CONSTRUCCIÓN DE UN MODELO DE PREDICCIÓN PARA APOYO AL DIAGNÓSTICO DE DIABETES

CONSTRUCTION OF A PREDICTION MODEL TO SUPPORT THE DIABETES DIAGNOSIS

Orlando Adrián Chan May
TecNM / Instituto Tecnológico Superior del Sur del Estado de Yucatán
oacmay@hotmail.com

Jimmy Josué Peña Koo
TecNM / Instituto Tecnológico Superior del Sur del Estado de Yucatán
jimjpk@hotmail.com

Jean Marie Vianne Kinani
TecNM / Instituto Tecnológico Superior de Huichapan
jmvianney@iteshu.edu.mx

Manuel Abraham Zapata Encalada
TecNM / Instituto Tecnológico de Chetumal
mzapata_00@hotmail.com

Resumen

En el presente trabajo se procesaron datos relevantes de 768 pacientes para apoyar en la predicción de diabetes de las personas. Entre las variables consideradas en el estudio se emplearon: glucosa, presión sanguínea, insulina, edad, entre otros, por medio de sistemas de aprendizaje automático y sistemas expertos con aprendizaje supervisado para generar árboles de decisión, así como el análisis de resultados del algoritmo de predicción J48, con las herramientas BigML y Weka, respectivamente. Esta investigación será la base para desarrollar un sistema experto que apoye los diagnósticos de diabetes en comunidades rurales que carezcan de personal médico o equipo.

Palabra(s) Clave: Aprendizaje supervisado, BigML, Machine Learning, Sistema experto, Weka.

Abstract

In the present work, relevant data of 768 patients were processed to help the prediction of diabetes in people. Among the variables considered in the study were

used: glucose, blood pressure, insulin, age, to name a few, through of machine learning and expert systems with supervised learning to generate decision trees, as well as the analysis of results of prediction algorithm J48, using the BigML and Weka tools, respectively. This research will be the start for developing an expert system that helps diabetes diagnoses in marginalized people that lack doctors or equipment.

Keywords**:** *BigML, Expert system, Machine Learning, Supervised learning, Weka.*

1. Introducción

El aprendizaje automático o Machine Learning se engloba dentro de las disciplinas de la Inteligencia Artificial. Es un método científico que permite usar las computadoras y otros dispositivos con capacidad computacional para que aprendan a extraer los patrones y relaciones que hay en nuestros datos por sí solos. Esos patrones se pueden usar luego para predecir comportamientos y en la toma de decisiones, aplicando para ello la integración de diferentes recursos tecnológicos..

Un sistema experto puede definirse como un sistema informático (hardware y software) que simula a los expertos humanos en un área de especialización dada. Aunado a lo anterior, la diabetes es el principal problema de salud en México. El número creciente de casos, el elevado porcentaje que desarrollan complicaciones tardías y el costo del tratamiento hacen insuficientes los esfuerzos para confrontarla. La prevención de la diabetes es la forma más plausible para modificar el crecimiento de la epidemia. Los sujetos que desarrollarán la enfermedad pueden ser detectados y existen intervenciones que disminuyen la incidencia de la enfermedad [Aguilar y Gómez, 2006]. Por esta razón se origina el presente estudio, con la finalidad de ser una herramienta auxiliar en el diagnóstico de dicho padecimiento.

2. Métodos

De acuerdo con Hernández *et al.* [2010], la investigación es aplicada por su propósito con un enfoque cuantitativo, de tipo pre-experimental, debido a que se

manipula intencionalmente la variable independiente, el proceso de decisión, para analizar las consecuencias sobre la variable dependiente, padecimiento de la diabetes.

Para el desarrollo de sistemas expertos existen varias metodologías y métodos, entre las principales están las propuestas por Buchanan, Grover, Weiss y Kulikowski y la metodología IDEAL. Estas metodologías cuyas fases principales se pueden observar en la tabla 1, estructuran el desarrollo en etapas desde el planteamiento del problema hasta la evaluación [Ferrer *et al.*, 2015].

Tabla 1 Metodologías para el desarrollo de SE.

Metodología de Buchanan	**Metodología de Grover**
Etapa 1: Familiarización con el problema y el dominio. Etapa 2: Delimitación del sistema. Etapa 3: Obtención de la estructura de inferencia del Sistema Experto. Etapa 4: Definición del Sistema Experto prototipo. Etapa 5: Depuración del sistema prototipo. Etapa 6: Optimización del Sistema Experto prototipo.	Etapa 1. Definición del dominio. Etapa 2. Formulación fundamental del conocimiento. Etapa 3. Consolidación del conocimiento basal.
Metodología de Weiss y Kulikowski	**Metodología IDEAL**
Etapa 1. Planteamiento del problema. Etapa 2. Encontrar expertos humanos que puedan resolver el problema. Etapa 3. Diseño de un Sistema Experto. Etapa 4. Elección de la herramienta de desarrollo. Etapa 5. Desarrollo y prueba de un prototipo. Etapa 6. Refinamiento y generalización. Etapa 7. Mantenimiento y puesta al día.	Fase 1. Identificación de la tarea. 1.1. Plan de requisitos y adquisición de conocimientos. 1.2. Evaluación y selección de la tarea. 1.3. Definiciones de las características de la tarea. Fase 2. Desarrollo de los prototipos. 2.1. Concepción de la solución. 2.2. Adquisición de conocimientos y conceptualización de los conocimientos. 2.3. Formalización de los conocimientos. 2.4. Selección de la herramienta e implementación. 2.5. Validación y evaluación del prototipo. 2.6. Definición de nuevos requisitos, especificaciones y diseño. Fase 3. Ejecución de la construcción del sistema integrado. Fase 4. Actuación para conseguir el mantenimiento perfectivo. Fase 5. Lograr una adecuada transferencia tecnológica.

Fuente: [Ferrer et al., 2015].

