UNIVERSIDAD PERUANA UNIÓN

FACULTAD DE INGENIERÍA Y ARQUITECTURA

Escuela Profesional de Ingeniería de Sistemas

Una Institución Adventista

Métodos de aprendizaje supervisado para la predicción de diabetes: una revisión sistemática de la literatura

Por:

Yerry Dany Aguirre Ascona

Asesor:

Mg. Nemias Saboya Rios

Lima, Diciembre

DECLARACIÓN JURADA

DE AUTORÍA DEL TRABAJO DE

INVESTIGACIÓN

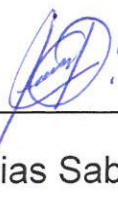
Mg. Nemias Saboya Rios, de la Facultad de Ingeniería, Escuela Profesional de Ingeniería de Sistemas, de la Universidad Peruana Unión.

DECLARO:

Que el presente trabajo de investigación titulado: "MÉTODOS DE APRENDIZAJE SUPERVISADO PARA LA PREDICCIÓN DE DIABETES: UNA REVISIÓN DE LA LITERATURA" constituye la memoria que presentan el estudiante Yerry Dany Aguirre Ascona para aspirar al grado de bachiller en Ingeniería de sistemas, cuyo trabajo de investigación ha sido realizado en la Universidad Peruana Unión bajo mi dirección.

Las opiniones y declaraciones en este trabajo de investigación son de entera responsabilidad del autor, sin comprometer a la institución.

Y estando de acuerdo, firmo la presente declaración en Lima, a los (03, diciembre) del 2019



Mg. Nemias Saboya Rios

Métodos de aprendizaje supervisado para la predicción de la diabetes: Una revisión sistemática de la literatura

Trabajo de investigación

Presentado para optar al grado de bachiller en Ingeniería de Sistemas

JURADO CALIFICADOR

Dra. Erika Inés Acuña Salinas
Presidente

Mg. Fernando M. Asín Gómez
Secretario

Mg. Leonardo Paucar Curasma
Vocal

Ing. Fredy Abel Huanca Torres
Vocal

Mg. Nemias Saboya Ríos
Asesor

Lima, 02 de diciembre del 2019

Métodos de aprendizaje supervisado para la predicción de diabetes: una revisión de la literatura

Yerry Dany Aguirre Ascona

[1] Universidad Peruana Unión, Lima, Perú
danyaguirre@upeu.edu.pe

Resumen. La inteligencia artificial (IA) y sus beneficios en el campo de la medicina han generado gran revolución. Es por este motivo que se quiere identificar los métodos de aprendizaje supervisado (una sub-área de la inteligencia artificial) y los factores empleados para la predicción de la diabetes que han sido más significativos en cuanto a técnica (de los cuales resaltan árbol de decisión y sus derivados) y resultados. Para la identificación de estos métodos se realizó una revisión sistemática de la literatura. De todos los artículos encontrados se extrajo los métodos de maching learning para considerarlos como antecedentes. Existen diversos métodos de aprendizaje supervisado que pueden predecir la diabetes en los que algunos son híbridos y otros puros, uno mejores que otros según sea el caso de estudio. Finalmente, después de una revisión de los artículos seleccionados se destaca la etapa del pre-procesamiento en el desarrollo de estos modelos para alcanzar una mayor puntuación en la precisión.

Palabras claves: inteligencia artificial, aprendizaje supervisado, predicción de diabetes.

1 Introducción

Con la aparición de DENTRAL[1], el primer sistema experto, cuyo objetivo era la inferencia de estructuras moleculares se dio a conocer el enorme beneficio que traería el investigar sobre cómo mejorar los métodos existentes o desarrollar nuevos métodos de inteligencia artificial. Por ese motivo se han aplicado al campo medico con el objetivo de aumentar el acierto en los casos de medicina [2][3], como el diagnóstico de enfermedades o la detección de células cancerígenas. Debido a esto han surgido muchos métodos de Machine Learning (de aprendizaje supervisado y no supervisado) para distintos casos clínicos. Este artículo se referirá a la diabetes.

La utilización de la inteligencia artificial es más cercana de lo que se puede pensar, particularmente en la medicina, ya se usa como apoyo para los médicos al detectar lesiones potencialmente letales en las mamografías. Además, se puede mencionar la colaboración de AliveCor y Mayo Clinic, para desarrollar una soluciona de machine learning para determinar los niveles de potasio en la sangre de una persona a través de una señal electrocardiográfica (ECG) de reloj inteligente (smartwatch). Y los gigantes tecnológicos, Google AI y su división DeepMind que han realizado un trabajo que

permite evaluar con precisión las afecciones oculares urgentes, predecir resultados en el entorno hospitalario y un importante estudio prospectivo de slides de patología en cáncer.[4]

En 1980 el número de casos con diabetes era 108 millones de personas y esto aumento a 422 millones en 2014. A nivel global los casos con diabetes han prevalecido en adultos mayores de 18 años aumentando del 4,7% en 1980 al 8,5% en 2014. En 2015, aproximadamente 1.6 millones de muertes fueron causadas directamente por la diabetes. Otros 2.2 millones de muertes fueron atribuibles a un alto nivel de glucosa en la sangre en 2012. Todos estos datos son de la organización mundial de la salud (OMS).[5]

Se encuentran tres tipos principales de diabetes: la diabetes tipo 1 o insulina dependiente, es decir el páncreas no produce suficiente insulina para contribuir al metabolismo; la segunda, la diabetes tipo 2 o diabetes mellitus o insulina independiente, es decir el problema es la utilización ineficaz de la insulina; la tercera, la diabetes gestacional, es decir hiperglucemia durante el periodo de embarazo y la posibilidad de padecer diabetes tipo 2 por parte de los hijos y de la madre.[5]

Por otra parte, el uso subjetivo de un método sobre otro puede ser contraproducente. Es decir, replicar el método sin considerar los factores con los cuales se obtuvo esa precisión con un método, la cual se señaló como el mejor para el caso probado, puede arrojar resultados que son relativamente falsos.

Esta investigación tiene como objetivo conocer los métodos y factores más significativos para la predicción de la diabetes y que herramientas se usa para implementar estas soluciones. Esto a su vez dará una mejor idea al momento de aplicar un método de aprendizaje supervisado para predecir la diabetes.

2 Revisión de la literatura

En esta sección se conceptualizan algunos términos que se usa posteriormente.

2.1 Inteligencia artificial

Para definir este término primero hay que conceptualizar inteligencia, del latín “inteligencia”. Se comprende que inteligencia no es una sino varias inteligencias, y que cada persona posee en cierto grado más de una que de otra, esta teoría propuesta por Howard Gardner en 1983, en define 7 inteligencias pero se han agregado otras hasta tener 9 tipos de inteligencias[3]: inteligencia lógica-matemática, inteligencia visual-espacial, inteligencia verbal-lingüística, inteligencia intrapersonal, inteligencia interpersonal, inteligencia corporal/cinestésica, inteligencia naturalista, inteligencia musical e inteligencia existencial o espiritual.

Debido a la amplitud de las inteligencias es difícil medirlas, aun con el test de cociente intelectual (CI), que solo es para inteligencia lógico-matemática y visual-espacial.[3]

La mejor definición seria: la inteligencia es la capacidad de adaptarse, permitiendo de este modo resolver los problemas con los que se encuentran.[3]

Entonces se dirá que la inteligencia artificial "es que la máquina de la impresión de ser inteligente cuando resuelve un problema, imitando por ejemplo el comportamiento humano o implementando estrategias más flexibles que las propia permitidas por la programación clásica" [3]. Aquí se encuentra cierta noción de la adaptabilidad lo que concuerda con la definición anterior de inteligencia.

