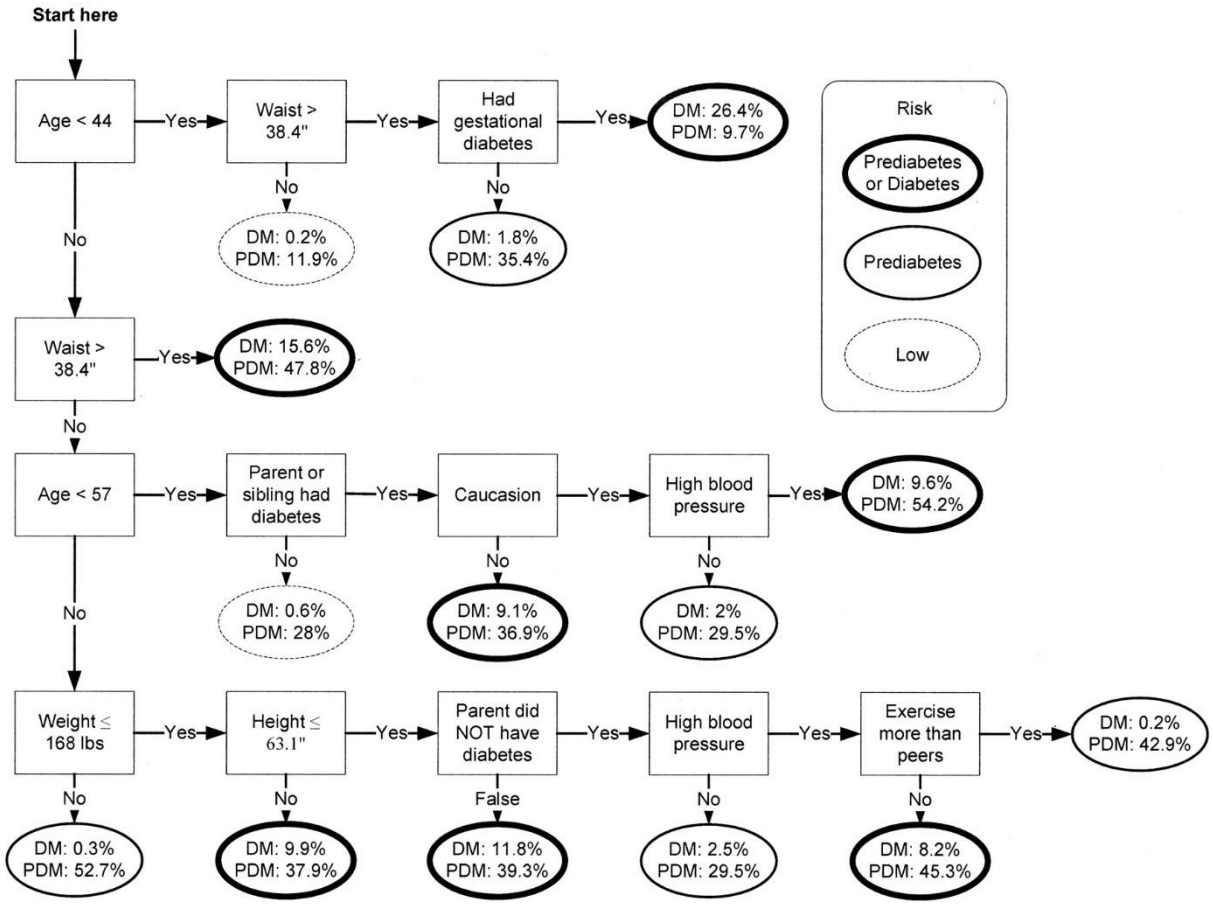
A. Asociación Americana de Diabetes – Test de riesgo de diabetes (ADA, 2017)



Anexo A. ADA – Test de riesgo de diabetes (Tamaño real)

Fuente: (ADA, 2017)

B. Árbol de Clasificación Para Identificar la Prediabetes o la Diabetes no Diagnosticada (Heikes, 2008)



Anexo B. Árbol de clasificación para identificar la prediabetes o la diabetes no diagnosticada

