UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Análisis comparativo de técnicas de Machine Learning sobre el método de muestreo para la predicción de diabetes

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE:

Ingeniero de Sistemas

AUTORES:

Chira Bohorquez, Piero Alejandro (orcid.org/(0000-0003-0844-766X)

Rivera Munive, Kevin (orcid.org/(0000-0003-4273-0143)

ASESOR:

Dr. Daza Vergaray Alfredo (orcid.org/0000-0002-2259-1070)

LÍNEA DE INVESTIGACIÓN:

Sistema de Información y Comunicaciones

LÍNEA DE RESPONSABILIDAD SOCIAL UNIVERSITARIA:

Desarrollo económico, empleo y emprendimiento

Lima - Perú

2023

Dedicatoria:

Con inmenso amor y cariño a toda mi familia por ser mi soporte y fuente de inspiración, principalmente a mis padres Juan y Teodocia, a mi esposa Marisol, por brindarme su apoyo, comprensión y fortalezas para seguir adelante y cumplir mis objetivos, asimismo a mi hijo Adriel, quien descansa en paz a lado de nuestro divino creador, “mi angelito de alas mágicas”.

Kevin

A nuestras familias, principalmente a mi hijo Alexander y a mi esposa Cristal por guiarme hacia el camino de superación, por su apoyo incondicional en todo momento con el esfuerzo y dedicación que han hecho que nuestros sueños se hagan realidad y ver hacia un futuro mejor.

Piero

Agradecimiento

A la Universidad Cesar Vallejo por abrirnos las puertas hacia la educación profesional y brindarnos acceso a las herramientas y bases de datos más grandes que existen, asimismo a todos nuestros docentes, principalmente al Dr. Alfredo Daza Vergaray, por las enseñanzas, el tiempo, la paciencia sobre todo las herramientas brindadas para culminar la etapa final del desarrollo de la presente tesis.

Kevin

Asimismo, un agradecimiento a nuestra familia por el apoyo constante que nos brindan y confiar en nosotros; así como a la Universidad Cesar Vallejo y a nuestros docentes por guiarnos durante todo el proceso de desarrollo del presente proyecto quienes fueron clave para concluirlo.

Piero

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ECUACIONES

RESUMEN

En siguiente trabajo se realizó con el objeto de aplicar un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo para la predicción de la diabetes. Para esto, se realizó una investigación usando un método de enfoque cuantitativo, aplicado a los datos de un repositorio de base de datos de Kaggle de medición de factores de diabetes en mujeres de al menos 21 años de herencia indígena Pima, la misma que consta de 768 ítems, las mismas que han sido considerados como población para posteriormente ser usada como muestra. Asimismo, el estudio es de tipo aplicada, con un diseño de investigación experimental de tipo pre-experimental de un solo grupo, ya que luego de aplicar las técnicas de Machine Learning a través de métricas como rendimiento; exactitud, precisión, especificidad, sensibilidad y F1 Score, se podrá verificar los resultados y realizar la medición.

Para ello, se consideró aplicar la metodología Knowledge Discovery in Databases (KDD), la misma que está divida de 5 etapas, la primera comienza con la selección de datos, la segunda y tercera etapa, con el preprocesamiento y transformación de los datos, en la cuarta etapa se efectúa la minería de datos, aplicado a la presente investigación, haciendo el entrenamiento en 6 algoritmos de aprendizaje automático Árbol de decisiones (DT), Random Forest (RF), máquina de vectores de soporte (SVM), Gradient Boosting Machine (GBM), K-vecino más cercano (K-NN) y Redes Neuronales (ANN), basando los resultado en los mejores hiperparámetros y por último en la quinta etapa, se diseñó un software para apoyar en la detección de la diabetes en función a 5 métricas, obteniendo los resultados en base a 6 algoritmos.

Como resultado se obtuvo que el modelo Random Forest (RF), Gradient Boosting Machine (GBM) y Árbol de Decisiones (DT) superaron a los demás modelos, el modelo Random Forest obtuvo un 79,22%, en cuanto a la métrica exactitud, mientras que el modelo GMB obtuvo un 75,32%, de exactitud, del mismo modo el árbol de decisiones (DT) obtuvo un 74.09% en cuanto a la precisión. Por otro lado, el KNN, ANN y SVM fueron los modelos de menor rendimiento en la mayoría de las cinco métricas, KNN con un 74.02%, ANN con un 63.63 % y SVM con un 73.10% de exactitud. Finalmente, en función a los resultados obtenidos por las métricas evaluadas se puede afirmar que el uso de Técnicas de Machine Learning para la predicción de la diabetes, son favorables para el sector salud.

Palabras clave: Machine learning, diabetes, análisis, métricas de precisión.

ABSTRACT

The following work was carried out in order to apply a comparative analysis of Machine Learning techniques on the sampling method for the prediction of diabetes. For this, an investigation was conducted using a quantitative approach method, applied to data from a repository of Kaggle database measuring diabetes factors in women of at least 21 years of Pima indigenous heritage, which consists of 768 items, the same that have been considered as a population to be subsequently used as a sample. Likewise, the study is applied, with an experimental research design of pre-experimental type of a single group, since after applying Machine Learning techniques through metrics such as performance; accuracy, precision, specificity, sensitivity and F1 Score, it will be possible to verify the results and perform the measurement.

For this, it was considered to apply the Knowledge Discovery in Databases (KDD) methodology, which is divided into 5 stages, the first begins with the selection of data, the second and third stage, with the preprocessing and transformation of the data, in the fourth stage data mining is performed, applied to this research, training in 6 machine learning algorithms Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting Machine (GBM), K-Nearest Neighbor (K-NN) and Neural Networks (ANN), basing the results on the best hyperparameters and finally in the fifth stage, software was designed to support the detection of diabetes based on 5 metrics, obtaining the results based on 6 algorithms.

As a result, the Random Forest (RF) and Gradient Boosting Machine (GBM) models outperformed the other models, the Random Forest model obtained 79.22% for the accuracy metric, while the GBM model obtained 78.66% for specificity, likewise the decision tree (DT) obtained 74.09% for accuracy. On the other hand, the KNN, ANN and SVM were the lowest performing models in most of the five metrics, KNN with 74.02%, ANN with 63.63 % and SVM with 73.10% accuracy. Finally, based on the results obtained by the evaluated metrics, it can be affirmed that the use of Machine Learning Techniques for the prediction of diabetes is favorable for the health sector.

Keywords: Machine learning, diabetes, analysis, precision metrics.

I. INTRODUCCIÓN

Durante las últimas décadas, las tecnologías de la información siempre han estado a la par con las investigaciones y nuevos métodos que impulsan un cuidado relacionado estrechamente en el ámbito de la salud. La diabetes es un padecimiento muy grave y común; puesto que la ingesta de alimentos con altos niveles de azúcar que perjudican al ser humano. Esto a su vez ha generado muy altos índices de personas que padecen de esta enfermedad, siendo un reto actual la forma en la que la predecimos; sin embargo, un diagnóstico anticipado podría ayudar a tratarla y prevenir futuras complicaciones. A manera de resumen las causas y efectos que se pudo identificar se describe en el anexo 4.

Al respecto, también en las investigaciones se considera que la diabetes es un diagnóstico que se relaciona a través de manifestaciones hiperglucemiantes, (aumento del azúcar en la sangre), tal como lo considera («American Diabetes Association» 2022), es el resultado de la interrelación de dos principales síntomas son el defecto de las funciones del páncreas y la reducción de la insulina en los tejidos corporales; principalmente ligados al estilo y forma de vida que lleva cada persona. Considerando lo anteriormente mencionado, somos nosotros mismos que a fin de afrontar esta dura enfermedad en toda la sociedad, nos alimentamos de cosas que no perjudican nuestra salud.

Por otro lado, dentro de las afectaciones que esta pueda tener, se considera que la diabetes se está propagando vertiginosamente y que puede llevar a ser crónica y compleja, la causa principal es el modo de vida que lleva cada persona, debido a la mala alimentación al consumir comidas denominadas “chatarra” o con altos porcentajes de grasa, es aquí, donde se observa que la actual alimentación está produciendo altos niveles de concentración de sodio, azúcares y grasas que están perjudicando fuertemente la salud en las personas (Mujumdar y Vaidehi 2019) .

En ese sentido, (Vizcarra y Ordóñez 2018), mencionan que la diabetes mellitus está ligada a las enfermedades metabólicas las mismas que presentan a la hiperglicemia como principal característica, la cual se debe al aumento de la segregación de insulina. La hiperglucemia crónica, principalmente es asociada a un diagnóstico que hace daño la insuficiencia y disfunción de varios órganos, afectando especialmente a los vasos sanguíneos, insuficiencia renal, corazón, sistema nervioso y produce ceguera.

Es preciso indicar que para la (International Diabetes Federation 2021), la diabetes de tipo 1 denominada DM1, es definida como la problemática en la que los niveles de insulina producidos por el páncreas son menores al promedio. Así mismo la de tipo 2 denominada DM2 basa su causa en que se produce resistencia a la insulina, lo cual no produce efectos. Se considera que, en los últimos 30 años, la DM2 ha sufrido un aumento considerable alrededor de todos los países. En general está siendo catalogada como una emergencia sanitaria, durante el siglo XXI se manifiesta que, en el 2021, una cantidad de 537 millones de personas tienen detectada la diabetes, se calcula que llegue a 643 millones en el año 2030 y 783 millones para el año 2045.

En ese contexto, (Alegre-Díaz et al. 2019), considera y precisa que si no es controlada la diabetes, la probabilidad de presentar complicaciones graves se vuelve extremadamente alta y que puede conllevar hasta la muerte, del mismo modo, los pacientes con tales síntomas, poseen altas probabilidades de presentar complicaciones como tuberculosis cardiovasculares, tal como lo demuestra en su investigación acerca de la toma de muestras en 100 000 mil mujeres y 50,000 hombres en la ciudad de México, acerca de las causas de mortalidad que conlleva a la diabetes, considerando con poca asociación a la cirrosis o el cáncer pulmonar.