Además, no puede faltar la observación del matemático, criptógrafo, científico de la computación y conocido como uno de los padres de la ciencia de la computación, Alan Turing. Para él, discutir el significado de las palabras "pensar" e "inteligencia" para conceptualizar las capacidades de la computadora le parecía un fastidio y un desperdicio de tiempo. Es por ello que ideo la prueba de Turing, basada en el juego de la imitación, para comprobar si una entidad era inteligente o no. La prueba de Turing en sí consiste en hacer preguntas a la entidad y conforme a las respuestas que le da, decidir si es inteligente o no; tan "simple" como sostener una conversación para identificar si conoce del tema o no. [6]

2.2 Modelos, Tareas y métodos de la inteligencia artificial

El conocimiento extraído en forma de relaciones, patrones o reglas inferidos de los datos y (previamente) desconocidos, o bien en forma de una descripción más concisa (es decir, un resumen de los mismo). Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Los modelos pueden ser predictivos y descriptivos.[7]

Las tareas son el motivo por los cuales se quiere llegar con los métodos o técnicas de inteligencia artificial (árbol de decisión, redes neuronales, bayes ingenua, lógica difusa, etc.) [7]. Por ejemplo: existe la necesidad de predecir el estado del clima; esto sería la tarea, la predicción del clima, y el método seria el cómo se va a conseguir cumplir la tarea.

Según [7], pg.139 hace mención de forma indiferente a métodos y técnicas. Por otro lado, en la pg.146 y 147 hace referencia como si las técnicas fueran una familia es decir: Técnicas bayesianas (Algoritmo EM y naive bayes). Luego en la pg.151 y 152 se identifican a las técnicas mencionadas anteriormente como métodos. Por lo tanto en este trabajo de investigación se referirá a métodos o técnicas de forma equivalente.

Además, se considera que para "una misma técnica se han desarrollado diferentes algoritmos que difieren en la forma y criterios concretos con los que se construye el modelo". [7]

2.3 Algoritmo de Machine Learning.

Machine Learning es una área de la inteligencia artificial como se puede observar en la Figura 1. Cuando se habla de maching learning, "aprendizaje de maquina", se refiere a la capacidad de las computadoras para reconocer patrones o aprender a partir de los datos ingresados[8]. Y cuando se refiere a algoritmos de maching learning se entenderá como datos de entrada, que representan cierta experiencia, que resultaran en otro aporte a la experiencia[8]. Generalmente el algoritmo esta implementado como la solución interpretable por la computadora.[3]

Un modelo de Machine Learning está implícito en el algoritmo de Machine Learning.

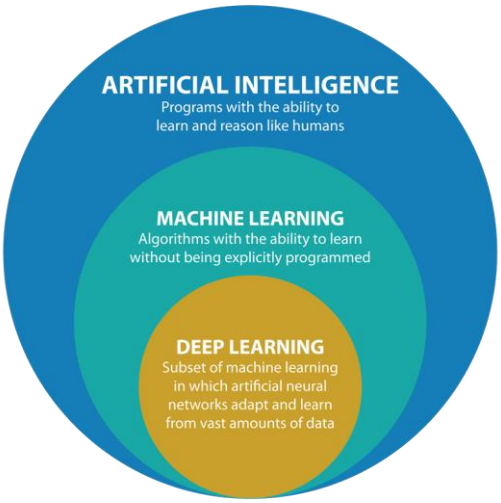


Fig. 1. Inteligencia artificial, maching learning, deep learning

2.4 Aprendizaje Supervisado.

Se entiende de aquel proceso en la que hay una o varias salidas esperadas para unos valores de entrada, realizada en la fase de entrenamiento[8][3]. Los datos de entrenamiento son pares de objeto, es decir, uno son los datos de entrada y el otro, el resultado esperado.

3 Método de la revisión sistemática de la literatura

3.1 Necesidad de la revisión sistemática

Las soluciones de Machine Learning han estado madurando con el pasar de los años como se puede ver en la Fig. 2[9] y Fig. 3[10] sobre las tendencias en soluciones de inteligencia artificial.

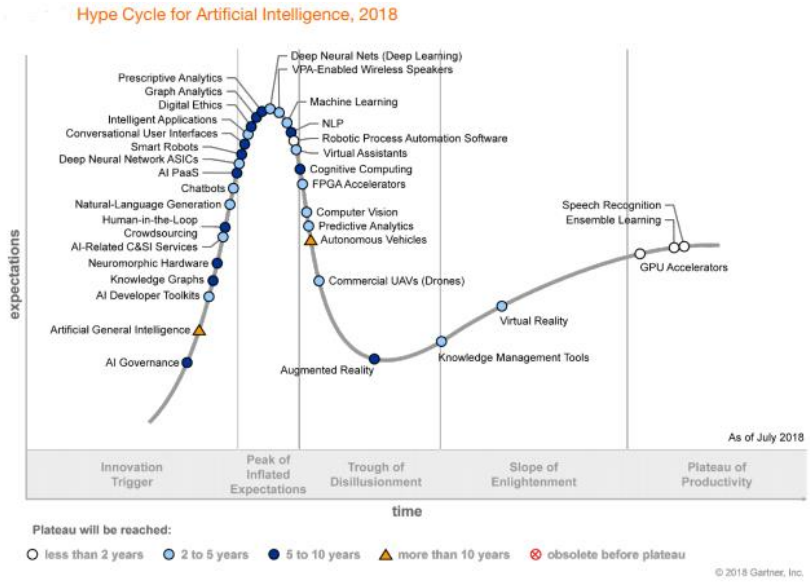


Fig. 2. Hype Cycle de Gartner para la inteligencia artificial, 2018

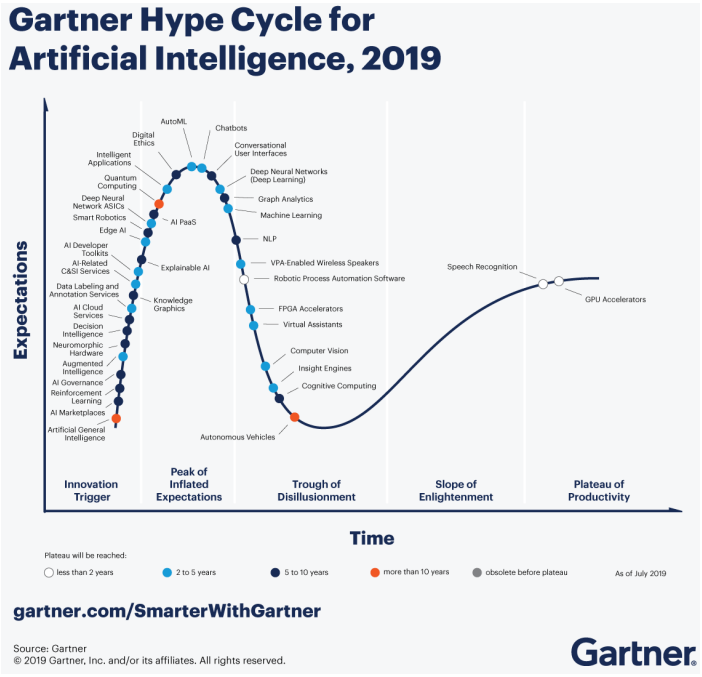


Fig. 3. Hype cycle de Gartner para la inteligencia artificial, 2019

Muchas de estas soluciones, como las aplicaciones del Deep learning (ver Figura 1 y 2), han apoyado en el campo de la medicina.

Por otra parte, el uso subjetivo de un método o algoritmo sobre otro puede ser contraproducente. Es decir, replicar el método o algoritmo sin considerar los factores con los cuales se obtuvo esa precisión con un modelo, la cual se señaló como el mejor para el caso probado, puede arrojar resultados que son relativamente falsos.

Es por eso que se planea una revisión de la literatura para determinar cuáles son los métodos de aprendizaje supervisado de Machine Learning más significativos a la predicción, en este caso la diabetes. Y se complementara con lo relacionado a los algoritmos, las herramientas y los factores de Machine Learning.

Finalmente, se requiere identificar las herramientas que se utilizan al desarrollar y/o implementar la solución de Machine Learning propuesta.

3.2 Preguntas para la revisión sistemática

Para definir y estructurar las preguntas de investigación se tomó referencia de la sección anterior. En la siguiente tabla (Tabla 1) se presenta las preguntas propuestas y la motivación de cada una.

Adicionalmente, en la Tabla 2 se presentan las preguntas de bibliometría propuestas con el objetivo de visualizar la evolución y tendencia de los estudios en el tiempo.