Fuente: (Heikes, 2008)

C. Vistas Diseñadas Para Generar el Modelo Entidad - Relación

A continuación, se muestra el código correspondiente a la vista que genera el Modelo Entidad-Relación asociado a las tablas de los años 2015-2016.

```
SELECT
      DM.ID
     ,(CASE
         WHEN DM.RIAGENDR = 1 THEN 'Hombre'
         WHEN DM.RIAGENDR = 2 THEN 'Mujer'
     END) AS Genero
     ,DM.RIDAGEYR AS Edad_Valor
     ,MA.BMDAVSAD AS DiamSagitalAbd_Valor
     ,GHB.LBXGH AS Glicohemoglobina_Valor
     ,INS.LBXIN AS Insulina_Valor
     ,TOLG.LBXGLT AS ToleranciaGlucosa_Valor
     ,GLU.LBXGLU AS GlucosaPlasma_Valor
     ,PBIOQ.LBXSTR AS Trigliceridos_Valor
     ,(CASE
         WHEN DIAB.DIQ175A = 10 THEN 'Si'
         WHEN MCND.MCQ300C = 1 THEN 'Si'
         WHEN MCND.MCQ300C = 2 THEN 'No'
         ELSE NULL
     END) AS HistoriaFamiliar
     ,(CASE
         WHEN ACF.PAD615 >= 15 OR ACF.PAD660 >= 15 OR ACF.PAD630 >= 30 OR ACF.PAD675 >= 30 THEN 'Si'
         WHEN ACF.PAD615 < 15 OR ACF.PAD660 < 15 OR ACF.PAD630 < 30 OR ACF.PAD675 < 30 THEN 'No'
     END) AS ActividadFisica
     ,SNIO.SLD012 AS HorasSueno_Valor
     ,PBIOQ.LBXSATSI AS AlanineALT_Valor
     ,ROUND(INS.LBXIN * GLU.LBXGLU / 405, 2) AS HOMAIr_Valor
     ,(CASE
         WHEN GLU.LBXGLU >= 125 THEN 'Diabetes'
         WHEN GHB.LBXGH >= 6.5 THEN 'Diabetes'
         WHEN TOLG.LBXGLT >= 200 THEN 'Diabetes'

         WHEN GLU.LBXGLU >= 100 AND GLU.LBXGLU < 125 THEN 'Prediabetes'
         WHEN GHB.LBXGH >= 5.7 AND GHB.LBXGH < 6.5 THEN 'Prediabetes'
         WHEN TOLG.LBXGLT >= 140 AND TOLG.LBXGLT < 200 THEN 'Prediabetes'

         WHEN GLU.LBXGLU < 100 THEN 'Normal'
         WHEN GHB.LBXGH < 5.7 THEN 'Normal'
         WHEN TOLG.LBXGLT < 140 THEN 'Normal'

         ELSE NULL
         END) AS Diagnostico
FROM
    dbo.DEMO_I_2015_2016 AS DM
    LEFT OUTER JOIN dbo.BMX_I_2015_2016 AS MA ON DM.ID = MA.ID
    LEFT OUTER JOIN dbo.GHB_I_2015_2016 AS GHB ON DM.ID = GHB.ID
    LEFT OUTER JOIN dbo.INS_I_2015_2016 AS INS ON DM.ID = INS.ID
    LEFT OUTER JOIN dbo.OGTT_I_2015_2016 AS TOLG ON DM.ID = TOLG.ID
    LEFT OUTER JOIN dbo.GLU_I_2015_2016 AS GLU ON DM.ID = GLU.ID
    LEFT OUTER JOIN dbo.BIOPRO_I_2015_2016 AS PBIOQ ON DM.ID = PBIOQ.ID
    LEFT OUTER JOIN dbo.DIQ_I_2015_2016 AS DIAB ON DM.ID = DIAB.ID
    LEFT OUTER JOIN dbo.MCQ_I_2015_2016 AS MCND ON DM.ID = MCND.ID
    LEFT OUTER JOIN dbo.PAQ_I_2015_2016 AS ACF ON DM.ID = ACF.ID
    LEFT OUTER JOIN dbo.SLQ_I_2015_2016 AS SNIO ON DM.ID = SNIO.ID
```