En el presente trabajo se aplicó una variante de la metodología de Weiss y Kulikowski, propuesta por Favret *et al.* [2018], que consta de las siguientes etapas:

- **Etapa 1 Planteamiento del problema**. El problema principal es la falta de un diagnóstico o estudio médico adecuado de la diabetes en los pacientes que se encuentran en zonas rurales y con bajo presupuesto, debido a la carencia de equipos o porque no se cuenta con médicos en las comunidades que se encarguen de llevar a cabo dicho diagnóstico o

simplemente interpretar los resultados obtenidos de otros centros de salud, lo que obliga el envío de las personas a otras unidades de medicina familiar, produciendo pérdida de tiempo y gastos económicos innecesarios.

- **Etapa 2 Encontrar expertos humanos**. Con capacidad y disposición para resolver el problema. Considerando el planteamiento anterior, se ha observado que actualmente en las ciudades, los hospitales de tamaño mediano y grande, poseen principalmente medidores de glucosa sanguínea como instrumento para realizar un diagnóstico de manera automatizada y determinar si una persona es propensa a padecer diabetes. En muchas ocasiones, estos medidores no generan resultados confiables y por ende, se transfiere a los pacientes a estudios de laboratorio más completos donde se analizan otras variables, además de la glucosa. Por consiguiente, cuando el sistema experto esté concluido, será posible tener el apoyo de personal médico para la interpretación de los valores que resulten de los diagnósticos de laboratorio para corroborar aquellos que se generen por medio del sistema experto, como parte de los objetivos futuros.
- **Etapa 3 Diseño de un sistema experto**. El sistema experto contendrá reglas de inferencia o predicción, a partir del proceso y análisis de un conjunto de datos muestra del Center for Machine Learning and Intelligent Systems de la Universidad de Irvine, California (UCI, por sus siglas en inglés), el cual dispone de las características más importantes de personas que pudieran padecer o no la diabetes [Dua & Karra, 2017].

  Es importante mencionar, que los repositorios de aprendizaje automático de UCI son colecciones de bases de datos, que se pueden utilizar libremente por estudiantes o investigadores para el análisis empírico de algoritmos de aprendizaje automático, propósito principal de su creación. Además, dispone de una política de citas, donde se aclara que al publicar material a través de estos repositorios se deben incluir las citas a los donantes de dichas fuentes, lo que ayudará a otros a obtener los conjuntos de datos y replicar en otros experimentos. En trabajos futuros, cuando el sistema experto se encuentre en fase de experimentación real, se realizarán los

oficios y permisos necesarios tanto para el tratamiento con los médicos, los pacientes y los centros médicos, respectivamente, para evitar problemas.

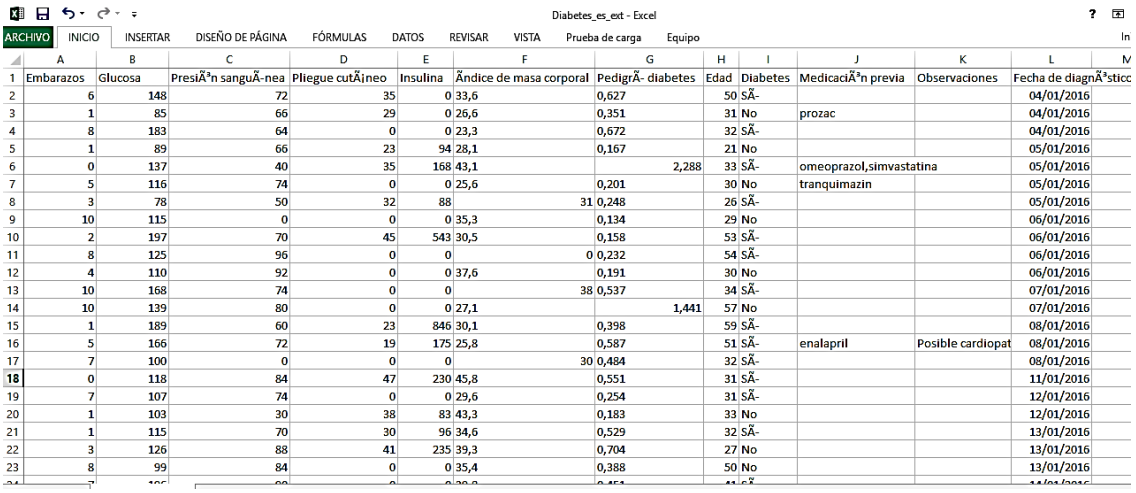
El conjunto de datos se trabajó en formato Attribute Relation File Format (ARFF) para el caso de Weka, como puede observarse en la figura 1. Para la herramienta BigML, se procesó en formato de Valores Separados por Comas (CSV), mismo que se observa en la figura 2. En ambos casos el número de instancias que se procesaron fue de 768 registros, que corresponden al mismo número de pacientes, cuyos atributos relevantes fueron nueve: número de embarazos, glucosa, presión sanguínea, piel, insulina, índice de masa corporal, pedigrí diabetes, edad y finalmente el atributo objetivo, la diabetes.

diabetes.arff

```
% Number of Instances: 768
%
% For Each Attribute: (all numeric-valued)
%    1. Number of times pregnant
%    2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
%    3. Diastolic blood pressure (mm Hg)
%    4. Triceps skin fold thickness (mm)
%    5. 2-Hour serum insulin (mu U/ml)
%    6. Body mass index (weight in kg/(height in m)^2)
%    7. Diabetes pedigree function
%    8. Age (years)
%    9. Class variable (0 or 1)
%
% Class Distribution: (class value 1 is interpreted as "tested positive for
%    diabetes", 0 is interpreted as "tested negative )
%
%    Class Value  Number of instances
%    0            500
%    1            268
%
@relation pima_diabetes
@attribute 'preg' numeric
@attribute 'plas' numeric
@attribute 'pres' numeric
@attribute 'skin' numeric
@attribute 'insu' numeric
@attribute 'mass' numeric
@attribute 'pedi' numeric
@attribute 'age' numeric
@attribute 'class' { tested_negative, tested_positive}
@data
6,148,72,35,0,33.6,0.627,50,tested_positive
1,85,66,29,0,26.6,0.351,31,tested_negative
8,183,64,0,0,23.3,0.672,32,tested_positive
1,89,66,23,94,28.1,0.167,21,tested_negative
0,137,40,35,168,43.1,2.288,33,tested_positive
5,116,74,0,0,25.6,0.201,30,tested_negative
3,78,50,32,88,31,0.248,26,tested_positive
```

Fuente: Elaboración propia.