Tabla 1. Preguntas de investigación y motivación

ID	PREGUNTA	MOTIVACIÓN
PI-01	¿Cuáles son las los método, modelos y algoritmos de aprendizaje supervisado más significativos a la predicción de la diabetes?	Identificar las los métodos, modelos y algoritmos utilizadas durante el diagnostico predictivo de la diabetes
PI-02	¿Cuáles son los factores más significativos que se tomaron para la predicción de la diabetes en los resultados?	Identificar las los factores por las cuales se benefició la predicción de enfermedades utilizadas durante el diagnostico predictivo de la diabetes.
P1-03	¿Cuáles son las herramientas para implementar una solución de aprendizaje supervisado para predicción de la diabetes?	Identificar las herramientas (hardware, software y metodología) para implementar una solución de predicción de la diabetes.

Tabla 2. Preguntas de bibliometría

ID	PREGUNTA	MOTIVACIÓN
PB-01	¿Determinar cuál es la cantidad de publicaciones por tipo de artículo?	Precisar la cantidad de estudios publicados por tipo de artículo para poder identificar la concentración de estos.
PB-02	¿Cómo ha evolucionado a lo largo del tiempo la continuidad de las publicaciones sobre el tema?	Identificar la continuidad de las publicaciones con el objetivo de establecer la relevancia del tema en el tiempo.
PB-03	¿En qué publicaciones se han encontrado estudios relacionados con el tema?	Identificar en qué dominio de aplicación se concentra la mayor cantidad de publicaciones sobre este tema

3.3 Definición de las cadenas de búsqueda

Para la elaboración de la cadena de búsqueda se usó la estrategia PICOC a través de un proceso iterativo en la que se realizaron ajustes para la selección de los resultados. A continuación, se desglosará PICOC.

Población:

Entidad: métodos de aprendizaje supervisado

Termino principal 1: métodos

Términos alternos: técnicas

Justificante: La selección del termino fue debido al objeto de estudio de la revisión a ejecutar y se determinan los términos alternos que representan cercanos al término principal[7].

Termino principal 1: Modelos

Justificante: Se selecciona el término debido a la popularidad relativa con que se refiere a las soluciones de Machine Learning.

Termino principal 2: Algoritmos

Justificante: Se selecciona el término debido a la particularidad para los casos que son empleados. Hay cierta diferencia entre algoritmo y método.

Termino principal 3: aprendizaje supervisado

Termino alternos: inteligencia artificial

Justificante: se selecciona el término debido al tipo de análisis a ejecutar y se obtienen dichos términos alternos por ser relacionado al término principal que se quiere encontrar.

Intervención:

Entidad: predicción de la diabetes

Termino principal 1: predicción

Términos alternos: diagnóstico

Justificante: se selecciona el término debido al tipo de análisis a ejecutar y se obtienen dichos términos alternos por estar relacionados al término principal.

Entidad: diabetes

Termino principal 1: diabetes

Términos Alternos: diabetes tipo 1, diabetes tipo 2, diabetes mellitus

Justificante: se selecciona el término debido al tipo de análisis a ejecutar y se obtienen dichos términos alternos por ser los principales tipos de diabetes.

Comparación: no aplica debido que en la RSL no se hace contraste alguno.

Resultados:

Entidad: Propuesta y experiencia de predicción de software

Termino principal 1: propuesta

Términos alternos: experiencia

Justificante: se colocan dichos términos porque es lo que se busca obtener como resultado de la búsqueda.

Idioma. Se escogió el inglés como idioma para la cadena de búsqueda por su continua utilidad para la elaboración de artículos en las bases de datos de prestigio.

Usando las recomendaciones de la estrategia PICOC, se consiguió como resultado la cadena de búsqueda a partir de operadores lógicos entre los elementos definidos previamente: (Población) AND (Intervención) AND (Comparación) AND (Resultado).

En la Tabla 3 se muestra la cadena obtenida por cada uno de los elementos de la estrategia PICOC, a partir de los cuales se elabora la cadena de búsqueda.

Tabla 3. Términos en inglés y conectores lógicos que se usaran en la búsqueda

CONCEPTO	TÉRMINOS
Población	(method OR technical) AND model AND algorithm AND (supervised learning OR artificial intelligence)
Intervención	prediction AND (diabetes OR diabetes type 1 OR diabetes type 2 OR mellitus diabetes)
Comparación	no aplica
Resultado	(Proposal OR experience)
Contexto	no aplica

3.4 Criterios de inclusión y exclusión

Siguiendo con la guía elaborada por Kitchenham[11], luego de ejecutar la cadena de búsqueda en las librerías indexadas, los resultados deberán ser sometidos a evaluación con el fin de determinar los estudios primarios que responden directamente las preguntas de investigación planteadas. Se tomó en consideración los siguientes criterios para la evaluación de los estudios:

Criterios de inclusión:

CI.1. Todos aquellos artículos prove nientes de librerías digitales y fuentes indexadas serán tomadas en consideración.
CI.2. Se aceptarán artículos que contengan estudios de métodos de predicción o resultados de análisis comparativos de métodos de predicción.
CI.3. Se aceptan todos los artículos dentro del rango de temporalidad definido.
CI.4. Se consideran artículos que provienen de revistas científicas, journals, procedimientos y conferencias.
CI.5. Los artículos deben provenir del área de inteligencia artificial y relacionado.

Criterios de exclusión:

CE.1. Los artículos duplicados serán excluidos.
CE.2. Serán excluidos los artículos que no se encuentren en idioma inglés.
CE.3. Serán excluidos los artículos de contenido similar, quedándose solo los que tengan el contenido más completo.
CE.4. Se rechazan los estudios secundarios y resúmenes.
CE.5. Se rechazan los artículos cuyo título no tenga relación con el objeto de estudio.

Temporalidad. Se considera los estudios desarrollados en los últimos 5 años debido a que se requieren analizar técnicas y métodos de predicción de la diabetes que se mantengan vigentes. Además, se lo considera debido al creciente avance que hubo en Machine Learning (especialmente en aprendizaje supervisado), después de los años 70(setenta).[1]

Fuentes de Datos. Las librerías digitales indexadas tenidas en consideración debido a su relevancia científica para la selección de artículos fueron:

- IEEE Xplore (http://www.ieee.org/web/publications/xplore/)
- ACM Digital Library (https://dl.acm.org/)
- Science Direct (http://www.sciencedirect.com)
- Google Scholar (https://scholar.google.com/)

Procedimiento para la selección de artículos. Se considera el siguiente procedimiento para la selección de artículos en la RSL:

- Paso 1: se ejecutó la cadena de búsqueda PICOC, en las bases de datos indexadas que fueron seleccionadas, aplicando los criterios de inclusión y exclusión según la Tabla 4. Las referencias de los artículos resultantes fueron guardadas para su posterior refinamiento.

- Pasó 2: De los artículos encontrados en el Paso 1 se filtró por rango de temporalidad y los que pertenecían a las áreas especificadas en los criterios de inclusión de la Tabla 4 y excluyendo solamente los que no se encuentren en idioma inglés.

- Paso 3: Se aplicó los criterios de inclusión y exclusión definidos en Tabla 4.

- Paso 4: Se concluyó aplicando los criterios de inclusión y exclusión presentados en la Tabla 4.

Tabla 4. Procedimiento y criterios de inclusión y exclusión

Procedimiento	Criterios de Selección
Paso 1	ci1, ci4, ce2
Paso 2	ci1, ci3, ci5, ce2
Paso 3	ci1, ci3, ci5, ce5
Paso 4	ci2, ci5 , ce2, ce3, ce4,

3.5 Criterios de calidad

Siguiendo con los lineamientos establecidos en la guía de Kitchenham, se evalúa la calidad de los estudios seleccionados[11].

Se define una lista de criterios con el fin de comprobar el cumplimiento de cada artículo. Cada criterio se acompaña de un puntaje basado en la escala de Rouhani, él cual se explica a continuación: Sí cumple (S) = 1, Cumple parcialmente (P) = 0.5 y No cumple (N) = 0. Se presenta el esquema en la Tabla V.

Estrategia para la extracción de datos. Con la finalidad de extraer toda la información necesaria para responder las preguntas de investigación planteadas, se diseñó un formulario, Tabla 6.