Para (Krasteva et al. 2018), considera que esta enfermedad si no se trata en la medida de lo posible puede ocasionar diferentes complicaciones para el paciente con diabetes produciendo malestar y molestias, ya que al ser comúnmente diagnosticada en adultos generan una alta cantidad de estrés y preocupación; asimismo todas estas complicaciones de los pacientes generan una sobrecarga en las atenciones del sector salud, lo que generalmente es resultado que la insulina no es adaptada por el organismo y éste a su vez se hace resistente a su asimilación o su producción es insuficiente.

En relación a los niveles de Hiperglucemia, (Stawarz et al. 2023), considera que dichos niveles cuando se encuentran elevados, pueden comprometer complicaciones a futuro, como ceguera y daños a los nervios, mientras que los niveles de glucemia muy bajos pueden provocar pérdida de conocimiento, convulsiones, coma y hasta la muerte. El autocontrol de la diabetes generalmente se relaciona directamente con la verificación del nivel estándar de glucosa que hay en la sangre y los factores del estilo de vida como una buena alimentación, así como con una rutina que físicamente sea activa y que pueda ser realizada en

varias oportunidades durante el día, si bien los médicos son los protagonistas en el apoyo a la atención de esta enfermedad, el autocontrol permanente depende principalmente de las decisiones propias que toma cada persona.

Según (Zhao et al. 2023), en su estudio precisa que para que la población no incremente la gravedad de su condición médica en relación a la diabetes, las Tecnologías de la Información han sido fundamentales, tal es el caso que ahora tienen acceso a medidores continuos de Glucosa (MCG), en aplicaciones para smartphones. Considera también que las Tecnologías de la Información tienen presencia con altos porcentajes en los mecanismos y sistemas de control de diabetes, basando en datos que contienen características clínicas, las mediciones de laboratorio y los medicamentos de los pacientes, además que se proporcionan las lecturas de monitoreo continuo de glucosa.

Asimismo, (Bergman, Stefanovski y Kim, et al. 2019), basa su investigación en la existencia de los estudios y técnicas que han logrado avances en su investigación orientados a la detección de la prediabetes, incluidos los estudios de asociación del genoma completo y la metabolómica. Si bien estos enfoques crean un gran volumen de datos con solamente una sola muestra de sangre, dichos estudios han demostrado que se obtiene un éxito con limitaciones en el uso de la información genética y metabolómica para la identificación del riesgo de enfermedad, por lo que asevera que la predicción de diabetes es la necesidad de predecir anticipadamente el desarrollo de esa enfermedad, para evitar la intervención y retrasar el resultado de la progresión de la enfermedad y el trastorno metabólico.

Por otro lado, (Malpartida, et al. 2022), precisa que hacer cambios en la ingesta de alimentos y las rutinas físicas adecuadas, reducen el progreso de ser portador de la diabetes tipo 2. Además, señala que la capacidad de la inteligencia artificial y las técnicas de Machine Learning para analizar conjuntos de datos complejos ayuda a los médicos en la predicción temprana de la diabetes, de tal forma que contribuirá a la atención planificada de los pacientes, lo que resultará en la mejora de los resultados en las atenciones médicas. La utilización y relevancia de la Inteligencia Artificial para las decisiones clínicas, el establecimiento de una alerta temprana y la calificación de los riesgos, son las áreas más prometedoras del desarrollo del análisis de datos.

Por otra parte, (Russell 2018, p. 8), precisa que el concepto Machine Learning, se define como la práctica de programación de computadoras para aprender de los datos. De ello se puede precisar que, este tipo de aprendizaje automático (ML), es un programa que aprende, no siendo necesariamente para ello, interesante desde el punto de vista predictivo y adaptativo a la selección de información que nos puede precisar una toma de decisión.

Según (Dami et al. 2021), el Machine Learning (ML) es un tipo de Inteligencia Artificial (IA), adaptativa y de autoaprendizaje en función al tiempo, identificando ciertos patrones de entrada y dado los algoritmos que contiene, evoluciona con el tiempo. En la actualidad hay diversos tipos de modelos predictivos y para ello se cuenta con técnicas de Aprendizaje Automático que son de ayuda a las organizaciones a través de estos modelos predictivos.

Por otra parte (Salamanca et al. 2021), considera la interpretación del aprendizaje automático como la identificación automatizada o proceso automatizado que extrae patrones en los datos. En los últimos años, se ha utilizado como una herramienta generalmente común y con buenos resultados, en todo lo relacionado con extracción de información de gran conjunto de datos.

Posteriormente, esta situación nos obliga a plantear nuestra problemática general ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo permitirá predecir la diabetes?, especialmente cuando se emplean métodos de aprendizaje automático. Los conjuntos de datos pueden contribuir al desarrollo de algoritmos o modelos basados en datos y tecnologías de control/gestión de la diabetes. Asimismo, como problemas específicos se planteó de la siguiente manera; Problema Específico 1: ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función a la sensibilidad permitirá predecir la diabetes?, Problema Específico 2: ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función a la precisión permitirá predecir la diabetes?, Problema Específico 3: ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función a la especificidad permitirá predecir la diabetes?, Problema Específico 4: ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función a la exactitud permitirá predecir la diabetes?, Problema Específico 5 : ¿Cómo un análisis comparativo de

técnicas de Machine Learning sobre el método de muestreo en función al F1 Score permitirá predecir la diabetes?

En ese contexto, el proyecto se justifica de forma teórica, debido a que está respaldado por artículos científicos, revistas indexadas y fuentes de búsqueda de información segura, así como la plataforma MyLOFT, que dio acceso a distintas bases de datos científicas y bibliotecas digitales para el desarrollo del proyecto, en ese sentido luego de comparar las distintas técnicas de Machine Learning a través de algoritmos y métodos, se podrá establecer y determinar cuál es el más adecuado para el manejo de la información acerca de los factores que determinan en el diagnóstico de la diabetes, de tal forma que se podrá establecer una identificación temprana evitando de esta manera, los problemas colaterales que esta ocasiona dicha enfermedad.

El presente proyecto presenta una justificación social dado que, debido a que dicha enfermedad está relacionada al deterioro de varios órganos, su predicción evitará que el tratamiento sea una preocupación, así como reducir las atenciones médicas relacionadas con esta enfermedad, se fomentará la conciencia social para la reducción de los casos.

Asimismo, la investigación desde el punto de vista económico se justifica en que nuestro trabajo, facilitará el reconocimiento de los factores sintomatológicos de la diabetes, de tal forma que se pueda evitar los largos y costosos tratamientos que acarrea el padecimiento de esta enfermedad. También se justifica tecnológicamente por la implementación de un sistema predictivo que contribuirá con las investigaciones futuras para el desarrollo de nuevas técnicas de desarrollo.

Para el presente trabajo se plantea el siguiente Objetivo General Aplicar un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes y también se plantearon los siguientes objetivos específicos Objetivo Específico 1: Emplear análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes en función a la sensibilidad. Objetivo específico 2: Emplear un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes en función a la precisión. Objetivo Específico 3: Emplear un análisis comparativo con técnicas de Machine Learning sobre el método de

muestreo para predecir la diabetes en función a la especificidad. Objetivo Específico 4: Emplear un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes en función a la exactitud Objetivo Específico 5: Emplear un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes en función a F1 Score.

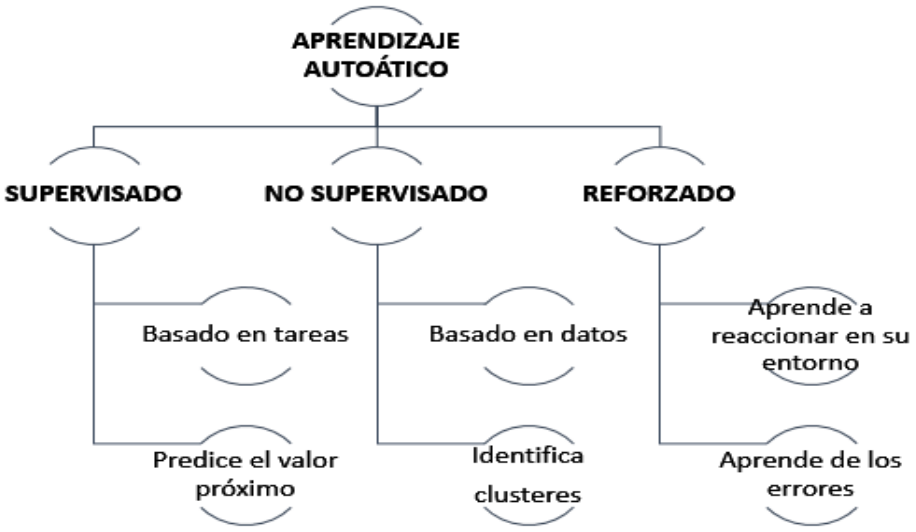
En esta investigación se establece como Hipótesis General: La aplicación de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice la diabetes y también se establece las siguientes hipótesis específicas; Hipótesis Específica 1: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con sensibilidad la diabetes. Hipótesis Específica 2: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con precisión la diabetes. Hipótesis Específica 3: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con especificidad la diabetes. Hipótesis Específica 4: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con exactitud la diabetes. Hipótesis Específica 5: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con F1 Score la diabetes.

II. MARCO TEÓRICO

En este capítulo, se realizó una búsqueda de información relevante en artículos y revistas indexadas a través de la plataforma MyLOFT, así como otros motores de búsqueda, brindando acceso a distintas bases de datos científicas como; Scopus, IEEE Xplore, Dovepress, Open Acces, Scientific reports y Scielo y Elsevier, las mismas que han servido para obtener referencias de investigaciones previas que hacen posible que la sustentación de la presente investigación sea más sólida y precisa, donde se encontró antecedentes tanto de nivel nacional e internacional, tal como se muestra en el Anexo 2, el cual muestran estudios relacionados con la predicción de esta enfermedad, empleando técnicas de aprendizaje automático a través de distintos algoritmos y métricas de rendimiento para evaluar la predicción de la diabetes.