Figura 1 Dataset de diabetes en formato ARFF.



Diabetes_es_ext - Excel

ARCHIVO | INICIO | INSERTAR | DISEÑO DE PÁGINA | FÓRMULAS | DATOS | REVISAR | VISTA | Prueba de carga | Equipo

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Embarazos	Glucosa	PresiÃ³n sanguÃ-nea	Pliegue cutÃ¡neo	Insulina	Ãndice de masa corporal	PedigrÃ- diabetes	Edad	Diabetes	MedicaciÃ³n previa	Observaciones	Fecha de diagnÃ³stico	
2	6	148	72	35	0	33,6	0,627	50	SÃ-			04/01/2016	
3	1	85	66	29	0	26,6	0,351	31	No	prozac		04/01/2016	
4	8	183	64	0	0	23,3	0,672	32	SÃ-			04/01/2016	
5	1	89	66	23	94	28,1	0,167	21	No			05/01/2016	
6	0	137	40	35	168	43,1	2,288	33	SÃ-	omeoprazol,simvastatina		05/01/2016	
7	5	116	74	0	0	25,6	0,201	30	No	tranquimazin		05/01/2016	
8	3	78	50	32	88	31	0,248	26	SÃ-			05/01/2016	
9	10	115	0	0	0	35,3	0,134	29	No			06/01/2016	
10	2	197	70	45	543	30,5	0,158	53	SÃ-			06/01/2016	
11	8	125	96	0	0	0	0,232	54	SÃ-			06/01/2016	
12	4	110	92	0	0	37,6	0,191	30	No			06/01/2016	
13	10	168	74	0	0	38	0,537	34	SÃ-			07/01/2016	
14	10	139	80	0	0	27,1	1,441	57	No			07/01/2016	
15	1	189	60	23	846	30,1	0,398	59	SÃ-			08/01/2016	
16	5	166	72	19	175	25,8	0,587	51	SÃ-	enalapril	Posible cardiopat	08/01/2016	
17	7	100	0	0	0	30	0,484	32	SÃ-			08/01/2016	
18	0	118	84	47	230	45,8	0,551	31	SÃ-			11/01/2016	
19	7	107	74	0	0	29,6	0,254	31	SÃ-			12/01/2016	
20	1	103	30	38	83	43,3	0,183	33	No			12/01/2016	
21	1	115	70	30	96	34,6	0,529	32	SÃ-			13/01/2016	
22	3	126	88	41	235	39,3	0,704	27	No			13/01/2016	
23	8	99	84	0	0	35,4	0,388	50	No			13/01/2016	

Fuente: Elaboración propia.

Figura 2 Dataset de diabetes en formato CSV.

Los datos se estructuran de manera que cada fila representa un paciente y las columnas cada propiedad o atributo que se emplea para el aprendizaje del sistema experto. Otro aspecto de diseño del sistema experto consiste en la utilización de soluciones basadas en aprendizaje automático con un enfoque supervisado, por la naturaleza de los conjuntos de datos donde se encuentra información histórica importante para el entrenamiento del sistema experto, mismo que genera árboles de decisión.

No se utilizaron las variables medicamentos previos, observaciones del médico y fecha de diagnóstico, porque no intervienen directamente como medidas críticas para determinar si una persona puede padecer o no la enfermedad. Como ejemplo concreto se puede mencionar la fecha de diagnóstico: toda persona puede acudir cualquier día a realizarse un estudio de diabetes y puede resultar positivo o negativo. En el caso del atributo de número de embarazos, 111 pacientes resultaron con cero embarazos, lo que no necesariamente significa que sea hombre o mujer, pero, a diferencia de la fecha del diagnóstico, éste dato en combinación con los demás, resulta sobresaliente para la predicción de la enfermedad, sin importar el sexo de la persona.

- **Etapa 4 Elección de la herramienta de desarrollo**. Para verificar el funcionamiento del sistema experto se trabajó con Weka y BigML,

respectivamente para corroborar los resultados obtenidos en las predicciones. En ambos casos se trabajó, en primera instancia, con árboles de clasificación, uno de los más utilizados en el tema de aprendizaje automático y están dentro de los métodos de clasificación supervisada, formada por una variable dependiente (clase), cuyo objetivo es averiguar dicha clase para casos nuevos. El modelo así obtenido puede servir para clasificar casos cuyas clases se desconozcan o, simplemente, para comprender mejor la información de la que se dispone. Otros algoritmos expertos de clasificación, similares a los árboles, son el algoritmo CART, el IDE3, C45 y el J48, que ayudan al proceso de clasificación. Como segunda instancia de clasificación se empleó el algoritmo J48, en los formatos RTFF y CSV, como se había comentado en la etapa 3, para clasificar si un paciente resulta positivo o negativo a la diabetes.

Otras plataformas existentes para aprendizaje automatizado son: AmazonML, AzureML y Google Prediction API, los cuales forman parte de ecosistemas más extensos desde servicios web, almacenamiento en la nube, automatización de la implementación y mucho más, características innecesarias para el presente estudio. Por el contrario, se decidió trabajar con Weka y BigML, por las características que se describen a continuación:

- ✓ **WEKA.** Este entorno que se presenta en la figura 3, fue creado por la Universidad de Waikato, Nueva Zelanda. Reúne una colección de algoritmos de aprendizaje máquina; su uso más común es para minería de datos ya que permite trabajar directamente con un conjunto de datos o datos provenientes de aplicaciones java; tiene una Licencia GPL (GNU PublicLicense) y trabaja con un formato ARFF, que se compone por cabecera (@ relation pima_diabetes); declaración de atributos (@attribute ‘preg’ numeric,...) y sección de datos (@data 6, 148, 72, 35, 0, 33.6, 0.627, 50, tested_positive, ...), como se presentó en la figura 2.