Tabla 5. Criterios de evaluación de calidad

Nro.	Criterios de evaluación de calidad
1	**¿El método seleccionado para llevar a cabo el estudio ha sido documentado apropiadamente?** S: Se ha documentado apropiadamente el método seleccionado. P: Se ha documentado parcialmente el método seleccionado. N: No se ha documentado el método seleccionado.
2	**¿El estudio aborda las amenazas a la validez?** S: El estudio aborda las amenazas totalmente. P: El estudio aborda las amenazas parcialmente. N: No se detallan amenazas.
3	**¿Es clara la documentación de las limitaciones del estudio?** S: Las limitaciones se han documentado claramente. P: Las limitaciones se han documentado parcialmente. N: No se han documentado las limitaciones.
4	**¿Los aportes del estudio para las comunidades científica, académica o para la industria han sido descritos?** S: Los aportes del estudio han sido mencionados claramente. P: Los aportes del estudio han sido mencionados parcialmente. N: No se han mencionado aportes.
5	**¿Los resultados han contribuido a responder las preguntas de investigación planteada?** S: Se han respondido todas las preguntas de investigación con los resultados. P: Los resultados han contribuido a responder algunas preguntas de investigación. N: Los resultados no han contribuido a responder las preguntas de investigación.

Tabla 6. Formulario para la extracción de datos

Criterio	Detalle	Relevancia
Identificador		
Fuente		
Título		
Autores		
Publicación		
Año de publicación		
Tipo de publicación		
Tipo de análisis comparativo		
Objetivo del análisis		
Elementos comparados		
Criterios de comparación utilizados		
Dominio de aplicación		

La tabla 6 es un ejemplo de plantilla para la extracción de datos de los artículos que se encontró al final de la búsqueda, los cuales proporcionan información para resolver las preguntas de bibliometría y de investigación.

4 Resultados

4.1 Resultados de la búsqueda

El primer paso para la selección de estudios consiste en la ejecución de la cadena de búsqueda en las librerías digitales seleccionadas. En la Tabla 7 se muestran los resultados y las cadenas de búsquedas empleadas. Para la base de datos IEEE Xplorer se hizo un ajuste para mejorar la búsqueda.

Tabla 7. Resultados de búsqueda

Base de Datos	Fecha	Total
Cadena de búsqueda		
SCIENCE DIRECT	**Mayo 2019**	**3694**
(method OR technical) AND model AND algorithm AND (supervised learning OR artificial intelligence) AND prediction AND (diabetes OR diabetes type 1 OR type 2 diabetes OR mellitus diabetes) AND (proposal OR experience)		
IEEE Xplorer	**Mayo 2019**	**145**
(method OR model OR algorithm) AND (supervised learning OR artificial intelligence) AND prediction AND (diabetes OR diabetes type 1 OR diabetes type 2 OR mellitus diabetes)		
ACM Digital Library	**Mayo 2019**	**1633**

(method OR technical) AND model AND algorithm AND (supervised learning OR artificial intelligence) AND prediction AND (diabetes OR diabetes type 1 OR type 2 diabetes OR mellitus diabetes)		
Google Scholar	**Mayo 2019**	**16900**
(method OR technical) AND model AND algorithm AND (supervised learning OR artificial intelligence) AND prediction AND (diabetes OR diabetes type 1 OR type 2 diabetes OR mellitus diabetes)		

4.2 Resultados de filtros aplicados

Selección de estudios primarios.
Seguidamente, se presenta el detalle de los pasos realizados para la selección de estudios:

Paso 1: Se procedió a ejecutar la cadena de búsqueda y en cuanto al idioma se limitó a los escritos en inglés. La librería indexada que produjo la mayor cantidad de resultados fue Google Scholar. Asimismo, sobre dicha lista fueron aplicados los criterios de inclusión y exclusión presentados para este paso según la Tabla 4.

Paso 2: Sobre la lista de resultados del Paso 1 se procedió con la exclusión e inclusión de los artículos para el objeto de estudio de acuerdo a lo definido en los criterios descritos según la Tabla 4.

Paso 3: Los artículos provenientes del Paso 2, fueron excluidos por no coincidir con el título definido en los criterios descritos y se aplicaron los criterios de inclusión según la Tabla 4.

Paso 4: Para proceder con la descarga de los artículos se revisó el contenido de los artículos restantes tomando en consideración el resumen, la introducción y las conclusiones, y fueron excluidos los que no tenían relevancia de acuerdo a lo definido en los criterios presentados según la Tabla 4.

En la Tabla 8 se muestran los resultados que se obtuvo de la selección de artículos.

Tabla 8. Resultados del proceso de selección de estudios

Base de Datos	Artículos Descubiertos	Paso1	Paso 2	Paso 3	Paso 4
Science Direct	3694	2303	298	8	2
IEEE Xplore	145	145	103	47	21
ACM digital library	1633	193	128	4	2
Google scholar	16900	14300	8640	27	3
Total	**22372**	**16941**	**9169**	**86**	**28**

4.3 Preguntas Bibliométricas

¿Cuál es la cantidad de publicaciones por tipo de artículo?

En la Fig. 2 se muestra la cantidad de publicaciones por tipo de artículo. Se observa que los artículos de conferencia (Conference) representan el 67.9% del total de artículos seleccionados; seguidamente los artículos en revista (Jornal) con un 17.9%; además los artículos de investigación (research articles) reflejan el 7.1% del total de los artículos seleccionados; y por último se puede observar a los artículos provenientes de procedimientos (Proceeding) representan el 7.1% del total. De este análisis se puede concluir que los artículos de conferencias y de investigación son la mayor fuente para esta RSL.

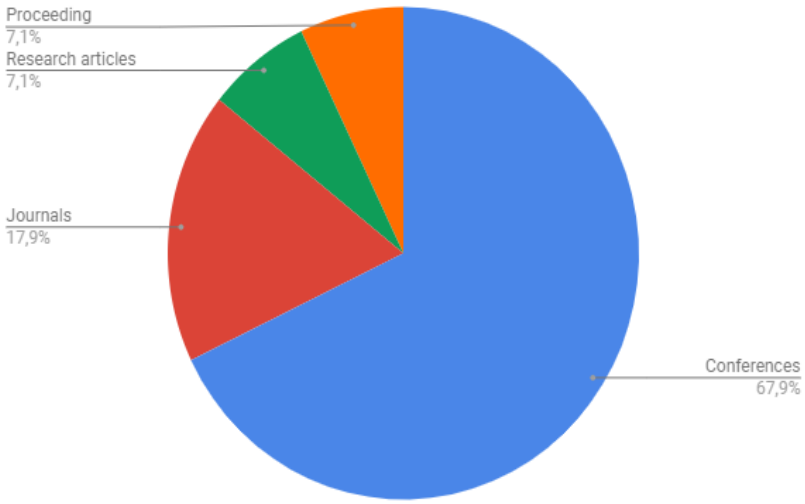


Fig. 4. Cantidad de publicaciones por tipo. Elaboración del autor

¿Cómo ha evolucionado en el tiempo la frecuencia de las publicaciones sobre este tema?

Se puede observar en la Fig. 3 un decremento en el número de publicaciones que describen algunos temas relacionados al estudio en esta RSL a partir del año 2014 y en adelante. De un total de 32 artículos, 11 (39.11%) han sido publicados a lo largo de los últimos 3 años y 17 (60.71%) han sido publicados entre 2014 y 2016. Aunque se ve un crecimiento a partir del año 2018.