Anexo C. Generación del Modelo Entidad-Relación (Vista Vw_NHANES_2015_2016)

Fuente: Elaboración propia

D. Vista Diseñada Para Agrupar los Años de Estudio

A continuación, se muestra el código correspondiente a la vista que une la información correspondiente a cada uno de los años que se utilizan para el modelo de clasificación.

```
SELECT
    *
FROM [dbo].[Vw_NHANES_2011_2012]

UNION

SELECT
    *
FROM [dbo].[Vw_NHANES_2013_2014]

UNION

SELECT

    *
FROM [dbo].[Vw_NHANES_2015_2016]
```

Anexo D. Vista diseñada para agrupar los años de estudio

(Vista Vw_NHANES_2011_2016_UNION)

Fuente: Elaboración propia

E. Vista Diseñada Para Limpieza y Filtrado de la Información

A continuación, se muestra el código correspondiente a la vista que limpia y filtra la información descargada del NHANES para que sea analizada por Weka e ingresada al algoritmo de clasificación C4.5.

```
USE NHANES

SELECT
      [ID]
     ,[Genero]
     ,[Edad_Valor]
     ,[Edad]
     ,[DiamSagitalAbd_Valor]
     ,[Insulina_Valor]
     ,[Trigliceridos_Valor]
     ,[HistoriaFamiliar]
     ,[ActividadFisica]
     ,[HorasSueno_Valor]
     ,[AlanineALT_Valor]
     ,[HOMAIr_Valor]
     ,[Diagnostico]

FROM [dbo].[Vw_NHANES_2011_2016_UNION]

WHERE
[Edad_Valor]>=20 AND
[Edad_Valor]<=65 AND
[Genero] IS NOT NULL AND
[Etnia] IS NOT NULL AND
[DiamSagitalAbd_Valor] IS NOT NULL AND
[Insulina_Valor] IS NOT NULL AND
[Trigliceridos_Valor] IS NOT NULL AND
[HistoriaFamiliar] IS NOT NULL AND
[ActividadFisica] IS NOT NULL AND
[HorasSueno_Valor] IS NOT NULL AND
[HorasSueno_Valor]<=12 AND
[AlanineALT_Valor] IS NOT NULL AND
[HOMAIr_Valor] IS NOT NULL AND
[Diagnostico] IN ('Normal','Prediabetes')
```