Podemos considerar que el Machine Learning, es parte de la inteligencia artificial, la cual, a través de algoritmos, que tiene como característica principal la de aprender y no tener que ser programados explícitamente, (Sandoval et al. 2018). En otras palabras, sólo hay que proporcionar al algoritmo una gran cantidad de datos, para que pueda aprender y saber actuar en casos diferentes, (Shalev-Shwartz et al. 2019). También manifiesta que el aprendizaje automático está típicamente organizado en tres ramos principales; el aprendizaje supervisado, el aprendizaje no supervisado y aprendizaje reforzado, tal como se muestra en la figura 1.

Figura 1: Tipos de aprendizaje automático



Fuente: (Leidy-Esperanza et al. 2021).

Respecto a los considerandos relacionados con el aprendizaje supervisado (Leidy-Esperanza et al. 2021), considera que en los algoritmos que usan datos previamente identificados para indicarle cómo tiene que ser categorizada la nueva información, el agente toma conocimiento de los datos de entrada y salida pretendiendo converger con el mejor clasificador posible. En el aprendizaje no supervisado, los algoritmos no usan datos identificados previamente para indicarle al algoritmo cómo va a ser la clasificación de la información, el algoritmo debe encontrar la manera de realizar la clasificación; por consiguiente, no requiere de una persona que retroalimente el algoritmo. El aprendizaje por refuerzo es aquel que funciona con intervención humana mediante el proceso. los algoritmos aprenden de la experiencia, es decir, se debe dar un refuerzo positivo o incentivo cada vez que aciertan.

Para (Management Solutions España, et al. 2018), manifiesta que un modelo de regresión puede efectuar la predicción de una cantidad continua de datos, mientras que los de clasificación predicen una etiqueta. Puesto que estas aplicaciones son muy complejas, una persona no puede planear con tal envergadura la realización de dichas actividades por la complejidad, es por ello que tiene que hacer uso de las computadoras y concederle las habilidades de efectuar un aprendizaje en base a las experiencias obtenidas y poder tener un comportamiento adaptativo ante nuevas situaciones.

Según (Amin 2021), menciona que existen diversas técnicas de aprendizaje dentro de las cuales considera que el Modelo de Regresión Lineal, es el más utilizado para efectuar la clasificación binaria, porque es un modelo denominado básico, que puede ser extendido a problemas de múltiples etiquetas. Para (Tusell 2018), esta técnica basa su función en la predicción de una variable dependiente y en razón a una o muchas variables independientes (x), a partir de una línea recta que más se adecúe a los datos que se le ha proporcionado.

Ecuación 1: Hipótesis de modelo de regresión lineal

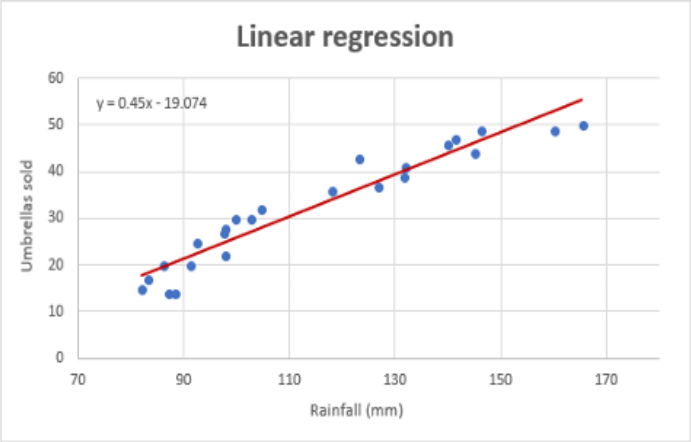
$$h(x) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots.$$

Fuente: (Tusell 2018)

(Tusell et al. 2018), considera que cuando en la regresión lineal existe una variable independiente es simple y cuando existe más de una, es denominada

múltiple. Es considerado el modelo más rápido y robusto, pero para que exista un buen funcionamiento, se debe de asegurar la relación lineal entre la entrada y la salida, en la figura 2, se muestra un ejemplo de la gráfica del modelo de RL.

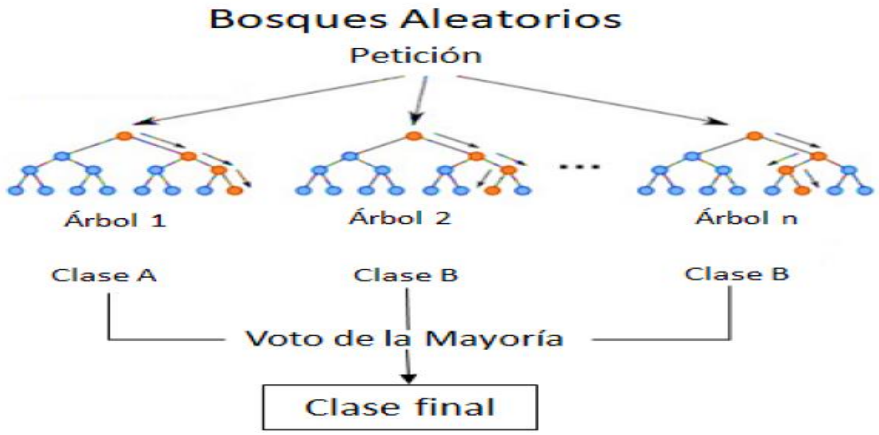
Figura 2: Gráfica de modelo de regresión lineal



Fuente: (Tusell 2018).

Para (Pineda-Jaramillo 2019), considera que los árboles de decisión, son gráficos orientados y estructurados por un número determinado de nodos que inician de los nodos raíz, son métodos no paramétricos con una estructura a un diagrama de flujo o a un árbol y se pueden utilizar para clasificar problemas, en el grafico 3 se muestra la gráfica del modelo bosque aleatorio.

Figura 3: Bosques Aleatorios.



Fuente: (Pineda-Jaramillo 2019).

Asimismo, para (González 2023), el modelo de regresión Polinomial de tercer grado y de una variable, busca encontrar un polinomio de grado n, que sea más adaptable a la distribución de datos a través de una curva. Es muy útil cuando

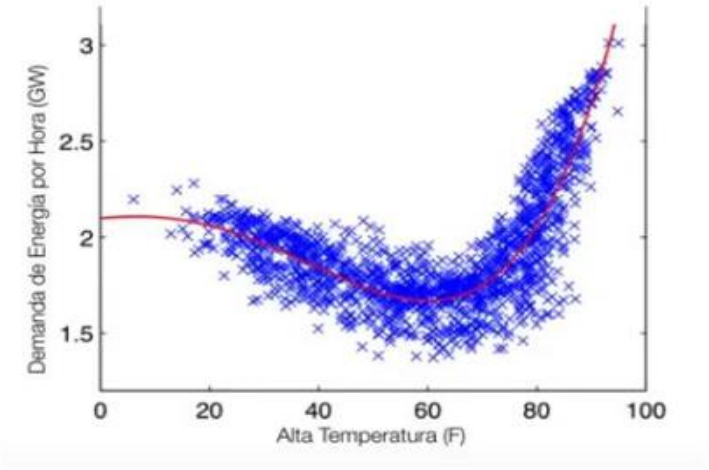
en el modelo no se ajustan los datos debido a algún tipo que no asegure mantener la linealidad entre ellos.

Ecuación 2 Hipótesis de modelo de regresión Polinomial

$$h(x) = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3$$

Fuente: (González 2023).

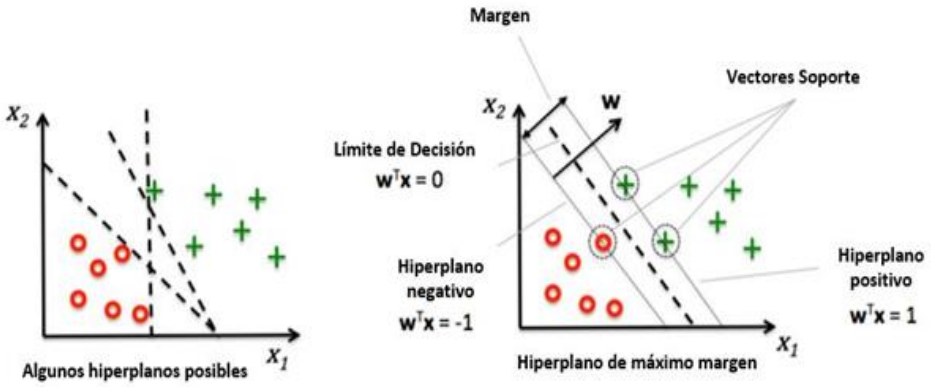
Figura 4: Ejemplo de regresión Polinomial.



Fuente: (González 2023).

Asimismo, en lo respecto a Vectores de soporte (Support Vector Machine, SMV), (Gandhi 2018), considera como objetivo principal un hiperplano en un espacio de N-dimensiones (representando N el número de variables independientes), que haga máxima la distancia existente entre los datos de ambas clases. En otras palabras, en lo que es concerniente a la clasificación de los datos es el límite de decisión. Para (Román 2019), las técnicas de mapeo y kernelización deben ser usadas en estos problemas de clasificación no lineales.

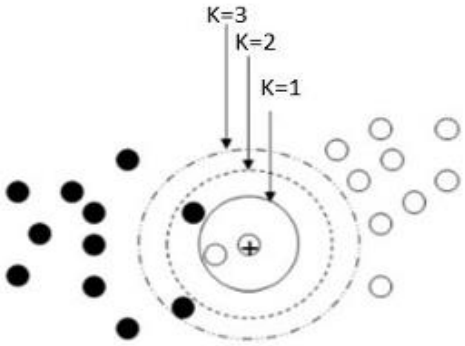
Figura 5: Representación Gráfica de SMV.