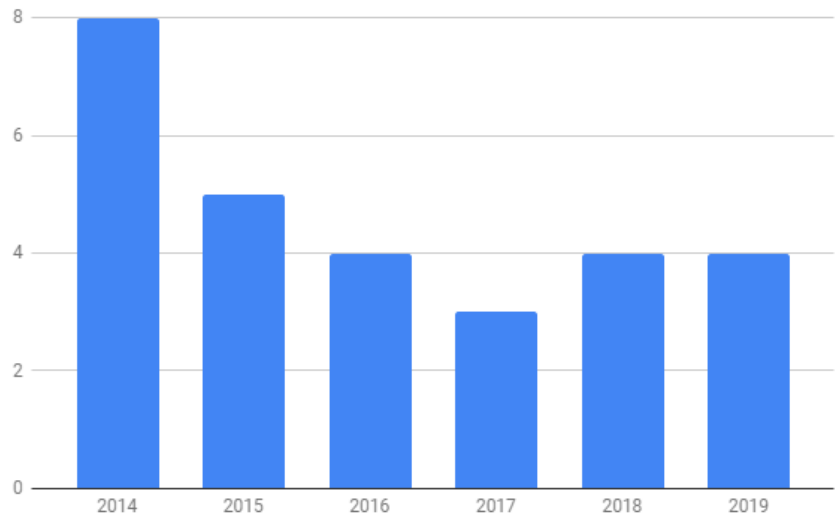


Fig. 5. Frecuencia de publicaciones. Elaboración del autor.

¿Cuáles son las publicaciones en las que se han encontrado estudios relacionados al tema?

En la Tabla 9 se presentan las publicaciones de donde se han extraído los artículos seleccionados. A partir de este análisis se puede observar que existe una recurrencia de publicaciones del dominio IEEE Journal of Biomedical and Health Informatics. Siendo estos dominios de aplicación los que concentran la mayoría de los artículos elegidos. Adicionalmente, también se puede observar la presencia de otros dominios tales como: Artificial Intelligence in Medicine, Journal of Biomedical Informatics, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), ICMLC '19 Proceedings of the 2019 11th International Conference on Machine Learning and Computing, ICMLSC 2019 Proceedings of the 3rd International Conference on Machine Learning and Soft, Computer and Communication Engineering (ECCE), entre otros.

Table IX. Publicaciones correspondientes a los artículos seleccionados

Título de Publicación	Cant.
Artificial Intelligence in Medicine	1
Journal of Biomedical Informatics	1
IEEE Journal of Biomedical and Health Informatics	2
Sixth International Symposium on Embedded Computing and System Design (ISED)	1
8th IEEE International Conference on Software Engineering and Service Science (ICSESS)	1
International Conference on Advances in Computer Engineering and Applications	1
International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)	1
1st International Conference on Next Generation Computing Technologies (NGCT)	1
Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)	1
IEEE International Advance Computing Conference (IACC)	1
International Conference on Information Technology Systems and Innovation (ICITSI)	1
Amity International Conference on Artificial Intelligence (AICAI)	1
International Conference on Electrical, Computer and Communication Engineering (ECCE)	1
IEEE International Conference on Bioinformatics and Biomedicine (BIBM)	1
IEEE International Conference on Bioinformatics and Bioengineering	1
International Conference on Advances in Computing and Communication Engineering (ICACCE)	1
IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)	1
IEEE Congress on Evolutionary Computation (CEC)	1
12th International Conference on Frontiers of Information Technology	1
21st International Conference of Computer and Information Technology (ICCIT)	1
4th Mediterranean Conference on Embedded Computing (MECO)	1
ICMLC '19 Proceedings of the 2019 11th International Conference on Ma-	1

chine Learning and Computing	
ICMLSC 2019 Proceedings of the 3rd International Conference on Machine Learning and Soft Computing	1
MOBILESoft '16 Proceedings of the International Conference on Mobile Software Engineering and Systems	1
Computers & Electrical Engineering	1
AI & SOCIETY	1
Medical Decision Making	1

4.4 Preguntas de investigación

En la tabla 11 se muestra el resumen de los artículos (no todos) encontrados detallando el método y la precisión de cada uno de ellos. Considerar que este resumen no es una comparación entre estos artículos debido a que no todos ellos han usado la misma base de datos o están constituidos con las mismas variables exógenas. Pero si se puede tomar en consideración la precisión que hay entre los métodos de los mismos artículos y los métodos que las conforman.

La tabla 10 especifica las variables de uno de los conjuntos de datos más utilizados. Siendo de la Universidad de California, Irvine (UCI) de Machine Learning repository. El conjunto de datos contiene datos de pacientes con diabetes y sujetos normales.[12]

El conjunto de datos contiene información de examen de los signos vitales de participantes de la herencia indios Pima obtenido durante un período de cinco años durante un programa de control de salud similar al NHS Health Check.[13]

Tabla 10. Variables de la fuente de data Pima indias diabetes

Pima Indians Diabetes
Número de veces embarazadas
La glucosa en plasma concentración a 2 horas en una prueba de tolerancia oral a la glucosa.
La presión arterial diastólica (mm Hg)
Tríceps espesor del pliegue de la piel (mm)
2 horas suero de insulina (mu U / ml)
Índice de masa corporal
función de la diabetes pedigrí
Edad

Tabla 11. Resumen de los métodos encontrados en los artículos seleccionados a criterio

Articulo	Técnica	Precisión	Variables	Fuente de data	Pre-procesamiento
Art01 [14]	- Naïve Bayes - Regresion Logistica	- NB = 0,653, LR =0,661 - NB = 0,73, LR = 0,735	- Cintura-cadera + TG en los hombres - Costilla-cadera + TG en mujeres	Korean Health and Genome Epidemiology Study database	No se aplico
Art02 [15]	K-neigborth nest	100%	Ver Tabla 10	Pima Indian Diabetes	-Limpieza de datos -Reducción de datos
Art03 [16]	- Regresion Logistica(LR) - Naïve Bayes(NB)	- 0,741 y 0,739 en las mujeras. - 0.687 y 0,686 en los hombres.	- AGE, HEIGHT, WEIGHT, FC, NC, RFcR, ANcR, and WRcR (mujeres) - AGE, HEIGHT, NC, AC, NFcR, RFcR, NHcR, RAcR, WAcR, WCcR, CHcR, HRcR, and HWcR (hombres)	Korean Health and Genome Epidemiology Study database	técnica minoría sintético sobre muestreo (SMOTE)
Art04 [17]	Árbol de decisión J48	90.04%	Ver Tabla 10	Pima Indian Diabetes	- Reemplazar los valores perdidos y valores imposibles con la media. - Usamos K-means para eliminar las muestras incorrectamente clasificadas
Art05 [18]	Levenberg-Marquardt algoritmo	0,71	Ver Tabla 10	Pima Indian Diabetes	No se aplico
Art06 [19]	- Neuro-Fuzzy (ANFIS) - Algoritmo de backpropagation Levenberg-Marquard	- 90,32 - 71,10	- Edad - IMC - Presion arterial diastolica - Diabetes Pediggri - Concentracion de plasma de glucosa	Bhubaneswar, Odisha, India	No se aplico
Art07 [20]	- Random forest - Bayes ingenuo - Algoritmo ID3 - Algoritmo AdaBoost	- 85% - 79.89% - 78.57% - 84.19%	- Edad - Altura - Peso - Cintura - Cadera La circunferencia de cintura (CC) o cintura-cadera (WHR) discriminan mejor los casos de diabetes entre	Conjuntos de datos de la Universidad de Virginia	- Todas las funciones no relacionadas son eliminados, incluyendo el colesterol total, lipoproteína de alta densidad, glucosa estabilizada, lipoproteína de alta densidad, el colesterol / HDL

			los que no tienen, en comparación con IMC.		ratio y la presión arterial sistólica primer lugar. - Se han eliminado los valores faltantes - Desratización
Art08 [21]	- J48 - Naïve Bayes - SVM con Polykernel - SVM con RBFkernel - Perceptrón multicapa	Puntaje AUC - 0.928 - 0.915) - 0.942 - 0.827 - 0.911	- Edad, sexo, polidipsia, polifagia, poliuria, los antecedentes familiares, alta presión arterial, la dieta, - La actividad física, la visión borrosa, el hábito de fumar, pérdida de peso, altura y peso (IMC), circunferencia de la cintura	Datos reales recogido de un hospital de renombre en el estado de Chhattisgarh de la India	Los valores faltantes se remplazan con la mediana
Art09 [22]	- Ensemble Perceptrón Algoritmo (EPA) - algoritmo de Perceptron (PA)	- 0.75 - 0.72	- Edad - IMC	Encuesta Nacional de Salud y Nutrición (NHANES) de los Estados Unidos	No se aplico
Art10 [23]	- AdaBoost algorithm with decision stump - Machine support vector - Naive Bayes - Árbol de decisión	- 80.72% - 79.687% - 79.687% - 77.6%	- Ver Tabla 10 - Tríceps espesor del pliegue de la piel *(en la data local se obtiene de forma indirecta)* - 2 horas suero de insulina *(en la data local se obtiene de forma indirecta)* - Indice de masa corporal *(en la data para validacion se obtuvo con altura y peso)* - función de la diabetes pedigrí *(en la data local se obtiene de forma indirecta)*	- Pima Indian Diabetes - Data local	Los valores perdidos se sustituyen con los atributos de valor media correspondientes en el conjunto de datos global
Art11 [24]	Extreme learning machine(ELM) BackPropagation	- 0.5964 - 0.0575	Ver Tabla 10	Pima Indian Diabetes	Se normaliza para que tengan un cierto rango de valores
Art12 [25]	- Backpropagation - Arbol de desicion J48 - Bayes Ingenuo - Vector machines	- 83.11 - 78.26 - 78.97 - 81.69	Ver Tabla 10	Pima Indian Diabetes	- Técnica de normalización Min-max - Seleccion de caracteristicas con chi-cuadrado