100 %

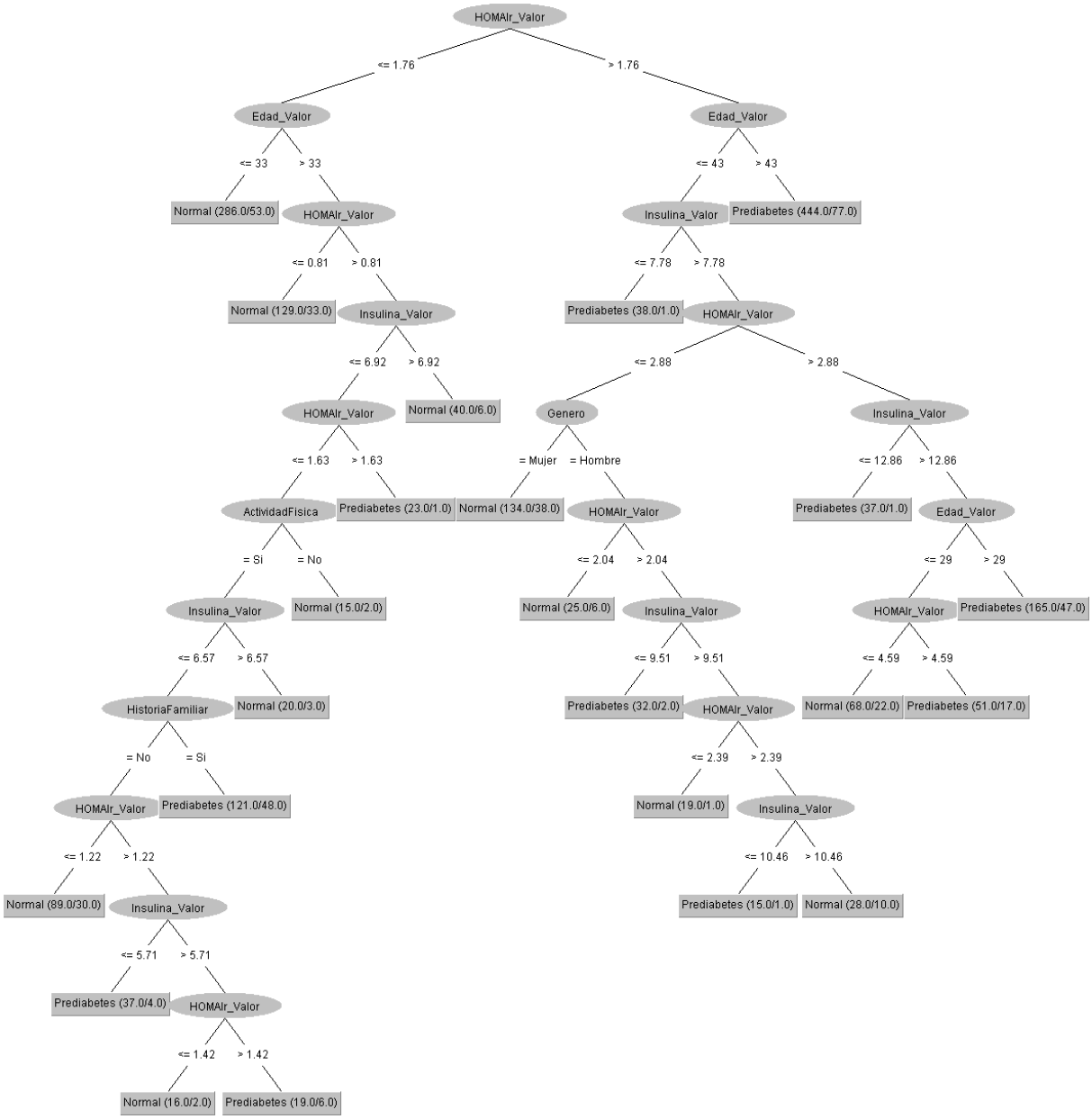
Anexo E. Vista diseñada para limpieza y filtrado de la información

(Vista Vw_NHANES_2011_2016WEKA_J48)

Fuente: Elaboración propia

F. Árbol de Clasificación en Modo Gráfico

A continuación, se muestra el árbol de clasificación obtenido en modo gráfico obtenido con Weka.



Anexo F. Árbol de clasificación en modo gráfico

Fuente: Elaboración propia

G. Código Contenido en el Documento HTML de la Visualización

A continuación, se muestra código contenido en el documento HTML de la visualización interactiva de la información.

```
<html>

    <head>


    </head>



    <body>


        <meta charset="utf-8">
        <style> /* set the CSS */

        .node circle {
          fill: #fff;
          stroke: steelblue;
          stroke-width: 3px;
        }

        .node text { font: 10px sans-serif; }

        .node--internal text {
          text-shadow: 0 1px 0 #fff, 0 -1px 0 #fff, 1px 0 0 #fff, -1px 0 0 #fff;
        }

        .link {
          fill: none;
          stroke: #ccc;
          stroke-width: 2px;
        }

        </style>


        <h1> Árbol de clasificación de la prediabetes</h1>


        <!-- load the d3.js library -->
        <script src="https://d3js.org/d3.v5.min.js"></script>

        <script src="SCRIPTarbol.js"></script>


        <h3> Por favor, ingrese los siguientes valores</h3>

        <h3 id="answer"></h3>

        <input type="text" id="Homa"> HOMA
        <br>
        <br>
        <input type="text" id="Edad"> Edad
        <br>
        <br>
        <input type="text" id="Insulina"> Insulina
        <br>
        <br>

        <datalist id="SiNo">
            <option value="Si">
            <option value="No">
        </datalist>

        <datalist id="Gen">
            <option value="Masculino">
            <option value="Femenino">
        </datalist>

        <input list="SiNo" type="text" id="Actividad"> Actividad Física
        <br>
        <br>
        <input list="SiNo" type="text" id="HFamiliar"> Historia Familiar
        <br>
        <br>
        <input list="Gen" type="text" id="Genero"> Género
        <br>
        <br>
        <button onclick="calculate()"> Calcular </button>

        <script>

            function calculate()
            {
                var field1=document.getElementById("Homa").value;
                var field2=document.getElementById("Edad").value;
                var field3=document.getElementById("Insulina").value;
                var field5=document.getElementById("Actividad").value;
                var field6=document.getElementById("HFamiliar").value;
                var field7=document.getElementById("Genero").value;