  Weka permite trabajar con Big Data a partir de Weka 3.7; utilizar el cliente simple o el KnowledgeFlow, así como la posibilidad de utilizar

carga de datos incremental y entornos distribuidos, como Hadoop y Spark [Durrant, B., Frank, E., Hunt, L. & Holmes, G., 2018].

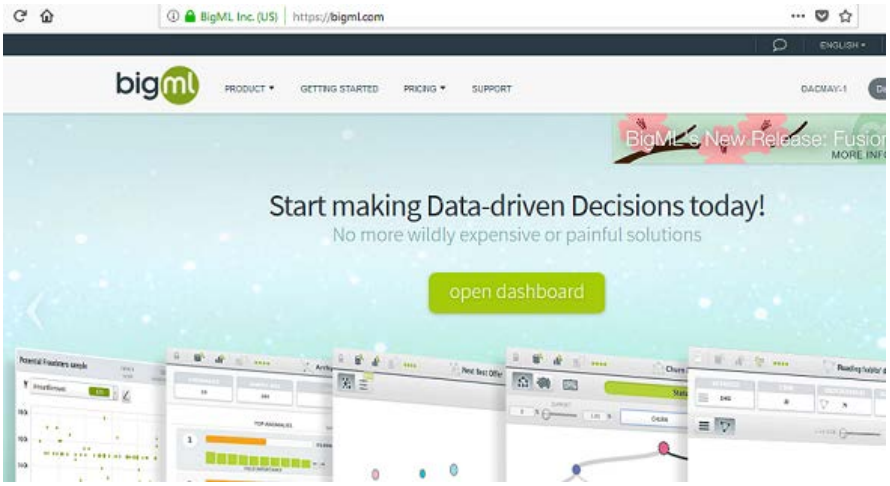
Fuente: University of Wikatao.

Figura 3 Panel principal de la herramienta Weka.

- ✓ **BigML.** Este entorno de trabajo ofrece una plataforma administrada para construir y compartir conjuntos de datos y modelos en forma de aprendizaje automático como servicio (MLaaS, por sus siglas en inglés), como consecuencia, BigML es una propuesta para hacer que el aprendizaje automático sea comprensible para los usuarios. La vista principal se muestra en la figura 4. También, BigML, logra explotar los beneficios de las soluciones de nube existentes. Por ejemplo, permite la importación de datos desde AWS S3, MS Azure, Google Storage, Google Drive, Dropbox, etc., lo que beneficia a los desarrolladores porque las infraestructuras de nube públicas podrían convertirse en un producto básico, es decir, una solución para los proveedores de diversos servicios.

  Además, al estar centrado solo en el aprendizaje automático, BigML ofrece un conjunto amplio de características, todas bien integradas dentro de una interfaz de usuario donde se pueden cargar conjuntos de datos, capacitar, evaluar modelos y generar nuevas predicciones, una por una o en un lote. Contiene una amplia galería de conjuntos de datos y modelos gratuitos para probar, bien organizados en categorías y accesibles al público, al igual que, algoritmos de

agrupación y visualización para crear modelos de alta calidad [Casalboni, 2015].



Fuente: https://bigml.com/

Figura 4 Panel principal de BigML.

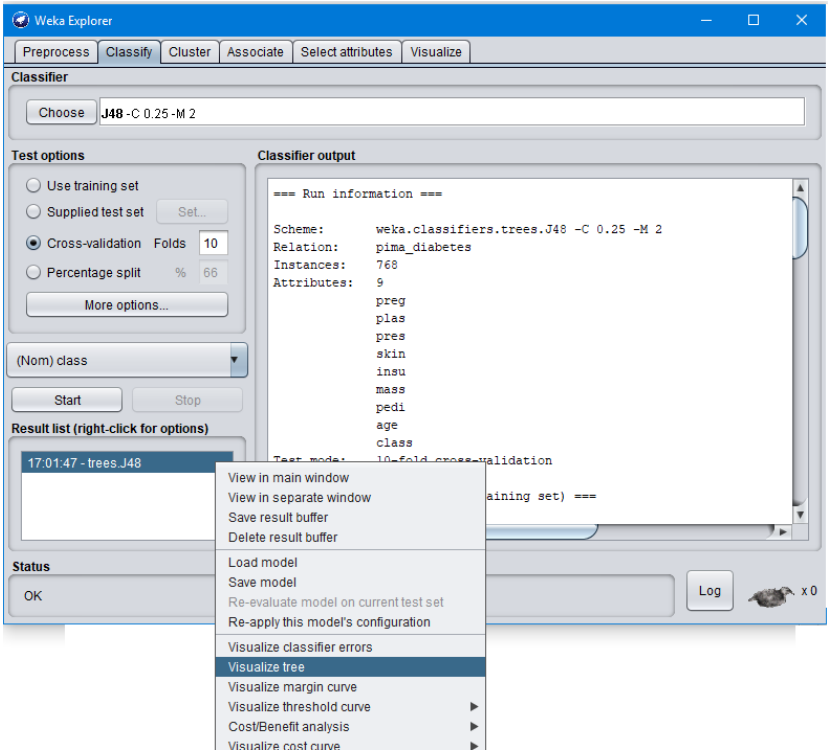
- **Etapa 5 Desarrollo y prueba de un prototipo**. Con la ejecución en ambas plataformas se generaron árboles de decisiones con una confiabilidad por encima del 70%. Después se analizó la información para el entrenamiento y predicción del padecimiento o no de la diabetes a través de un sistema de aprendizaje supervisado y automático.