Fuente: (Roman 2019).

Según (García-Laencina et al. 2017), este modelo asume que los objetos que tienen similitud están muy cercanos unos con otros y su similitud son considerada en base a las distancias entre puntos en una gráfica. Es así que plantean que la variable (k), equivale al número de vecinos más cercanos que se eligen con el fin de efectuar la clasificación, de acuerdo a ello, se obtendrán diferentes predicciones.

Figura 6: Ejemplo de KNN, clasificación de acuerdo a parámetro



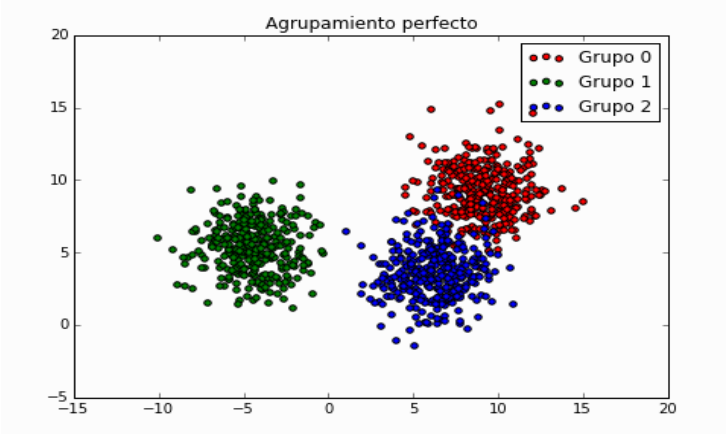
Fuente: (García-Laencina et al. 2017).

Respecto al aprendizaje no supervisado, según (Russell y Norvig 2018), mencionan que el agente identifica los patrones que existen dentro de los datos de entrada sin ser necesario tener en consideración los datos de salida. En ese sentido, el objeto es obtener y recabar información representativa de los datos de entrada, los mismos que no tienen etiquetas y su estructura no es conocida. Como problemática se presenta dos tipos: el agrupamiento o también llamado clustering, y la reducción dimensional. En primer lugar, se crean conjuntos de objetos de similares características, en tanto que el de reducción busca redundancia en la

información de los datos, para efectuar la reducción de la cantidad de variables y con ello, así como ampliar el espectro de la mejorando significativamente el rendimiento computacional (Roman 2019).

Para (López Briega 2018), menciona que los modelos K-Means es uno de los modelos más conocidos en este método, dentro de su concepción tiene varias etapas, considerando la primera en identificar la variable (k), que vendría a ser la cantidad de clústeres, para posteriormente de forma aleatoria elegir (k) datos del set los cuales se les llamaran centroides los mismos que harán su desplazamiento hacia el punto que equivalga a la media entre las distancias del dato y su centroide. Este proceso será iterativo hasta que el centroide apenas logre moverse entre cada interacción, pudiendo obtener la denominada convergencia.

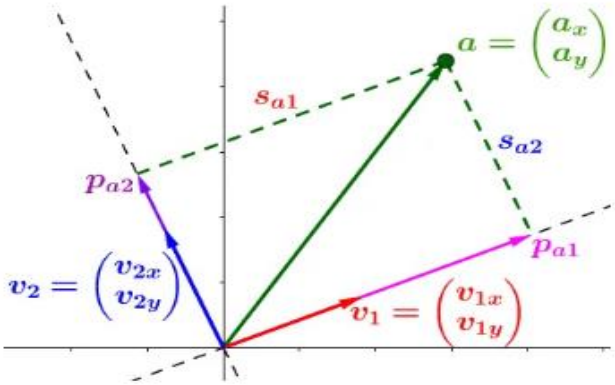
Figura 7: Representación de modelo K-Means



Fuente: (López Briega 2018)

Para (Abdulrahman 2019), el método de descomposición en Valores Singulares (Singular Valúa Decomposition o SVD): es un método mediante el cual a una matriz real o compleja se le aplica una factorización con la finalidad de reducir las dimensiones. Su principio es descomponer vectores, de tal forma que se puedan expresar en dos variables, por lo que la dirección de la proyección es el unitario y proyección indica la longitud.

Figura 8: Representación de modelo SVD



Fuente: (Abdulrahman 2019).

Para (Febrian et al. 2023), en su investigación usó la comparación de técnicas de Machine Learning, usando algoritmos de K-Nearest Neighbor (KNN) y Naïve Bayes (NB) para predecir la enfermedad de la diabetes, mediante la cual usó un método de enfoque cuantitativo que encamina la medición de los datos existentes, de una base de datos pública de Kaggle, en función a varios atributos de salud, bajo las métricas de accuracy, precisión y Recall, para poder determinar cuál de los dos es más adaptativo para la predicción de la diabetes, el algoritmo Naive Bayes supera a KNN, con un valor promedio de 76,07 % de precisión, 73,37 %, de recuperación y en el caso del Naive Bayes un valor promedio de 73,33 % de precisión y 70,25 % de recuperación, se concluye que al comparar el algoritmo k-Nearest Neighbor y el algoritmo Naive Bayes, de acuerdo con los resultados de sus experimentos a través del sistema de medición de Matriz de Confusión, el algoritmo Naive Bayes es preferible para predecir la diabetes utilizando el conjunto de datos.

Según (Mansoori et al. 2023), en su estudio precisa que existe un gran problema por el que viene padeciendo la población en general a consecuencia de la diabetes Mellitus tipo 2 (DM2, su objetivo principal de este estudio fue anticipar el diagnóstico de la diabetes usando la comparación de los modelos sobre una muestra de 9000 adultos de 35 a 55 años, utilizando tres modelos de aprendizaje automático; regresión lineal, árbol de decisión y bosque aleatorios, para investigar la relación entre los predictores hematológicos y las variables de respuesta binaria (diabéticos y no diabéticos), tomando como método el algoritmo SMOTE para equilibrar las clases, bajo las métricas de accuracy, precisión y especificidad, para predecir la diabetes usaron el conjunto de datos que se dividió aleatoriamente en

dos partes: datos de entrenamiento y datos de prueba (75 % frente a 25 %). finalmente, el estudio mostró que el modelo bosque aleatorios presentó un mejor rendimiento para la predicción de la diabetes obteniendo un 97,43% de precisión a comparación de los modelos de regresión lineal y árbol de decisión, con 67,28% y 66,26% de precisión, de acuerdo con los resultados, se puede concluir que algunos de los factores hematológicos podrían ser una herramienta valiosa en la predicción de T2DM.

Según (Orlando y Karina 2022), en su investigación aplicó modelos de aprendizaje automático para la detección anticipada de la diabetes tipo 2, utilizando 5 modelos, el K-vecino más cercano (K-NN), Bernoulli Naïve Bayes (BNB), el árbol de decisión (DT), la regresión logística (LR) y la máquina de vectores de soporte (SVM), con el objetivo de identificar y clasificar si un paciente tiene diabetes o no ,utilizando modelos ML y seleccionar el mejor modelo de clasificación para predecir la diabetes, asimismo para determinar el rendimiento de los modelos de clasificación se utilizaron diferentes métricas, como la puntuación F1, exactitud, precisión y recuperación, tomando como método el algoritmo SMOTE para equilibrar las clases**,** los resultados muestran que los modelos K-NN y BNB superan a los demás modelos, el modelo K-NN obtuvo la mejor precisión en la detección de diabetes, con un 79,6% de precisión, mientras que el modelo BNB obtuvo un 77,2% de precisión en la detección de la diabetes, el cual concluye que, en base a los resultados obtenidos, los dos mejores modelos para identificar y clasificar la diabetes tipo 2 mediante modelos ML son K-NN y BNB, garantizando que el uso de modelos ML para la detección temprana de diabetes es muy prometedor en el sector salud.

Según (Rajput y Khedgikar et al. 2022), teniendo como objetivo el predecir la diabetes, a través del uso de cinco diferentes tipos de algoritmos de Aprendizaje Automático, como Vectores de Soporte, Vecino más cercano, árbol de decisión, bosque aleatorio, regresión logística y aumento de gradiente demostrando que los algoritmos de aumento de gradiente estocástico y árbol de decisión, obtuvieron valores que superaron al rendimiento y lograron un mayor índice de precisión para la obtención de datos a comparación de los tipos de algoritmos 77%, planteando incrementar cinco algoritmos para mejorar su precisión tales como Regresión logística multinomial, Naïve Bayes, Árboles de decisión, Bosque aleatorio y random forest, concluyendo que los mejores resultados fueron obtenidos para el

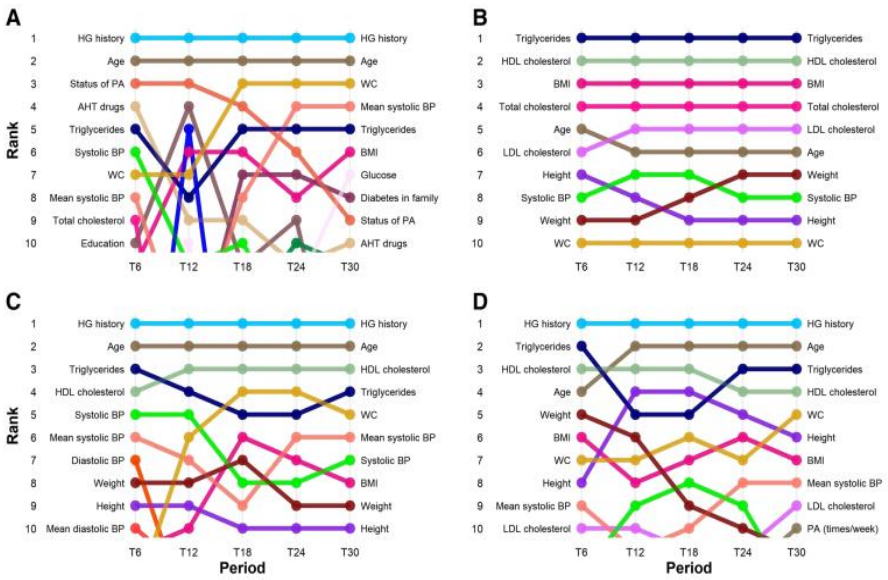
árbol de decisión y el algoritmo de aumento de gradiente estocástico, llegando a concluir el índice de masa corporal (IMC) alto y la edad avanzada son factores importantes en el desarrollo del riesgo de diabetes.