```

```
                if(parseFloat(field2)<20)
                    {
                        document.getElementById("answer").innerHTML="Para utilizar este sistema debe tener entre 20 y 65 años"
                        console.log("1")
                    } else if(parseFloat(field2)>65)
                    {
                        document.getElementById("answer").innerHTML="Para utilizar este sistema debe tener entre 20 y 65 años"
                        console.log("2")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)<=33)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 84.4%"
                        console.log("3")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)<=0.81)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 79.6%"
                        console.log("4")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)>6.92)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 87%"
                        console.log("5")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)<=6.92 &
                    parseFloat(field1)>1.63)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        95.8%"
                        console.log("6")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)<=6.92 &
                    parseFloat(field1)<=1.63 & field5=="No")
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 88.2%"
                        console.log("7")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)<=6.92 &
                    parseFloat(field1)<=1.63 & field5=="Si" & parseFloat(field3)>6.57)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 87%"
                        console.log("8")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)<=6.92 &
                    parseFloat(field1)<=1.63 & field5=="Si" & parseFloat(field3)<=6.57 & field6=="Si")
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        71.6%"
                        console.log("9")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)<=6.92 &
                    parseFloat(field1)<=1.63 & field5=="Si" & parseFloat(field3)<=6.57 & field6=="No" & parseFloat(field1)<=1.22)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 74.8%"
                        console.log("10")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)<=6.92 &
                    parseFloat(field1)<=1.63 & field5=="Si" & parseFloat(field3)<=6.57 & field6=="No" & parseFloat(field1)>1.22 & parseFloat(field3)
                    <=5.71)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        90.2%"
                        console.log("11")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)<=6.92 &
                    parseFloat(field1)<=1.63 & field5=="Si" & parseFloat(field3)<=6.57 & field6=="No" & parseFloat(field1)>1.22 &
                    parseFloat(field3)>5.71 & parseFloat(field1)<=1.42)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 88.9%"
                        console.log("12")
                    } else if(parseFloat(field1)<=1.76 & parseFloat(field2)>33 & parseFloat(field1)>0.81 & parseFloat(field3)<=6.92 &
                    parseFloat(field1)<=1.63 & field5=="Si" & parseFloat(field3)<=6.57 & field6=="No" & parseFloat(field1)>1.22 &
                    parseFloat(field3)>5.71 & parseFloat(field1)>1.42)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        70%"
                        console.log("13")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)>43)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        85.2%"
                        console.log("14")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)>43)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        85.2%"
                        console.log("15")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)<=7.78)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        97.4%"
                        console.log("16")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)>2.88 & parseFloat(field3)
                    <=12.86)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        97.4%"
                        console.log("17")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)>2.88 &
                    parseFloat(field3)>12.86 & parseFloat(field2)>29)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        77.8%"
                        console.log("18")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)>2.88 &
                    parseFloat(field3)>12.86 & parseFloat(field2)<=29 & parseFloat(field1)>4.59)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        75%"
                        console.log("19")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)>2.88 &
                    parseFloat(field3)>12.86 & parseFloat(field2)<=29 & parseFloat(field1)<=4.59)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 75.6%"
                        console.log("20")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)<=2.88 &
                    field7=="Femenino")
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 77.9%"
                        console.log("21")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)<=2.88 &
                    field7=="Masculino" & parseFloat(field1)<=2.04)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 80.1%"
                        console.log("22")
```

```
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)<=2.88 &
                    field7=="Masculino" & parseFloat(field1)>2.04 & parseFloat(field3)<=9.51)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        94.1%"
                        console.log("23")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)<=2.88 &
                    field7=="Masculino" & parseFloat(field1)>2.04 & parseFloat(field3)>9.51 & parseFloat(field1)<=2.39)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 95%"
                        console.log("24")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)<=2.88 &
                    field7=="Masculino" & parseFloat(field1)>2.04 & parseFloat(field3)>9.51 & parseFloat(field1)>2.39 & parseFloat(field3)<=10.46)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        93.7%"
                        console.log("25")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)<=2.88 &
                    field7=="Masculino" & parseFloat(field1)>2.04 & parseFloat(field3)>9.51 & parseFloat(field1)>2.39 & parseFloat(field3)<=10.46)
                    {
                        document.getElementById("answer").innerHTML="Usted ha sido clasificado como prediabético, con una probabilidad de acierto del
                        93.7%"
                        console.log("26")
                    } else if(parseFloat(field1)>1.76 & parseFloat(field2)<=43 & parseFloat(field3)>7.78 & parseFloat(field1)<=2.88 &
                    field7=="Masculino" & parseFloat(field1)>2.04 & parseFloat(field3)>9.51 & parseFloat(field1)>2.39 & parseFloat(field3)>10.46)
                    {
                        document.getElementById("answer").innerHTML="Usted se encuentra sano con una probabilidad de acierto del 73.7%"
                        console.log("27")
                    } else
                    {
                        document.getElementById("answer").innerHTML="No se cuenta con información suficiente para el proceso de clasificación. Por
                        favor ingrese una mayor cantidad de variables y vuelva a intentar"
                        console.log("11")
                    }

            }


        </script>

    </body>



</html>
```