  Para el desarrollo y procesamiento del conjunto de datos con el entorno de trabajo Weka se realizaron los pasos siguientes:

  - ✓ Descarga del instalador desde el sitio web oficial de la aplicación en http://www.cs.waikato.ac.nz/ml/weka/downloading.html
  - ✓ Ejecutar Weka y seleccionar la opción "Explorer".
  - ✓ Desde preprocesado, abrir la base de datos (open file) que se encuentra en la carpeta data del directorio donde se instaló. Seleccionar la base de datos en formato ARFF.
  - ✓ Crear un árbol de decisión, mediante la pestaña Classify, algoritmo J48 [Choose-trees-J48].
  - ✓ Ejecutar y con clic derecho, opción Result list, seleccionar Visualize tree.

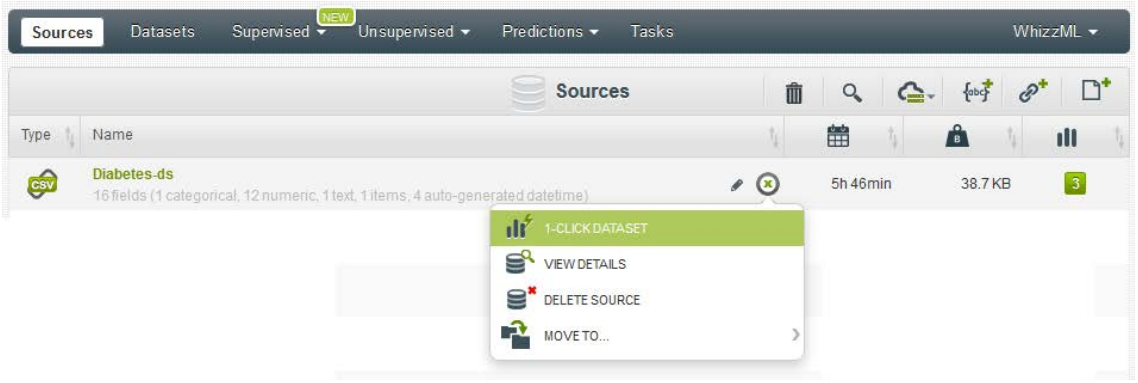
Lo anterior se puede ver en la figura 5. Para el procesado de la información con BigML, se siguieron los pasos que se describen a continuación:

- ✓ Registro en https://bigml.com/accounts/register/
- ✓ Activación de cuenta por enlace recibido en correo electrónico.
- ✓ Acceso a la plataforma: https://bigml.com
- ✓ En el panel principal o dashboard se subió el archivo CSV a través de la opción "Create a source from a URL" y escribiendo s3://bigml-public/csv/Diabetes_es_ext.csv en la casilla de texto. También es posible descargar la base de datos: https://archive.ics.uci.edu/ml /index.php.
- ✓ Utilizando el objeto Source se construyó un dataset con la acción 1-click dataset para analizar la distribución de los datos. Esto se presenta en la figura 6.
- ✓ Para generar el árbol de decisión, se hace clic en la opción dataset, 1-click supervised-model.



Fuente: Elaboración propia.

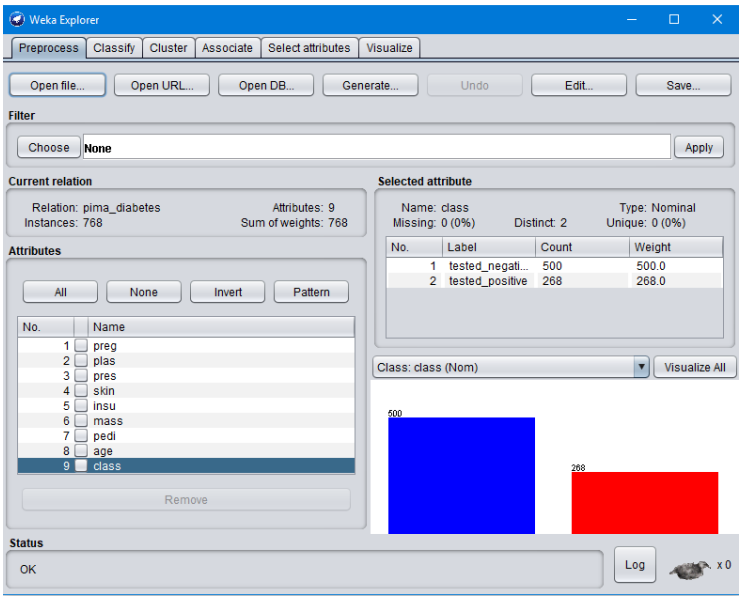
Figura 5 Prueba de prototipo con Weka.



Fuente: Elaboración propia.

Figura 6 Prueba de prototipo con BigML.