Art13 [26]	- Vector Machine (SVM) - Naive Bayes (NB) - K-vecinos más cercanos (KNN) - Árbol de decisión C4.5 (DT)	- 0.65 - 0.685 - 0.708 - 0.72	- Edad, Sexo, Peso, Dieta, Poliuria, Consumo de agua - Sed excesiva, Presion sanguiniea, Hipertension - Cansancio, Problema en la vision, Problema en el riñon - Perdida de audicion, Picazon en la piel, Genetica	Centro Médico de Chittagong (MCC), Bangladesh	El valor numérico exacto de los atributos no es significativo para predecir la diabetes. Como tal, convertimos los valores de atributos numéricos en nominal
Art14 [27]	- Random forest - Vector Máquinas (SVM) - Regresión logística (LR)	- 0.89 - 0.825 - 0.844	- Snps - IMC - Edad	Instituto de Investigación Biomedica de Girona	SNPs mas relevantes
Art16 [28]	Red neuronal feedforward utilizando el modelo de ventana	Error cuadrático medio 1.26 ml/d	niveles de glucosa monitoreado	AIDA, simulador matemático de la diabetes	No se aplico
Art17 [29]	Árbol de decisión difuso basado en índice de GINI	75.8%	Ver Tabla 10	Pima Indian Diabetes	Se eliminaron los registros con datos perdidos

A. ¿Cuáles son los métodos, modelos y algoritmos de aprendizaje supervisado más significativos a la predicción de la diabetes?

Los modelos más usados en este trabajo de revisión fueron el árbol de decisión [17] y sus derivados como el bosque aleatorio[27] y demás([17], [21]). Esto puede ser debido a la estimación de importancia de las variables[30], a la gran capacidad de manejar grandes datos y la estabilidad ante datos faltantes

De los artículos presentados en la Tabla 5, considerando que no hicieron pre-´procesamiento, se observa que [19] alcanzo 90.32 con un modelo hibrido, Neuro-Fuzzy (ANFIS). Además, [22] también presento un puntaje de 0.75 a diferencias de las otras metodologías que no hicieron pre-procesamiento. No se pude confirmar que una es mejor que otra debido a las distintas variables que usaron los modelos, pero esto podría dar un indicio.

En [31] se describe una técnica que utiliza de forma conjunta tres algoritmos de árbol de decisión (ID3, C4.5 and CART), con el objetivo de mejorar la precisión individual, que se combinan mediante Bagging (seleccionado entre otros métodos de combinación) la cual consiste en la votación por mayoría de los modelos generados para la clasificación. También se utilizó Stratified sampling para solucionar el desequilibrio de clases. Se probó esta técnica hibrida con dos conjuntos de datos, y dio un buen

resultado para ambos conjuntos de datos. En comparación con los otros método de combinación bagging dio los mejores resultados los cuales fueron para la precisión, la sensibilidad, la especificidad y f-medida 91.56%, 95.63%, 68.33% y 79.71%, respectivamente.

Cuando se junta la calidad del conjunto de datos y un modelo hibrido se puede obtener lo que [32] obtuvo. Se usó un hibrido de SVM y Naive Bayes en la cual se obtuvo un 97.6%, en la que ambos algoritmos tienen que coincidir en sus predicciones, de ser diferentes se dará un mayor monitoreo (según explica el articulo). El conjunto de datos es propia obtenida de Kosovo. Después de adquirir los datos iniciales de los pacientes, y después de los extensos exámenes de laboratorio y monitoreo continuo, como especifica el artículo.

La propuesta de The least angle regression(LARS) con PCA descrita en [33] es interesante, debido a que implementa una normalización de los datos y una selección de variables con PCA de forma implícita, todo esto dio un 89.53% de área bajo la curva ROC (AUC), que es una proporción entre sensibilidad y especificidad.

B. ¿Cuáles son los factores más significativos que se tomaron para la predicción de la diabetes en los resultados?

Uno de los factores son las variables empleadas en la elaboración del modelo de aprendizaje supervisado, se puede mencionar en el [16] y [27], en las cuales se demuestra que el uso de combinaciones de variables o medidas aumenta la precisión del modelo. En ello se consideró las variables que más aportaban, mediante técnicas como las propias del bosque aleatorio[27] y los propuestos por la técnica SMOTE[16], a la tarea de predicción, a partir de las cuales se introdujeron a la elaboración del modelo.

La calidad del conjunto de datos es fundamental para trabajar con Machine Learning, es por ello que en [34] se obtuvo 87.5% de precisión con el modelo random forest en la que al conjunto de datos no se aplicó ninguna técnica de preprocesamiento posiblemente debido a la ausencia de defecto alguno sobre el conjunto de datos (el artículo no hace mención sobre algún defecto). Es posible que sea el caso para [35]

Otro de los factores es la desigualdad o desequilibrio de los datos que se usaran para la elaboración del modelo, esto significa que los registros usados para el entrenamiento tienen una predisposición a una clase en particular, en otras palabras, hay considerablemente más registros de una clase en particular. Esto se puede resolver usando la técnica minoría sintético sobre-muestreo (SMOTE).[16][13]

Otra evidencia de SMOTE es en [36] donde se usó el conjunto de datos Pima Indias Diabetes del cual derivaron dos conjuntos de datos. Al primer conjunto se le aplico la eliminación de valores perdidos (missing value) y SMOTE para el sobre muestreo. Al segundo conjunto de datos se le aplicó una selección de características

complejas (5 algoritmos). Se hicieron las pruebas con ambos grupos de datos, y el segundo grupo de datos tuvo un puntaje relativamente mejor, es decir no mejoro en casi nada a los del primer conjunto de datos. De esta manera se demuestra que SMOTE puede dar un gran aporte sin mucho esfuerzo.

Además, en [37] se muestra una cualidad mas de SMTE donde se mejoro los puntajes obtenidos por árbol de decisión, red neuronal probabilístico(PNN) y naive bayes, mostraran mejoras de 64%, 51% y 5% respectivamente, cuando se le aplico SMOTE. Se puede mencionar que para el árbol de decisión que obtuvo 0.215(sensibilidad), 0.992(especificidad) y 0.336 (f-measure), después de aplicar SMOTE se obtuvo 0.726(sensibilidad), 0.802(especificidad) y 0.436(f-measure). El conjunto de datos es de Tehran Lipid Estudy and Glucose (TLG).

En [38] también utiliza una técnica de sobre-muestreo aunque no especifica cual. Con la cual random forest obtuvo una precisión de 84% para el conjunto de datos Pima Indians Diabetes.

Con respecto a artículo [15] se muestra que después de aplicar tres técnicas de pre-procesamiento (limpieza de datos, reducción de muestras y PCA), siendo PCA la técnica que reduzco de ocho variables a dos, la cual permitió que el algoritmo k-vecinos más cercanos alcanzara una precisión de 100% en la validación cruzada como en las validaciones convencionales. Cabe recalcar que después de la limpieza de datos y la reducción de muestras el dataset usando, pima indians diabetes (ver Tabla 10), se redujo de 768 a 696 muestras. Para ver el rendimiento de la metodología propuesta por el artículo ver la tabla 12, en la cual D1 significa que al dataset no se le aplica ninguna técnica de pre-procesamiento, D2 es el dataset después de aplicar la limpieza de datos y D3 significa el dataset luego de aplicar la reducción de muestras. Además de ello se utilizó la validación de AUC la cual dio 1, lo que indica el 100%.