Según (Li et al. 2022), evalúa los modelos de aprendizaje automático fundados en información compilada en el mediano plazo, para la predicción de la diabetes en pacientes con síndrome metabólico; para ello, basa su investigación en la recopilación de información desde el año 2008 hasta el 2020, los mismos que se encuentran en el repositorio de la base de datos del área de Gestión de Salud del Hospital de la Facultad de Medicina de la Unión de Pekín (PUMCH-HM), para lo cual se tomó una muestra de 4510 participantes, los mismos que fueron tomados de manera aleatoria y dentro del período considerado anteriormente los mismos que se evaluaron en tres algoritmos de clasificación convencionales: regresión logística, bosque aleatorio y Xgboost, utilizando la herramienta Python 3.8. Para dichos algoritmos se desarrollaron 5 modelos de riesgo, en todos los clasificadores se calcularon utilizando un valor de estado aleatorio fijo para garantizar resultados consistentes. Concluye que los modelos basados en datos longitudinales de varios años pueden proporcionar herramientas de evaluación más personalizadas para la evaluación del riesgo en pacientes con diabetes.

Para (Zhao et al. 2023), basa su investigación en la elaboración de datos denominados el ShangháiT1DM y ShangháiT2DM, que se elaboraron en función a pacientes que cuentan con diabetes mellitus tipo 1 y tipo 2. Considerando que la Diabetes Melitus de tipo 1 a partir de ahora denominada (DM1) representa del 5 al 10% del total de la población con diabetes y que la Diabetes Melitus de tipo 1 a partir de ahora denominada (DM2) basa su causalidad en cómo el organismo se resiste a la asimilación de la insulina y la deficiencia de está, teniendo como objetivo proporcionar evidencias para recomendaciones en el modo de vivencia de las personas y efectuar un seguimiento del control glucémico de estas, cada paciente se sometió a un examen físico que incluía la medición de la talla y el peso. El índice de masa corporal (IMC) se calculó como el peso dividido por la altura al cuadrado (kg/metro2). Cada paciente usó un dispositivo de monitoreo de glucosa flash (FreeStyle Libre H, Abbott Diabetes Care, Witney, Reino Unido) para efectuar la medición de los niveles de glucosa intersticial de forma continua hasta por 14 días.

Asimismo, (Kopitar et al. 2020), basa su estudio en la detección temprana de la DM2, utilizando modelos de predicción basados en aprendizaje automático, utilizando el método de regresión multivariada, para ello compara los modelos de aprendizaje automático como Glmnet, el cual es un método de regresión lineal; así como Random forest, XGBoost y LightGBM, que son métodos de árboles de decisiones, a manera de resultado los modelos más óptimos se considera a Glmnet con (0,859) % y XGBoost (0,881) %, como resultado en sus gráficas tenían cierta semejanza. (considerar las interpretaciones de los gráficos), utilizando el método Lasso, que asegura el rendimiento de los conjuntos de datos; así como bajo la recuperación de la curva de precisión, tal como se aprecia a continuación, considerando que ambos métodos de regresión lineal son aceptables para la predicción de la diabetes.

Figura 9: Importancia de las variables Ranking de variables para Glmnet (A), Light GBM (B), bosque aleatorio (C) y XGBoost (D) durante el período observado (T6–T30)

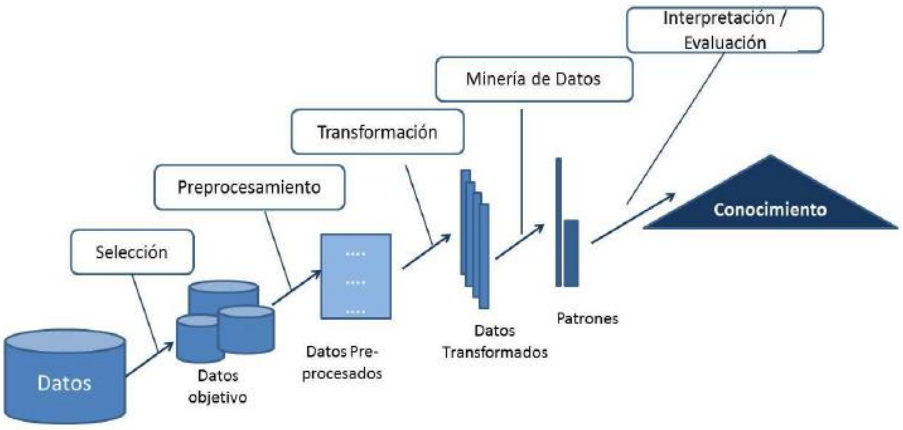


Fuente: (Kopitar et al 2020).

Por otro lado, dentro de las Métricas de Evaluación de los Modelos, se ha considerado evaluar las métricas de rendimiento, precisión, sensibilidad y F1 Score, a través de matrices de confusión en donde tambien (Abdulkadir y Derbew 2023), consideran que esta última es una herramienta comúnmente utilizada dentro del aprendizaje automático supervisado. En donde las instancias de una clase predicha son representadas en las columnas y las instancias de una clase real son representadas en cada fila.

Asimismo, la metodología que se va a emplear es KDD, en inglés (Knowledge Discovery in Database), representado en la figura 8, según Maria Consuelo et al. (2017), considera que es un proceso interactivo de tal forma que se puedan realizar cambios y repetir cada paso para obtener los mejores resultados, se basa en las disciplinas tradicionales para obtener información a través de datos del aprendizaje automáticos lo que engloba al proceso que pretende extraer conocimiento a partir de datos. Lo que a este proceso se le define como proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia entendibles en los datos, paso iterativo que consta de una serie de fases para la generación de conocimiento y la toma de decisiones. Asimismo, consideran que el proceso de la metodología Knowledge Discovery in Databases (KDD), es la preparación de datos, selección limpieza de los mismos, incorporación del conocimiento, e interpretación de los resultados. Estas fases comprenden de 5 etapas.

Figura 10: Etapas de la metodología KDD



Fuente: María Consuelo (2017).

Según (Maria y Daniel, 2017) La primera etapa es la selección, luego de un entendimiento del problema y definido las metas del proceso, se crea un conjunto de datos sobre el cual se buscará conocimiento nuevo, la segunda etapa es el pre-procesamiento / limpieza, básicamente se trata de una etapa de análisis de calidad de la data, donde se eliminan datos ruidosos, se utilizan estrategias para homogenizar datos desconocidos, datos nulos, duplicado la tercera etapa es transformación / reducción, tiene como objetivo eliminar variables no influyentes según la meta del proceso, para lo cual se utilizan técnicas de reducción para disminuir el número de variables la cuarta etapa es minería de datos, de la vista

minable generada en la etapa anterior se aplica técnicas con el fin de descubrir patrones o reglas. Finalmente, la etapa de interpretación, donde se comprende la interpretación de los patrones encontrados, visualizando y traduciendo los mismos en términos comprensibles por el usuario.

Para (KUMAR Dewangan y AGRAWAL et al. 2020), el rendimiento de un modelo se puede evaluar con diferentes medidas de rendimiento, tales como precisión, sensibilidad y especificidad, para su evaluación se usa verdadero positivo (TP), verdadero negativo (TN), falso positivo (FP) y falso negativo, tal como lo describe la tabla 1.

Tabla 1: Matriz de confusión

		Positive	Negative
Actual Class	**Positive**	(TP) True Positive	(FN) False Negative
	Negative	(FP) False Positive	(TN) True Negative

Fuente: (KUMAR Dewangan y AGRAWAL 2019).

- True Positives (TP): son las predicciones positivas que realmente son positivos para la clase.
- False Positives (FP): son las predicciones positivas que realmente son negativos para la clase.
- True Negatives (TN): son las predicciones negativas que realmente son negativas para la clase.
- False Negatives (FN): son las predicciones negativas cuando en realidad son positivas

Adicionalmente, (Zapeta Hernández et al. 2022), que la precisión mide la cantidad de verdaderos positivos y falsos positivos para hacer precisa su predicción basada en un índice porcentual de las predicciones acertadas, lo cual

resulta siendo de real importancia con el fin de evitar una confusión entre las muestras tanto positivas como negativas. De esta forma, los resultados obtenidos serán de calidad, en función a la naturaleza intuitiva de esta medición.

Ecuación 3: Fórmula para el cálculo de la precisión

$$Presición = \frac{TP}{TP + FP} * 100$$

Fuente: (Zapeta Hernández et al. 2022)

Dónde:
TP: True Positive
FP: False Positive

Adicionalmente, con el fin que el modelo sea más sensible, (KUMAR Dewangan y AGRAWAL 2020), considera que la citada métrica considera el porcentaje de los casos positivos que se identifican dentro de los parámetros indicados, a fin de reducir las probabilidades de error durante las mediciones.

Ecuación 4: Fórmula para calcular la sensibilidad

$$Sensibilidad = \frac{TP}{TP + FN} * 100$$

Fuente: (Zapeta Hernández et al. 2022)

Según (Malpartida 2019), menciona como especificidad o tasa de falsos negativos: éste es el número de casos negativos que el algoritmo identifica correctamente, gracias a este, el modelo podrá clasificar correctamente los casos con resultados negativos y prevenir errores dentro de lo posible.