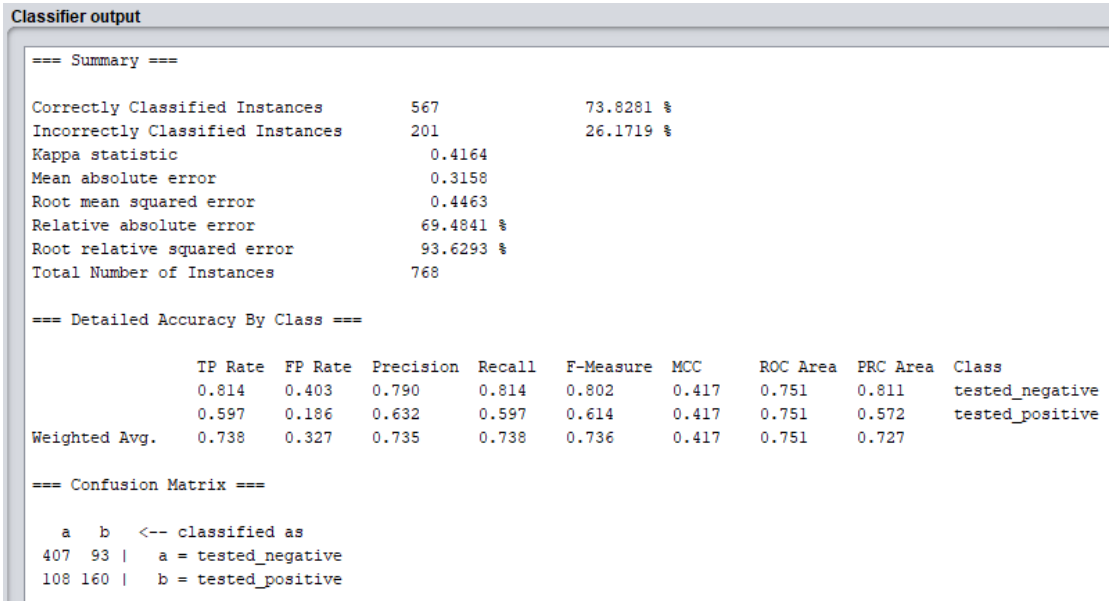
- **Etapa 6 Refinamiento y generalización**. Los resultados que se generaron en ambas herramientas permiten incluir nuevas posibilidades para el desarrollo futuro de un sistema experto, lo que permitirá afinarlo porque se obtuvieron de manera similar las predicciones con los datos procesados. Sin embargo, para llevarlo a la práctica, son necesarias estrategias para la prevención de la diabetes con la intervención de entidades gubernamentales y con el consenso de la sociedad. En esta etapa se utilizó Weka para el preproceso de 768 instancias o registros de pacientes, lo que puede observarse en la figura 7.



Fuente: Elaboración propia.

Figura 7 Procesado de instancias con Weka.

Como siguiente paso, se llevó a cabo una clasificación por medio del algoritmo J48, donde es posible observar que 567 instancias de pacientes fueron clasificadas de manera correcta, lo que representa un 73.82% del total de los pacientes en cuestión, así como una precisión del 79% para detección negativa de la diabetes y una precisión del 63% para detección positiva de la enfermedad. Esta información se visualiza en la figura 8.



Classifier output

```
=== Summary ===

Correctly Classified Instances         567               73.8281 %
Incorrectly Classified Instances       201               26.1719 %
Kappa statistic                          0.4164
Mean absolute error                      0.3158
Root mean squared error                  0.4463
Relative absolute error                 69.4841 %
Root relative squared error             93.6293 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.814    0.403    0.790      0.814    0.802      0.417    0.751     0.811     tested_negative
                 0.597    0.186    0.632      0.597    0.614      0.417    0.751     0.572     tested_positive
Weighted Avg.    0.738    0.327    0.735      0.738    0.736      0.417    0.751     0.727

=== Confusion Matrix ===

   a   b   <-- classified as
 407  93 |   a = tested_negative
 108 160 |   b = tested_positive
```

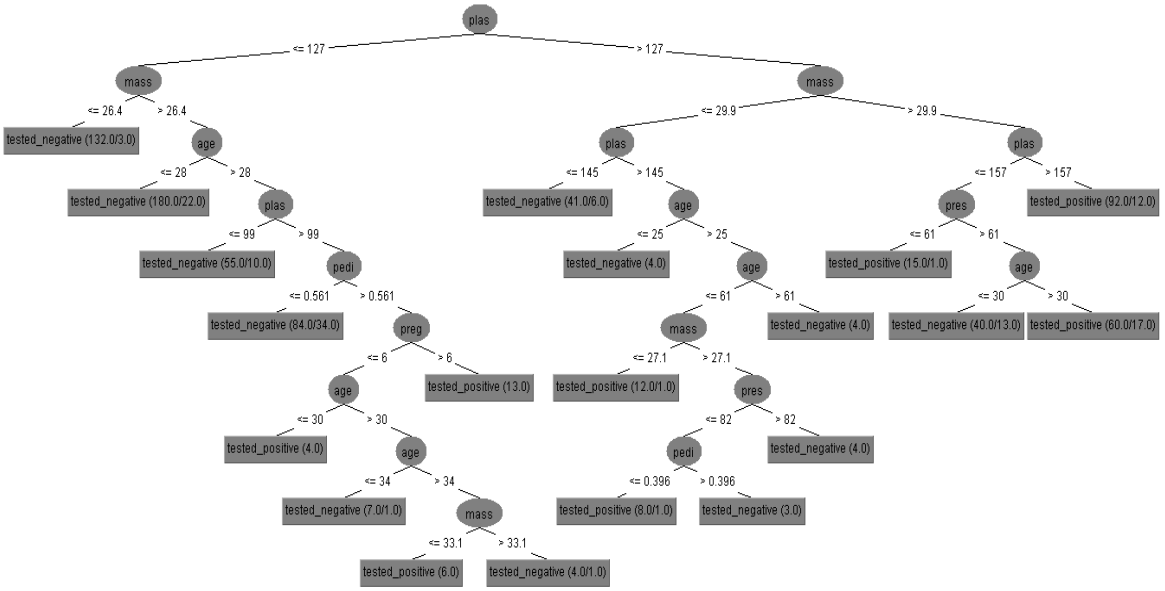
Fuente: Elaboración propia.

Figura 8 Procesado de instancias con Weka.

El árbol generado al ejecutar el algoritmo se presenta en la figura 9, donde los nodos son los atributos de cada paciente y los arcos sus reglas o dependencias para determinar si padece o no la diabetes, mismas que pueden tomarse como base para la generación de inferencias de los diagnósticos.