Tabla 12. Rendimiento de la metodología propuesta por [15]

Without Principle Component Analysis			
Evaluation technique	Classification Accuracy		
	D1	D2	D3
Conventional	100%	100%	100%
Cross-validation	60.5%	60.5%	100%
With Principle Component Analysis Conventional			
Conventional	100%	100%	100%
Cross-validation	58.3%	60.4%	100%

Con respecto al artículo [17] se observa que la aplicación de técnicas de pre-procesamiento como reemplazar valores perdidos por los valores media del atributo y eliminar las muestras mal clasificadas por el algoritmo K-means influyeron en la puntuación alcanzada. Sin embargo, es dudoso el criterio de haber eliminado por K-means algunas de las muestras, esto estaría contradiciendo la clasificación de la misma data, y presumiría que K-means seria suficiente para decidir cual muestra fue

erróneamente clasificada con lo que se vería en vano la elaboración del modelo J48 propuesto en dicho artículo.

Con respecto al artículo [21] donde se utiliza una técnica pre-procesamiento para remplazar los valores perdidos con la media de atributo para obtener la puntuación de 92.8% de AUC, a diferencias de los otros modelos probados por el artículo, de un conjunto de datos de 145 muestras. Con esto se entiende que J48 es buen clasificador incluso con pocos datos.

Considerando [28] en la que se usó un conjunto de datos proveniente del simulador de AIDA; en la cual se puede simular hasta 40 casos de estudio con diferentes grupos de edad, de enfermedades, y la ingesta de comida; que alcanzo el error cuadrático medio de 1.26 ml/d, siendo el promedio de los 10 casos probados en el artículo, es posible que por tratarse de los datos simulados no haya sido necesario de un pre-procesamiento.

Al comparar el ligero pre-procesamiento que se hizo en la [29], con los [17], [23] y [25] (por ejemplo), se puede observar que estos otros alcanzaron un puntaje superior siendo el modelo y el origen del conjunto de datos iguales.

Sin embargo, cabe mencionar a [39] donde se realizó una prueba de chi-cuadrado para validar la dependencia de las variables predictores que se usa, aunque no se aplicaron los resultados de dicha distribución, además se realizó una limpieza del conjunto de datos que tenían para su posterior clasificación con árbol de decisión con la que se obtuvo un puntaje de 75%. Es posible que la razón del puntaje bajo sea la deficiencia en la limpieza de los datos o que una sola técnica de pre-procesamiento no sea suficiente para el conjunto de datos que tienen. Con lo que se puede decir, que por el hecho de aplicar una técnica de pre-procesamiento no asegura la optimización de la clasificación; hay que aplicar las técnicas necesarias y hacerlo bien.

Es posible que [40] y su 82.35% de precisión sea engañoso, porque aunque se haya utilizado Min Max Scaling como técnica de normalización(que transforma los valores entre 0 a 1, para tener una distribución normal) no se ha solucionado el problema de desequilibrio que presenta el conjunto de datos Pima Indians diabetes que utilizo. Solo se especifica como métrica la precisión mas no la especificidad ni la sensibilidad u otros.

C. ¿Cuáles son las herramientas para implementar una solución de aprendizaje supervisado para predicción de la diabetes?

El uso del software WEKA para evaluar la efectividad del modelo propuesto[41] y dos de sus propuestas están programados en el leguaje C++, de los otros no hay mención. El [42] menciona la gran cantidad de métodos que posee la herramienta e indica la personalización de acuerdo a lo requisitos del estudio como una de sus mayores ventajas.

Para aplicación de sistemas de información web se valieron de tensorflow.js para la implementación del modelo de Machine Learning [40].

Para aplicar el estudio [25] desarrollaron los modelos en restudio con el lenguaje de programación R. Por otro lado, los experimentos de [30] se realizan con el lenguaje de programación R 3.2.1 y se usaron los paquetes para random forest y SVM.

En [16] se utiliza el software SPSS para el análisis de los resultados. Por otra parte en [25] se usa software RStudio con el lenguaje R para implementarlo.

Una herramienta basada en la nube de Microsoft es el Azure Machine Learning Studio(AMLS), que uso [43] para desarrollar un modelo de árbol de decisión, la cual cuenta con una interfaz gráfica de usuario para construir y poner en funcionamiento modelos de Machine Learning. Se facilita el trabajo con una serie de funciones de arrastrar y soltar en la interfaz. Y el despliegue se puede hacer con unos pocos clics.

5 Conclusiones

Para recopilar las diferentes fuentes, debidamente filtradas, que estén relacionadas a esta investigación fue necesaria la RSL, además de servir el método PICO para el planteamiento de las preguntas de investigación y de la literatura.

Cabe mencionar que utilizar una técnica hibrida como [41] que consiste en electromagnetism-like mechanism algorithm (EM) usando prueba de signo opuesto (OST) de tipo ROST combinado con 1NN, con un conjunto de datos sin valores faltantes pero con desequilibrio de datos; puede provocar un valor engañoso en la precisión. Es por eso que el resultado de la precisión es 73.03% mientras que el índice kappa es de 0.0389 (siendo el índice kappa de 0 a 1, donde 0.00 – 0.20 significa ínfima concordancia). Mostrando de esta manera discordancia.

Otro ejemplo es [12] que usa un modelo hibrido (SVM, ANN y Bayes Ingenuo) que obtuvo un puntaje de 58.3%(especificidad), 86.8%(sensibilidad) y 77%(exactitud) con el conjunto de datos Pima Indians Diabetes en el que se hizo una selección de las variables predictores con el software WEKA (el artículo no especifica el proceder). Este modelo hibrido pudo haber tenido un mayor puntaje si se hubiera considerado el desequilibrio de los datos, como se hizo en otros artículos.

Uno de los métodos más significativos de aprendizaje supervisado, de acuerdo a esta investigación, fue el árbol de decisión y sus derivados.

Para plantear un modelo de aprendizaje supervisado de predicción de la diabetes ya sea puro o hibrido se debe considerar como factores por sobre todo el pre-

procesamiento y las variables predictoras con las que se trabajara, esto involucra tener un poco la ayuda del experto de tema para orientar a las técnicas a aplicar.

La precisión no solo recae en el modelo que se use sino también en las variables con las que se trabaje por lo que se tendrá que determinar las variables más influyentes que aumente la precisión del modelo.

Entre las herramientas usadas para el aprendizaje supervisado, la que más resalto fue WEKA, debido al conjunto de herramientas que tiene para trabajar con machine learning y a su fácil uso de ellas.

Considerar que mientras más datos o casos se tengan el modelo de aprendizaje tendrá mayor precisión.

Una oportunidad para esta investigación es la implementación de métodos de árboles de decisión y sus derivados a las variables que indica [14] considerando las técnicas de pre-procesamiento que requiere el conjunto de datos seleccionado.

Referencias

[1] M. Casella, *Historia y evolución de la Inteligencia Artificial*. Marco Casella, 2015.

[2] F. H. Maldonado, *Procedimientos de inteligencia artificial en el estudio de las enfermedades infecciosas*. Díaz de Santos, 1999.

[3] V. Mathivet, *Inteligencia Artificial para desarrolladores Conceptos e implementación en C# (2a edición)*. Ediciones Eni, 2018.

[4] Anahad O'Connor, "How Artificial Intelligence Could Transform Medicine - The New York Times," 2019. [Online]. Available: https://www.nytimes.com/2019/03/11/well/live/how-artificial-intelligence-could-transform-medicine.html. [Accessed: 25-Jun-2019].

[5] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030.," *PLoS Med.*, vol. 3, no. 11, p. e442, 2006.

[6] J. Haugeland and I. T. de Firmani, *La inteligencia artificial*. Siglo XXI, 2001.