Anexo G. Código del documento HTML de la visualización interactiva

Fuente: Elaboración propia

H. Código Contenido en el Documento JS de la Visualización

A continuación, se muestra código contenido en el documento JS de la visualización, donde se encuentra la graficación del árbol de clasificación.

```
console.log("Inicio del archivo js")

var treeData =
{ "name": "HOMA",
  "children": [
  { "name": "Edad (HOMA ≤ 1.76)",
    "children": [
    { "name": "Normal (Edad ≤ 33)"},
    { "name": "HOMA (Edad > 33)",
      "children": [
      { "name": "Normal (HOMA ≤ 0.81)"},
      { "name": "Insulina (HOMA > 0.81)",
        "children": [
        { "name": "HOMA (Insulina ≤ 6.92)",
          "children": [
          { "name": "Actividad Física (HOMA ≤ 1.63)",
            "children" : [
            { "name" : "Insulina (Act.Física = Si)",
              "children" : [
              { "name" : "Historia Familiar (Insulina ≤ 6.57)",
                "children" : [
                { "name" : "HOMA (HFamiliar = No)",
                  "children" : [
                  { "name" : "Normal (HOMA ≤ 1.22)"},
                  { "name" : "Insulina (HOMA > 1.22)",
                    "children" : [
                    { "name" : "Prediabetes (Insulina ≤ 5.71)"},
                    { "name" : "HOMA (Insulina > 5.71)",
                      "children" : [
                      { "name" : "Normal (HOMA ≤ 1.42)"},
                      { "name" : "Prediabetes (HOMA > 1.42)"}
                      ]
                    }
                    ]
                  }
                  ]
                },
                { "name" : "Prediabetes (HFamiliar = Si)" }
                ]
              },
              { "name" : "Normal (Insulina > 6.57)"},
              ]
            },
            { "name" : "Normal (Act.Física = No)"},
            ]
          },
          { "name" : "Prediabetes (HOMA > 1.63)"}
          ]
        },
        { "name": "Normal (Insulina >6.92)"}
        ]
      }
      ]
    }
    ]
  },
  { "name": "Edad (HOMA>1.76)",
    "children" : [
    { "name" : "Insulina (Edad ≤ 43)",
      "children" : [
      { "name" : "Prediabetes (Insulina ≤ 7.78)"},
      { "name" : "HOMA (Insulina > 7.78)",
        "children" :[
        { "name" : "Género (HOMA ≤ 2.88)",
          "children" : [
          { "name" : "Normal (Género Femenino)"},
          { "name" : "HOMA (Género Masculino)",
            "children" :[
            { "name" : "Normal (HOMA ≤ 2.04)"},
            { "name" : "Insulina (HOMA > 2.04)",
              "children": [
              { "name" : "Prediabetes (Insulina ≤ 9.51)" },
              { "name" : "HOMA (Insulina > 9.51)",
                "children" : [
                { "name" : "Normal (HOMA ≤ 2.39)"},
                { "name" : "Insulina (HOMA > 2.39)",
                  "children" : [
                  { "name" : "Prediabetes (Insulina ≤ 10.46)"},
                  { "name" : "Normal (Insulina > 10.46)"}
                  ]
                }
                ]
              }
              ]
            }
            ]
          }
          ]
        }
        ]
      },
```

```
        { "name" : "Insulina (HOMA > 2.88)",
          "children" : [
          { "name" : "Prediabetes (Insulina ≤ 12.86)"},
          { "name" : "Edad (Insulina > 12.86)",
            "children" : [
            { "name" : "HOMA (Edad ≤ 29)",
              "children" : [
              { "name" : "Normal (HOMA ≤ 4.59)"},
              { "name" : "Prediabetes (HOMA > 4.59)"}
              ]
            },
            { "name" : "Prediabetes (Edad > 29)"}
            ]
          }
          ]
        }
        ]
      }
      ]
    },
    { "name" : "Prediabetes (Edad > 43)"}
   ]
 }
    ]
};


// set the dimensions and margins of the diagram
var margin = {top: 40, right: 10, bottom: 50, left: 0},
    width = 1700 - margin.left - margin.right,
    height = 768 - margin.top - margin.bottom;

// declares a tree layout and assigns the size
var treemap = d3.tree()
    .size([width, height]);

//  assigns the data to a hierarchy using parent-child relationships
var nodes = d3.hierarchy(treeData);

// maps the node data to the tree layout
nodes = treemap(nodes);

// append the svg obgect to the body of the page
// appends a 'group' element to 'svg'
// moves the 'group' element to the top left margin
var svg = d3.select("body").append("svg")
      .attr("width", width + margin.left + margin.right)
      .attr("height", height + margin.top + margin.bottom),
    g = svg.append("g")
      .attr("transform",
            "translate(" + margin.left + "," + margin.top + ")");

// adds the links between the nodes
var link = g.selectAll(".link")
    .data( nodes.descendants().slice(1))
  .enter().append("path")
    .attr("class", "link")
    .attr("d", function(d) {
       return "M" + d.x + "," + d.y
         + "C" + d.x + "," + (d.y + d.parent.y) / 2
         + " " + d.parent.x + "," +  (d.y + d.parent.y) / 2
         + " " + d.parent.x + "," + d.parent.y;
       });

// adds each node as a group
var node = g.selectAll(".node")
    .data(nodes.descendants())
  .enter().append("g")
    .attr("class", function(d) {
      return "node" +
        (d.children ? " node--internal" : " node--leaf"); })
    .attr("transform", function(d) {
      return "translate(" + d.x + "," + d.y + ")"; });

// adds the circle to the node
node.append("circle")
  .attr("r", 10);

// adds the text to the node
node.append("text")
  .attr("dy", ".35em")
  .attr("y", function(d) { return d.children ? -20 : 20; })
  .style("text-anchor", "middle")
  .text(function(d) { return d.data.name; });

```

Anexo H. Código del documento JS de la visualización interactiva

Fuente: Elaboración propia