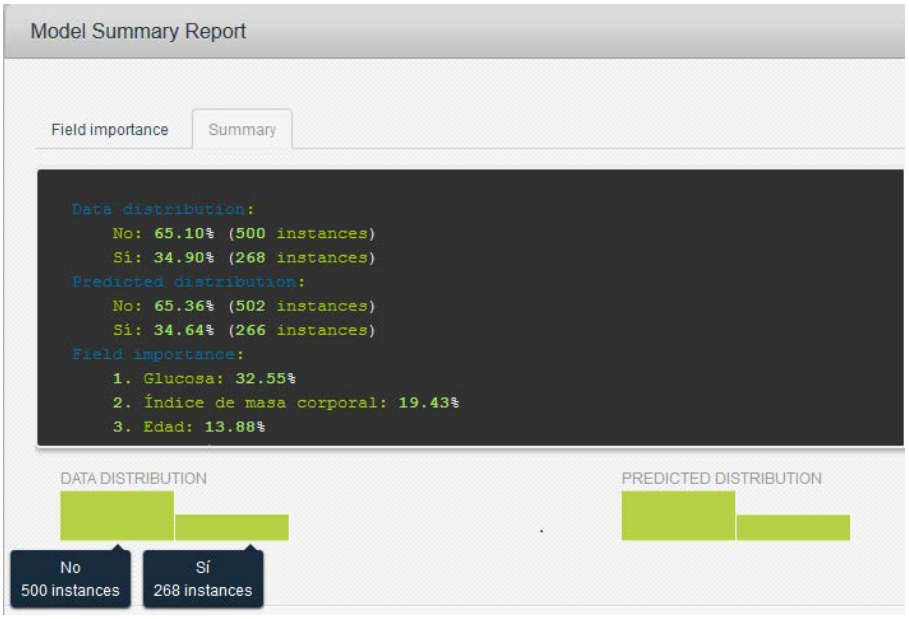
En la herramienta BigML se procesaron igual número de registros de pacientes (768), en donde la distribución de datos, 500 instancias resultaron negativos a la diabetes y 268 resultaron positivos, para el entrenamiento del sistema. En tanto la distribución de las predicciones resultaron con un "No" el 65.36% de las instancias, es decir, 502; mientras que el 34.64% (266 instancias), resultaron con una predicción "Si" como se observa en la figura

10. Además, es posible observar en la figura 11, el porcentaje de importancia de los atributos o variables de los pacientes, donde el que predomina es la variable de glucosa con 32.55% y el que menos importancia posee es la piel con 1.71%.



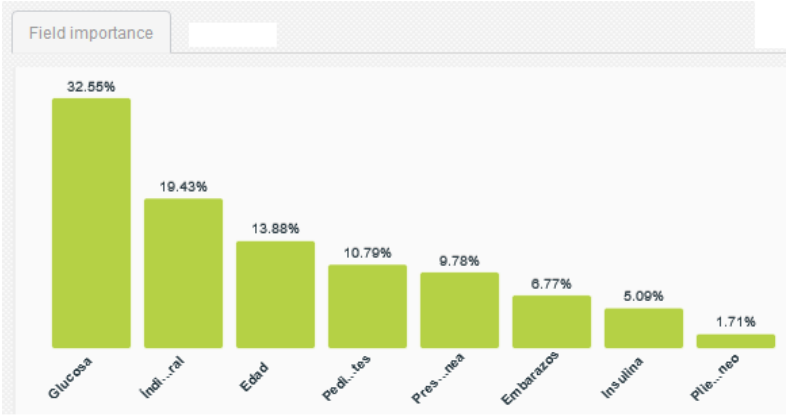
Fuente: Elaboración propia.

Figura 9 Representación de un árbol de decisión con Weka.



Fuente: Elaboración propia.

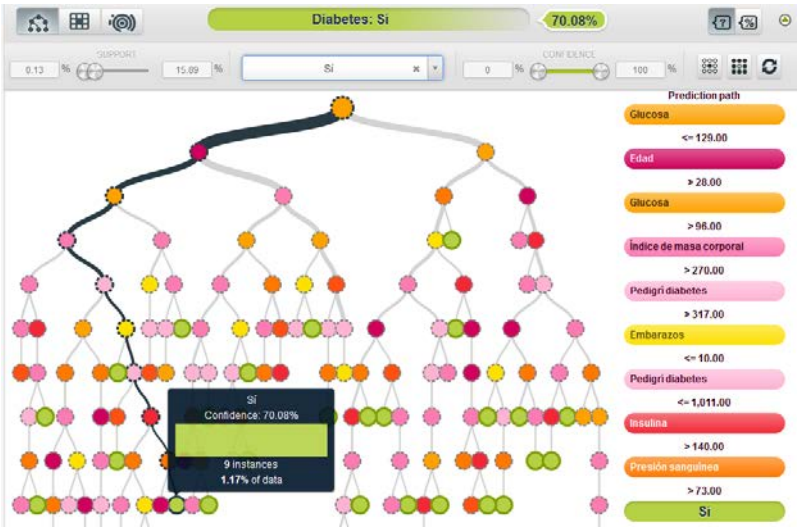
Figura 10 Resumen procesado con BigML.



Fuente: Elaboración propia.

Figura 11 Importancia de variables.

El árbol de decisión generado se puede ver en la figura 12, así como el nivel de confiabilidad de las reglas de predicción, mismo que incrementa conforme se recorre hasta llegar a un nodo u hoja objetivo, que en este es la diabetes.



Fuente: Elaboración propia.

Figura 12 Árbol de decisión obtenido con BigML.

- **Etapa 7 Mantenimiento y puesta al día**. Como siguiente paso, se pretende el desarrollo de un sistema experto con una interfaz de usuario comprensible para la ejecución del diagnóstico. Se espera la implementación de terminales de diagnóstico mediante sistemas expertos,

como un pre-diagnóstico que ayude en las comunidades rurales debido a la falta de médicos en los centros de salud.

3. Resultados

Con el conjunto de datos de diabetes se llevó a cabo un pre-procesado de la información para determinar si una persona puede padecer de manera positiva la enfermedad. Es importante recalcar que la medida de la glucosa sobresale de manera importante con el resto de las variables como medida de padecer de manera positiva la enfermedad, sin embargo, como se mencionó en la etapa 2 de la metodología, no siempre resulta confiable.