[7] J. H. Orallo, M. J. R. Quintana, and C. F. Ramírez, *Introducción a la minería de datos*. Pearson Educación, 2004.

[8] I. Freeman, A. Haigler, S. Schmeelk, L. Ellrodt, and T. Fields, "What are they Researching? Examining Industry-Based Doctoral Dissertation Research through the Lens of Machine Learning," *Proc. - 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2018*, pp. 1338–1340, 2019.

[9] Laurence Goasduff, "Artificial intelligence trends." [Online]. Available: https://www.gartner.com/smarterwithgartner/top-trends-on-the-gartner-hype-cycle-for-artificial-intelligence-2019/. [Accessed: 09-Oct-2019].

[10] S. Sicular and K. Brant, "Hype Cycle for Artificial Intelligence," *Gartner*, no.

Juli, pp. 5–6, 45–47, 2018.

[11] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering - A systematic literature review,” *Inf. Softw. Technol.*, vol. 51, no. 1, pp. 7–15, 2009.

[12] L. Li, “Diagnosis of diabetes using a weight-adjusted voting approach,” *Proc. - IEEE 14th Int. Conf. Bioinforma. Bioeng. BIBE 2014*, pp. 320–324, 2014.

[13] N. A. Nnamoko, F. N. Arshad, D. England, and J. Vora, “Meta-classification model for diabetes onset forecast: A proof of concept,” *Proc. - 2014 IEEE Int. Conf. Bioinforma. Biomed. IEEE BIBM 2014*, pp. 50–56, 2014.

[14] B. J. Lee and J. Y. Kim, “Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on Machine Learning,” *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 1, pp. 39–46, 2016.

[15] M. Panwar *et al.*, “K-nearest neighbor based methodology for accurate diagnosis of diabetes mellitus,” *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 1, pp. 1–5, 2017.

[16] B. J. Lee, B. Ku, J. Nam, D. D. Pham, and J. Y. Kim, “Prediction of fasting plasma glucose status using anthropometric measures for diagnosing Type 2 diabetes,” *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 2, pp. 555–561, 2014.

[17] W. Chen, S. Chen, H. Zhang, and T. Wu, “A hybrid prediction model for type 2 diabetes using K-means and decision tree,” *Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS*, vol. 2017-Novem, no. 61272399, pp. 386–390, 2018.

[18] S. A. Saji and K. Balachandran, “Performance analysis of training algorithms of multilayer perceptrons in diabetes prediction,” *Conf. Proceeding - 2015 Int. Conf. Adv. Comput. Eng. Appl. ICACEA 2015*, pp. 201–206, 2015.

[19] A. Swain and A. Sample, “Comparative Risk Analysis on Prediction of Diabetes Mellitus Using Machine,” pp. 3312–3317, 2016.

[20] W. Xu, J. Zhang, Q. Zhang, and X. Wei, “Risk prediction of type II diabetes based on random forest model,” *Proc. 3rd IEEE Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics, AEEICB 2017*, pp. 382–386, 2017.

[21] K. Sowjanya, A. Singhal, and C. Choudhary, “MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices,” *Souvenir 2015 IEEE Int. Adv. Comput. Conf. IACC 2015*, pp. 397–402, 2015.

[22] R. Mirshahvalad and N. A. Zanjani, “Diabetes prediction using ensemble perceptron algorithm,” *Proc. - 9th Int. Conf. Comput. Intell. Commun. Networks, CICN 2017*, vol. 2018-Janua, pp. 190–194, 2018.

[23] V. V. Vijayan and C. Anjali, “Prediction and diagnosis of diabetes mellitus - A machine learning approach,” *2015 IEEE Recent Adv. Intell. Comput. Syst. RAICS 2015*, no. December, pp. 122–127, 2016.

[24] J. J. Pangaribuan and Suharjito, “Diagnosis of diabetes mellitus using extreme learning machine,” *2014 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2014 - Proc.*, no. November, pp. 33–38, 2014.

[25] F. G. Woldemichael and S. Menaria, “Prediction of Diabetes Using Data Mining Techniques,” *Proc. 2nd Int. Conf. Trends Electron. Informatics, ICOEI 2018*, no. Icoei, pp. 414–418, 2018.

[26] M. F. Faruque, Asaduzzaman, and I. H. Sarker, “Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus,” in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–4.

[27] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, “Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction,” *Artif. Intell. Med.*, vol. 85, pp. 43–49, 2018.

[28] M. Asad, U. Qamar, B. Zeb, A. Khan, and Y. Khan, “Blood glucose level prediction with minimal inputs using feedforward neural network for diabetic type 1 patients,” *ACM Int. Conf. Proceeding Ser.*, vol. Part F1481, pp. 182–185, 2019.

[29] K. V. S. R. P. Varma, A. A. Rao, T. Sita Maha Lakshmi, and P. V. Nageswara Rao, “A computational intelligence approach for a better diagnosis of diabetic patients,” *Comput. Electr. Eng.*, vol. 40, no. 5, pp. 1758–1765, 2014.

[30] W. Xao, F. Shao, J. Ji, R. Sun, and C. Xing, “Fasting blood glucose change prediction model based on medical examination data and data mining techniques,” *Proc. - 2015 IEEE Int. Conf. Smart City, SmartCity 2015, Held Jointly with 8th IEEE Int. Conf. Soc. Comput. Networking, Soc. 2015, 5th IEEE Int. Conf. Sustain. Comput. Communic*, pp. 742–747, 2015.

[31] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, “An Efficient Rule-Based Classification of Diabetes Using ID3, C4.5, & CART Ensembles,” *Proc. - 12th Int. Conf. Front. Inf. Technol. FIT 2014*, pp. 226–231, 2015.

[32] Z. Tafa, N. Pervetica, and B. Karahoda, “An intelligent system for diabetes prediction,” *Proc. - 2015 4th Mediterr. Conf. Embed. Comput. MECO 2015 - Incl. ECyPS 2015, BioEMIS 2015, BioICT 2015, MECO-Student Chall. 2015*, pp. 378–382, 2015.

[33] S. Qiu, P. Wang, J. Li, X. Gao, and B. Chen, “An improved prediction method for diabetes based on a feature-based least angle regression algorithm,” *ACM Int. Conf. Proceeding Ser.*, pp. 232–238, 2019.

[34] S. Rallapalli and T. Suryakanthi, “Predicting the risk of diabetes in big data electronic health Records by using scalable random forest classification algorithm,” *Proc. - 2016 3rd Int. Conf. Adv. Comput. Commun. Eng. ICACCE 2016*, pp. 281–284, 2017.

[35] A. Sarwar and V. Sharma, “Comparative analysis of machine learning techniques in prognosis of type II diabetes,” *AI Soc.*, vol. 29, no. 1, pp. 123–129, 2014.

[36] N. Nnamoko, A. Hussain, and D. England, “Predicting Diabetes Onset: An Ensemble Supervised Learning Approach,” *2018 IEEE Congr. Evol. Comput. CEC 2018 - Proc.*, pp. 1–7, 2018.

[37] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, and D. Khalili, “The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes,” *Med. Decis. Mak.*, vol. 36, no. 1,

pp. 137–144, 2016.

[38] D. Dutta, D. Paul, and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," *2018 IEEE 9th Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEMCON 2018*, pp. 924–928, 2019.

[39] A. Anand and D. Shakti, "Prediction of diabetes based on personal lifestyle indicators," *Proc. 2015 1st Int. Conf. Next Gener. Comput. Technol. NGCT 2015*, no. September, pp. 673–676, 2016.

[40] S. K. Dey, A. Hossain, and M. M. Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm," *2018 21st Int. Conf. Comput. Inf. Technol. ICCIT 2018*, pp. 1–5, 2019.

[41] K. J. Wang, A. M. Adrian, K. H. Chen, and K. M. Wang, "An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus," *J. Biomed. Inform.*, vol. 54, pp. 220–229, 2015.

[42] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018.

[43] Y. Srivastava, P. Khanna, and S. Kumar, "Estimation of Gestational Diabetes Mellitus using Azure AI Services," *Proc. - 2019 Amity Int. Conf. Artif. Intell. AICAI 2019*, pp. 321–326, 2019.