Ecuación 5: Fórmula para calcular la especificidad

$$Especificidad = \frac{TP}{TP + FN} * 100$$

Fuente: (Malpartida et al. 2022)

En lo paralelo, también consideran que el Accuracy es una métrica utilizada para determinar cuántas predicciones correctas produjo un modelo a través de todo un conjunto de datos de prueba. Así como también se decide obtener una mejor solución individual y seleccionar una mejora solución producida por un algoritmo creando un modelo adecuado. (Malpartida (2019).

Ecuación 6: Fórmula para calcular el accuracy

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

Fuente: (Malpartida et al. 2022)

Asimismo, el mismo autor considera que la métrica F1-score mide la precisión del modelo en función de la sensibilidad y la precisión, generando que un valor más alto de F1-score indica que el modelo es más preciso, generando que puede ser un valor más objetivo en el cálculo.

Ecuación 7: Fórmula para calcular el F1 Score

$$F1\ Score = \frac{Precisión * Sensibilidad}{Precisió + * Sensibilidad}$$

Fuente: (Malpartida et al. 2022)

III. METODOLOGÍA

} En este capítulo, se detalla el tipo y diseño de investigación, incidiendo principalmente en aspectos relevantes como la definición y la forma en la que operan las variables. Adicionalmente, se establece la delimitación de la población; así como la determinación de la muestra y se profundiza acerca del método de análisis de datos y los aspectos éticos de la investigación,

3.1 Tipo y diseño de investigación

3.1.1 Tipo de investigación

La presente investigación es de tipo aplicada, ya que, se emplea un Sistema Inteligente con Machine Learning con la finalidad de recabar un resultado de la predicción de la diabetes en mujeres de al menos 21 años de herencia indígena Pima. Para (Teodoro y Nieto 2018), en su investigación acerca del método científico, considera que la investigación aplicada busca la resolución de los problemas de los procesos de bienes y servicios de la actividad de los seres humanos y considera que mejora, perfecciona y optimiza del funcionamiento de los sistemas relacionado a las tecnológicas actuales las mismas que están a la par con la ciencia y la tecnología; razón por la cual, no solamente está enfocado a encontrar un verdadero, falso o probable sino a la de eficiente, deficiente, ineficiente, eficaz o ineficaz.

Es de tipo cuantitativa, donde (Vega-Malagón et al. 2018), considera que el enfoque cuantitativo se basa en el planteamiento del problema y se fundamenta en el esquema lógico y deductivo, también pretende generalizar los resultados en base a lo consultado de las muestras representativas.

Tomando en cuenta el método de obtención de datos (Guevara et al. (2020), considera que, en el tipo de investigación experimental, es un conjunto de sujetos es sometido a diferentes circunstancias (variable independiente), con el fin de verificar los efectos que provocan (variable dependiente). Asimismo, considera el método exitoso si se observan cambios en la variable dependiente, producto de la manipulación de los datos de la variable independiente. Este estudio abordará la problemática, efectuando la comparación de los datos sobre el muestreo para la predicción de la diabetes.

3.1.2 Diseño de investigación

El diseño de la presente investigación se planteó un diseño experimental, de tipo preexperimental, ya que según (Hernández Sampieri et al. 2018), busca establecer un estímulo a un determinado grupo, a fin de posteriormente realizar una medición de una o más de las variables consideradas, a fin realizar una verificación en función al primer grupo.

Esta investigación basa su diseño de investigación experimental, de tipo pre - experimental, dado que se realizará, comparaciones con información, puesto que se busca establecer porcentajes de medición de la precisión de las técnicas de aprendizaje de un cierto grupo de personas que padecen la enfermedad de la diabetes, así como los resultados obtenidos de acuerdo a las sintomatologías presentadas.

Ecuación 8: Diagrama del diseño de investigación

$$G = X \rightarrow O_1$$

Pre-test – Tratamiento

Dónde:

G: Grupo experimental (mujeres de al menos 21 años de herencia indígena pima)

X: Tratamiento (O1: Mediciones pre-test de la solución de Machine Learning)

3.2 Variables y Operacionalización

Por otro lado, (Sánchez et al. (2018)), menciona que las variables son una pieza fundamental para la investigación, es una cualidad que se le atribuye a un objeto, a la que le podemos asignar una categoría o un valor; finalmente considera que de las variables se desprenden los indicadores. Para nuestra investigación se ha considerado como variable independiente “Técnicas de Machine Learning” y como variable dependiente “Predicción de la diabetes”.

3.2.1 Variable independiente

Como variable independiente se considera a las Técnicas de Machine Learning y como variable dependiente a la predicción de la diabetes, puesto que según (Morales et al. 2019), que estas son de suma importancia para la

investigación, de tal manera que considera que la variable independiente se puede clasificar por subclasificaciones, y considera que si la investigación es de carácter experimental la variable dependiente suele ser cuantitativa.

3.2.2 Variable dependiente

En función a lo descrito en el párrafo precedente se considera como variable dependiente a la predicción de la diabetes, la misma que será evaluada mediante la aplicación de cinco indicadores las cuales son: Precisión, Especificidad, Sensibilidad, Exactitud y F1 Score, tal como se aprecia en la Tabla de Operacionalización de Variables y en la Matriz de Consistencia que se visualiza en el Anexo N° 1 y 2.

3.3 Población muestra y muestreo

3.3.1 Población

Para (Sánchez et al. 2018), un conjunto de propiedades que tienen relación y son agrupadas en componentes, se le denomina población, es la sumatoria de conjunto de elementos, pudiendo ser personas, cosas o eventos, que tienen algún criterio que los relacione. Éstos pueden ser identificados a través de la examinación del interés deseado, es por ello que serán parte de la hipótesis de investigación. Es preciso indicar que, si trabajamos con grupos de personas, es conveniente denominarlos como “habitantes”, de lo contrario, es mejor llamarlos mundos de investigación. Como población se considera los datos de todas las personas con la enfermedad de diabetes. En particular, todos los pacientes de la muestra utilizada, son mujeres de al menos 21 años de herencia indígena pima, la cual consta de 768 registros.

3.3.2 Muestra

Se tomará a partir de la fórmula del muestreo aleatorio simple, tomando en consideración un 99% de nivel de confianza y un 1% de margen de error, el resultado es el siguiente:

Ecuación 9: Fórmula del Tamaño de la Muestra

$$Tamaño\ de\ la\ Muestra = \frac{\frac{z^2 * p(1-p)}{e^2}}{1 + (\frac{z^2 * p(1-p)}{e^2 N})}$$

Donde:

N=768

Z= nivel de confianza (99%) = 2.58

Ecuación 10: Tamaño de la Muestra con parámetros

$$Tamaño\ de\ la\ Muestra = 257$$

3.3.3 Muestreo

El presente trabajo aplica un tipo de muestreo no probabilístico, por lo mismo que se consideraron todos los datos.

3.4 Técnicas de instrumentos de recolección de datos

3.4.1 Técnicas

Para (Gonzáles et al. 2020), las técnicas e instrumentos a usar en una investigación son variantes, sin embargo, éstos son determinantes, en los últimos años ciertas técnicas con cada especialidad, lo cual permite poner a la investigación en contexto con la especialidad. En la presente investigación se aplicó la obtención de datos de fuente abierta, ya que los datos a utilizar se encuentran en un repositorio de base de datos experimental en Kaggle, la cual consta de 768 ítems de medición de factores de diabetes en mujeres de al menos 21 años de herencia indígena pima, la cual se obtuvo del siguiente link: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download

3.4.2 Instrumentos

Para la presente investigación y dado que los datos obtenidos son de fuente abierta, con la finalidad de contar con la instrumentación necesaria se elaborará una ficha de registro de datos. Asimismo, a través de instrumentos como las Fichas de Registro se utilizará la técnica de observación, a fin de consignarlos y compilar la precisión por cada algoritmo usado. Es por ello que se considera a la observación como un método en el cual se registran datos a través de reglas impuestas, tal y como se aprecia en las Fichas de Registro a emplear, las mismas que están en el ANEXO del 5 al 9.

3.5 Procedimiento

Todo comenzó con la exploración de información relacionada al tema a investigar, de fuentes nacionales como internacionales, de diferentes tipos de fuentes como tesis o artículos científicos con la finalidad de hacer un análisis comparativo del método de sobre muestreo para la predicción de la diabetes relacionado al Machine Learning, para luego relacionarlo con las variables consideradas, para que dichos antecedentes académicos sustenten el desarrollo de nuestra investigación.

Para la investigación y recuperación de los datos, se obtuvieron de fuentes abiertas, en este caso data para la investigación es la de diabetes en mujeres de al menos 21 años de herencia indígena pima; en base a ello, se analizarán mediante los algoritmos Decisión Tree, Random Forest, K-Nearest Neighboor (KNN), K-Means, Naïve Bayes, Regresión Lineal y Support Vector Machine para establecer el modelo de predicción y posteriormente la interfaz del Sistema inteligente. La implementación de sistema estará desarrollada en el lenguaje Python con el software Anaconda y la herramienta Jupyter.

3.6 Método de análisis de datos

El presente trabajo de investigación recopilará la información de la base de datos antes mención, para posteriormente digitalizar la información para poder volcarlo en resultados.

Posteriormente, se empleará las fichas de registro según las métricas de evaluación y los algoritmos seleccionados para obtener la precisión, para luego

analizarlo mediante el método analítico y descriptivo, a fin de analizar los datos recopilados en la búsqueda de patrones para considerar la predicción futura. De ello, se podrá considerar cuál de los modelos de predicción es el más adecuado.

3.7 Aspectos éticos

El presente trabajo se realiza aplicando la normativa de esta casa de estudios, considerada en la RESOLUCIÓN DE CONSEJO UNIVERSITARIO N° 0531-2021/UCV mostrado en el ANEXO 13, dando un enfoque de artículos para así garantizar un trabajo de calidad con ética moral y derechos de autor, logrando que pueda ser utilizado para extraer información en futuras investigaciones.