Por medio de las herramientas Weka y BigML, se generaron árboles de decisión, que consisten en un tipo de modelo predictivo donde se utiliza un grafo con estructura de árbol para la clasificación de los datos. Cada nodo del árbol simboliza una pregunta y cada rama corresponde a una respuesta concreta a dicha pregunta, es decir, un predicado. La prueba se ejecutó con 768 registros que corresponden a los datos de pacientes que han padecido o no la diabetes.

También, se aplicó el algoritmo de clasificación J48, donde más de 500 instancias de pacientes fueron clasificadas de manera correcta, lo que le permite al sistema tener una precisión por encima del 70% para la detección de pacientes que no tienen la enfermedad y una precisión del 63% para detección positiva de ésta. La matriz de confusión resultó con 93 casos de falsos positivos y 108 casos de falsos negativos. Esto es debido a que existen valores de registros como el número de embarazos, el cual en algunos casos resulta con valor cero. Por consiguiente el uso de la aplicación favorece en una medida aceptable el pronóstico de la diabetes, sin embargo, si se carece de alguna información, se considera como que no es propenso a padecer la enfermedad.

Las predicciones pueden mejorarse a través de un modelo de aprendizaje más robusto del sistema experto a través de las reglas de inferencia que se se obtuvieron en los árboles de decisión para su posterior aplicación en otros entornos de desarrollo.

4. Discusión

Al aplicar una metodología adecuada para el diseño y desarrollo de sistemas expertos se pueden conseguir objetivos de manera satisfactoria, como en el caso de la metodología Weiss y Kulikowski. Por otro lado, Weka y BigML, poseen varios algoritmos de máquinas de conocimiento, los cuales pueden ser útiles para ser aplicados sobre diversos conjuntos de datos mediante las distintas interfaces que ofrece, como la opción de Explorer y Datasets, que se trabajaron en este caso de estudio, o para ser incluidas dentro de otras aplicaciones. Además, ambas herramientas, contienen lo necesario para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización, porque están diseñadas como herramientas orientadas a la extensibilidad lo que permite añadir nuevas funcionalidades a un proyecto, debido a que se pueden combinar con otros lenguajes de programación como Prolog, para la generación de sistemas expertos más robustos.

En el trabajo de Cantón y Gibaja [2017], acerca del desarrollo de políticas de prevención de lesiones en México, mencionan que BigML simplificó la validación de anomalías y análisis de la distribución de los datos. También, fue utilizado para determinar la correlación entre variables. De manera similar en este estudio, la herramienta permitió identificar problemas de predicción. Otros autores, como Castillo, Gutiérrez y Hadi [2008], están de acuerdo en la existencia de diversos campos de aplicación de los sistemas expertos, entre los cuales se encuentran transacciones bancarias, control de tráfico, problemas de planificación, y por supuesto, diagnósticos médicos.

Como parte de los trabajos futuros, se pretende desarrollar un sistema experto que procese una nueva base de datos con información actualizada de pacientes en comunidades rurales del Sur de Yucatán. Para lograrlo, será necesario analizar a detalle cuáles fueron las principales reglas y cómo se generaron a través de Weka y BigML para su implementación correspondiente. De esta manera, se concuerda con los puntos de vista de Badaracco, Mariño y Alfonzo [2014], quienes comentan que los sistemas expertos son una de las técnicas de la Inteligencia Artificial ampliamente utilizada para la resolución de problemas comprendidos en diversos

dominios del conocimiento. Representan y explicitan el conocimiento obtenido de los sujetos utilizando diferentes mecanismos, como las reglas y las probabilidades. Además, proporcionan un marco para seleccionar acciones a seguir en situaciones complejas e inciertas, con miras a apoyar la toma de decisiones.

De esta manera, la implementación de las herramientas Weka y BigML, sólo es una muestra de la aplicación de la tecnología en ambientes de aprendizaje automático, big data, minería de datos o sistemas expertos como ramas dentro de la inteligencia artificial para apoyar en la toma de decisiones de los expertos.

5. Bibliografía y Referencias

[1] Aguilar, C. y Gómez, F. (2006). Declaración de Acapulco: propuesta para la reducción de la incidencia de la diabetes en México. Revista Investigación clínica. pp. 71-77.

[2] Badaracco, N., Mariño, S. y Alfonzo, P. (2014). Modelización de la asignación de aulas con técnicas simbólicas de la IA como ayuda a la toma de decisiones. Revista electrónica de Estudios telemáticos [Telematique]. Vol. 13.pp. 16-35.

[3] Cantón, R. & Gibaja, D. Development of Injuries Prevention Policies in Mexico: A big Data Approach (2017). International Journal of Interactive Multimedia and Artificial Intelligence. Special Issue on Big Data and e-Health. pp. 35-41.

[4] Casalboni, A. (2015). BigML: Machine Learning made easy: https://cloudacademy.com/blog/bigml-machine-learning

[5] Castillo, E. Gutiérrez, J. y Hadi, A. (2008). Sistemas expertos y modelos de redes probabilísticas.

[6] Dua, D. & Karra, E. (2017). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Recuperado de http://archive.ics.uci.edu/ml

[7] Durrant, B., Frank, E., Hunt, L. & Holmes, G. (2018). University of Waikato. Machine Learning Group: http://www.cs.waikato.ac.nz/ml/index.html.

[8] Favret, F., Eckert, K., Felten, A. & Sandberg, G. (2018). Determinación de los porcentajes de palo en la yerba mate mediante técnicas de inteligencia artificial. Iberoamerican Journal of Industrial Engineering. Vol. 10, pp. 177-198.

[9] Ferrer Y., Jiménez, K., Arguelles, D., Montes de Oca, A. (2015). Sistema experto para la elección del tipo de recuperación en canteras de materiales de construcción. Universidad de las Ciencias Informáticas. La Habana, Cuba: Ediciones Futuro. pp. 33-48.

[10] Hernández, R., Fernández, C. & Baptista, P. (2010). Metodología de la Investigación. México: Mc Graw Hill.