Por otro lado, se respeta la autoría de las fuentes consultadas, referenciando a los autores de las diferentes fuentes bibliográficas consultadas. Dichas referencias, se citan según el manual de la norma ISO 690 y 690-2 brindado por el Fondo Editorial de la Universidad César Vallejo.

La ética en un proyecto de investigación es considerada un tipo de ética práctica o aplicada, la cual consta de resolver problemas no necesariamente generales, sino, de carácter específico que surge en el desarrollo de la investigación. (Salazar et al., 2018).

IV. RESULTADOS

En este capítulo se menciona los resultados obtenidos de la investigación, en función a las métricas de rendimiento utilizadas en la presente investigación las mismas que son: sensibilidad, precisión especificidad, exactitud y F1 Score, las cuales fueron comparadas a través de 6 algoritmos los cuales son: Árbol de decisiones (DT), Random Forest (RF), máquina de vectores de soporte (SVM), Gradient Boosting Machine (GBM), K-vecino más cercano (K-NN) y Redes Neuronales (ANN), para determinar que algoritmo se adapta más a la predicción de la diabetes en personas mayores de edad y personas embarazadas.

Asimismo, esta sección comienza a través de la carga del conjunto de datos de Kaggle, que incluye una cantidad de 768 registros de personas con 9 características; número de embarazos, presión arterial diastólica, grosor de los pliegues de la piel, nivel de insulina, índice de masa corporal, antecedentes genéticos de diabetes, edad y resultado de diabetes) (sí/no), como se muestra en la Tabla2, usando la siguiente línea de código: data.info ().

Tabla N° 2: conjunto de datos de variables

Nº	COLUMNA	RECUENTO NO NULO	TIPO DE DATO
1	Embarazos	768 no nulo	Entero
2	Glucosa	768 no nulo	Entero
3	Presión arterial	768 no nulo	Entero
4	Grosor de la piel	768 no nulo	Entero
5	Insulina	768 no nulo	Decimal
6	IMC	768 no nulo	Decimal
7	DiabetespedegreeFunción	768 no nulo	flotador64
8	Edad	768 no nulo	Entero
9	Resultado	768 no nulo	Entero

Fuente: Elaboración propia.

Seguidamente, se realizó el análisis exploratorio de datos utilizando histogramas y bibliotecas para el proceso de limpieza, siendo de mucha ayuda por lo que permitió

identificar características que tienen valores cero y, a su vez, reemplazarlas con algún valor. Después de aplicar el histograma, el conjunto de datos consta de 268 diabéticos y 500 personas sin diabetes. Seguidamente se procedió a comprobar los valores estadísticos del conjunto de datos.

Una de las principales tareas en la detección y clasificación de la diabetes mediante modelos Machine Learning, es analizar cómo se relacionan entre sí las variables del conjunto de datos, para lo cual se utilizan técnicas de análisis de datos y herramientas de software. Para ello, se importó los datos en una base de datos en Microsoft SQL Server Management Studio la cual es una aplicación utilizada para la gestión y administración de los componentes dentro de SQL Server.

Asimismo, se utilizó el lenguaje de programación Python a través de la plataforma Jupyter la misma que se encuentra ubicada en la plataforma de código abierto más utilizada en la ciencia de datos y el aprendizaje automático, tal es el caso de Anaconda; todo ello, con el fin de importar el conjunto de datos, calcular la matriz de correlación y realizar la técnica en el análisis de correlación.

Posteriormente, se efectuó la conexión entre el motor de base de datos y la plataforma Jupyter, para efectuar el análisis exploratorio de los datos, obteniendo la matriz de correlación de los datos contenidos, de los cuales destacan algunos valores que superan la correlación promedio, la cual equivale a 0.5. Tal como es el caso entre edad y embarazos, donde se evidencia que el número de embarazos aumenta a medida que avanza la edad y se detiene a partir de cierta edad. Asimismo, se encuentra una correlación significativa entre la glucosa y la insulina, donde un aumento en los niveles de glucosa se asocia con una mayor probabilidad de diagnóstico de diabetes. Para la glucosa y la diabetes: cuanto mayor sea el nivel de glucosa, mayor será la cantidad de insulina necesaria para regularlo. Además, existe una relación entre el IMC y la grasa corporal: cuanto mayor es el IMC, mayor es el porcentaje de grasa del paciente.

Después de finalizar el análisis exploratorio del conjunto de datos, el siguiente paso fue ejecutar la capacitación. Este proceso comenzó dividiendo los datos en una proporción del 80% para el conjunto de entrenamiento y el 20% para el conjunto de prueba. Para esto utilizamos la biblioteca Sklearn.Model_selection.train_test_split() ya que esta biblioteca nos permite realizar el análisis simplemente especificando los tamaños de las pruebas. Al igual que en el análisis exploratorio, se observó que el conjunto de datos contiene datos nulos o faltantes. Por lo tanto, procedimos a corregirlos

eliminándolos del conjunto de datos, para lo cual utilizamos las librerías SimpleImputer() e impute.fit_transform (train, test), Seguidamente se realizó el procedimiento de Grid search, en la tabla 3 se presenta los hiperparámetros óptimos descubiertos utilizando el enfoque Grid search.

Tabla 3: Hiperparámetros para cada modelo utilizando grid search

Modelo	Hiperparámetros
GBM	'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 300
Árbol de decisión (DT)	max_depth: 3, criterion='entropy', min_samples_leaf: 1, min_samples_split: 2
Redes Neuronales (ANN)	['learning_rate_init'] ['hidden_layer_sizes']
K-vecino más cercano (K-NN)	'metric': 'euclidean', 'n_neighbors': 9
Máquina de vectores de soporte (SVM)	'C': 10, 'kernel': 'linear'
Random Forest	'ax_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100

Fuente: Elaboración propia.

A continuación, se expresa la validación de cada hipótesis mediante la matriz de confusión.

1. **Árbol de decisiones (DT)**

Tabla N° 4: Matriz de confusión – DT

Clase	1	2
1	88	10
2	35	21

Fuente: Elaboración propia.

Tabla N° 5: Matriz de observación – DT

Clases	Medidas			
	TP	TN	FP	FN

1	88	21	35	16
2	21	88	10	26

Fuente: Elaboración propia.

2. Máquina de vectores de soporte (SVM)

Tabla N° 6 Matriz de confusión – SVM

Clase	1	2
1	74	14
2	25	41

Fuente: Elaboración propia.

Tabla N° 7 Matriz de observación – SVM

Clases	Medidas			
	TP	TN	FP	FN
1	74	41	25	14
2	41	74	14	25

3. **Fuente:** Elaboración propia.

3. Random Forest (RF)

Tabla N° 8: Matriz de confusión – RF

Clase	1	2
1	98	10
2	22	24

Fuente: Elaboración propia.

Tabla N° 9: Matriz de observación – RF

Clases	Medidas			
	TP	TN	FP	FN

1	98	24	22	10
2	24	98	10	22

Fuente: Elaboración propia.

4. K-vecino más cercano (K-NN)

Tabla N° 10 Matriz de confusión – KNN

Clase	1	2
1	84	13
2	27	30

Fuente: Elaboración propia.

Tabla N° 11 Matriz de observación – KNN

Clases	Medidas			
	TP	TN	FP	FN
1	84	30	27	13
2	30	84	13	27

Fuente: Elaboración propia.

5. Redes Neuronales (ANN)

Tabla N° 12 Matriz de confusión – ANN

Clase	1	2
1	61	38
2	18	37

Fuente: Elaboración propia.

Tabla N° 13 Matriz de observación – ANN

Clases	Medidas			
	TP	TN	FP	FN

1	61	37	18	38
2	37	61	38	18

Fuente: Elaboración propia.

6. GBM

Tabla N° 14: Matriz de confusión – GBM

Clase	1	2
1	81	19
2	19	35

Fuente: Elaboración propia.

Tabla N° 15: Matriz de observación – GBM

Clases	Medidas			
	TP	TN	FP	FN
1	81	35	19	19
2	35	81	19	19

Fuente: Elaboración propia.

Hipótesis Específica 1: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con sensibilidad.

1. Árbol de decisión (DT)

Tabla N° 16 Cálculo de la sensibilidad con el algoritmo DT

Clases	$Sensibilidad = \frac{TP}{TP + FN} * 100$	Resultado
1	(88/(88+10)) * 100%	90.00%
2	(21/(21+35)) * 100%	48.50%
Total		70.80%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una sensibilidad de 70.80%, empleando el algoritmo Árbol de decisión.

2. Máquina de vectores de soporte (SVM)

Tabla N° 17 Cálculo de la sensibilidad con el algoritmo SVM

Clases	$Sensibilidad = \frac{TP}{TP + FN} * 100$	Resultado
1	(74/(74+14)) * 100%	84.09%
2	(41/(41+25)) * 100%	62.12%
Total		73.10%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una sensibilidad de 73.10%, empleando el algoritmo Máquina de vectores de soporte (SVM).

3. Random Forest (RF)

Tabla N° 18 Cálculo de la sensibilidad con el algoritmo Random Forest

Clases	$Sensibilidad = \frac{TP}{TP + FN} * 100$	Resultado
1	(98/(98+10)) * 100%	90.74/%
2	(24/(24+22)) * 100%	52.17%
Total		71.45%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una sensibilidad de 71.45%, empleando el algoritmo Random Forest.

4. K-vecino más cercano (K-NN)

Tabla N° 19 Cálculo de la sensibilidad con el algoritmo (K-NN)

Clases	$Sensibilidad = \frac{TP}{TP + FN} * 100$	Resultado
1	(84/(84+13)) * 100%	86.59%
2	(30/(30+27)) * 100%	52.63%
Total		69.61%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una sensibilidad de 69.61%, empleando el algoritmo KNN.

5. Redes Neuronales (ANN)

Tabla N° 20 Cálculo de la sensibilidad con el algoritmo (ANN)

Clases	$Sensibilidad = \frac{TP}{TP + FN} * 100$	Resultado
1	(61/(61+38)) * 100%	61.61%
2	(37/(37+18)) * 100%	67.27%
Total		64.44%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una sensibilidad de 64.44%, empleando el algoritmo ANN.

6. GBM

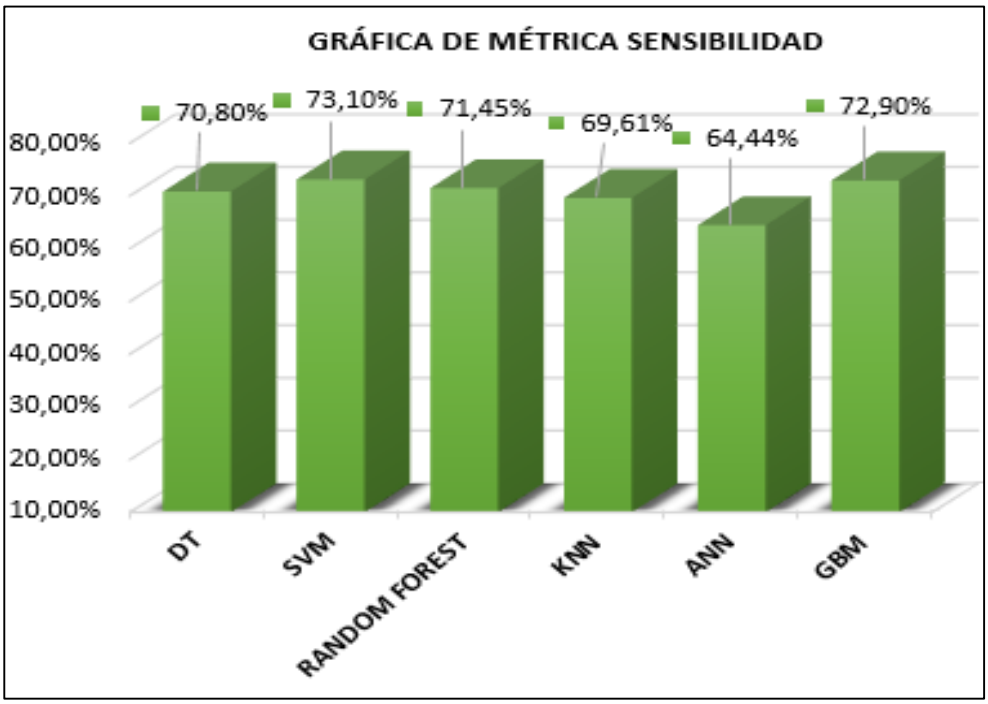
Tabla N° 21 Cálculo de la sensibilidad con el algoritmo GBM

Clases	$Sensibilidad = \frac{TP}{TP + FN} * 100$	Resultado
1	(81/(81+19)) * 100%	81.00%
2	(35/(35+19)) * 100%	64.81%
Total		72.90%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una sensibilidad de 72.90%, empleando el algoritmo GBM.

Figura N° 11: Resultados según la métrica de sensibilidad



Fuente: Elaboración propia.

Hipótesis Específica 1: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con precisión la diabetes.

1. **Árbol de decisión (DT)**

Tabla N° 22 Cálculo de la precisión con el algoritmo DT

Clases	$Presición = \frac{TP}{TP + FP} * 100$	Resultado
1	(88/(88+35)) * 100%	75.54%
2	(21/(21+10)) * 100%	69.80%
Total		74.09%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una precisión de 74.09%, empleando el algoritmo DT.

2. Máquina de vectores de soporte (SVM)

Tabla N° 23 Cálculo de la precisión con el algoritmo SVM

Clases	$Presición = \frac{TP}{TP + FP} * 100$	Resultado
1	(74/(74+25)) * 100%	74.74%
2	(41/(41+14)) * 100%	74.54%
Total		74.64%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una precisión de 74.64%, empleando el algoritmo SVM.

3. Random Forest (RF)

Tabla N° 24 Cálculo de la precisión con el algoritmo RF

Clases	$Presición = \frac{TP}{TP + FP} * 100$	Resultado
1	(98/(98+22)) * 100%	81.66%
2	(24/(24+10)) * 100%	70.58%
Total		%76.12

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una precisión de 76.12%, empleando el algoritmo RF.

4. K-vecino más cercano (K-NN)

Tabla N° 25 Cálculo de la precisión con el algoritmo KNN

Clases	$Presición = \frac{TP}{TP + FP} * 100$	Resultado
1	(84/(84+27)) * 100%	75.60%
2	(30/(30+13)) * 100%	69.76%
Total		72.72%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una precisión de 72.72%, empleando el algoritmo KNN.

5. Redes Neuronales (ANN)

Tabla N° 26 Cálculo de la precisión con el algoritmo ANN

Clases	$Presición = \frac{TP}{TP + FP} * 100$	Resultado
1	(61/(61+18)) * 100%	77.21%
2	(37/(37+38)) * 100%	49.50%
Total		63.27%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una precisión de 63.27%, empleando el algoritmo ANN.

6. GBM

Tabla N° 27 Cálculo de la precisión con el algoritmo GBM

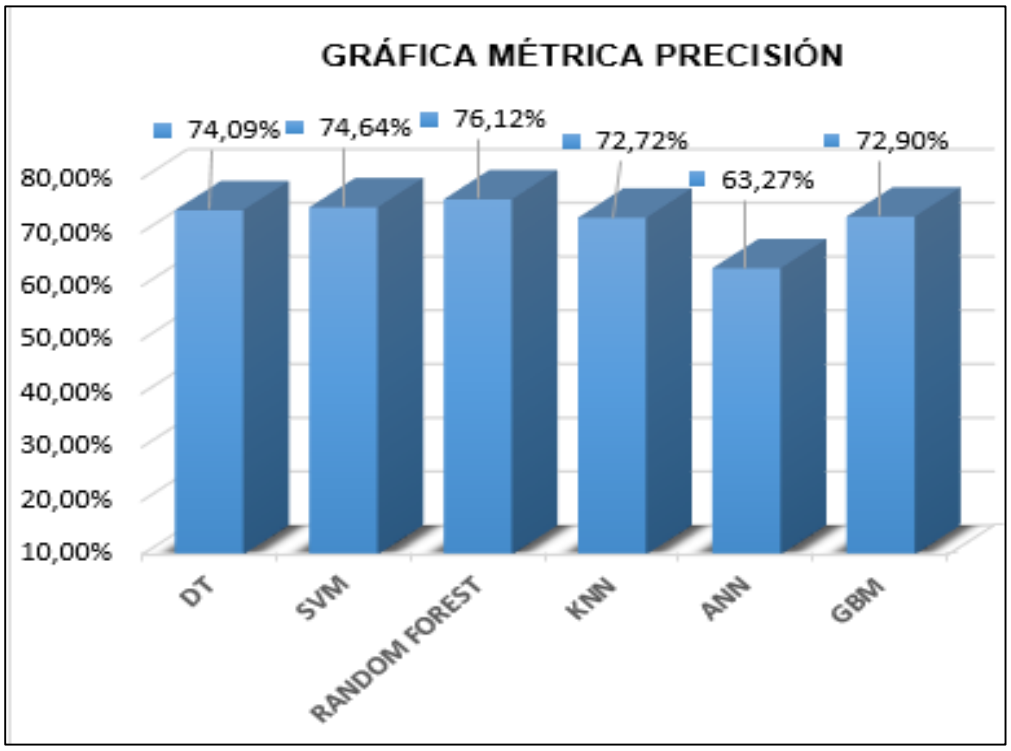
Clases	$Presición = \frac{TP}{TP + FP} * 100$	Resultado
1	(81/(81+19)) * 100%	81.00%

2	(35/(35+19)) * 100%	64.81%
Total		72.90%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predece la diabetes con una precisión de 72.90%, empleando el algoritmo GBM.

Figura N° 12: Resultados según la métrica de precisión



Fuente: Elaboración propia.

Hipótesis Específica 2: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá con especificidad la diabetes.

1. **Árbol de decisión (DT)**

Tabla N° 28 Cálculo de la especificidad con el algoritmo DT

Clases	$Especificidad = \frac{TN}{TN+FP} * 100\%$	Resultado

1	(21/(21+35)) * 100%	55.50%
2	(88/(88+10)) * 100%	89.79%
Total		70.80%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una especificidad de 70.80%, empleando el algoritmo DT.

2. Máquina de vectores de soporte (SVM)

Tabla N° 29 Cálculo de la especificidad con el algoritmo SVM

Clases	$Especificidad = \frac{TN}{TN + FP} * 100$	Resultado
1	(41/(41+25)) * 100%	62.12%
2	(74/(74+14)) * 100%	84.09%
Total		73.10%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una especificidad de 73.10%, empleando el algoritmo SVM.

3. Random Forest (RF)

Tabla N° 30 Cálculo de la especificidad con el algoritmo RF

Clases	$Especificidad = \frac{TN}{TN + FP} * 100$	Resultado
1	(41/(41+25)) * 100%	62.12%
2	(74/(74+14)) * 100%	84.09%
Total		71.45%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una especificidad de 71.45%, empleando el algoritmo RF.

4. K-vecino más cercano (K-NN)

Tabla N° 31 Cálculo de la especificidad con el algoritmo KNN

Clases	$Especificidad = \frac{TN}{TN + FP} * 100$	Resultado
1	(30/(30+27)) * 100%	52.63%
2	(84/(84+13)) * 100%	86.59%
Total		69.61%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una especificidad de 69.61%, empleando el algoritmo KNN.

5. Redes Neuronales (ANN)

Tabla N° 32 Cálculo de la especificidad con el algoritmo ANN

Clases	$Especificidad = \frac{TN}{TN + FP} * 100$	Resultado
1	(37/(37+18)) * 100%	67.27%
2	(61/(61+38)) * 100%	61.61%
Total		64.44%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una especificidad de 64.44%, empleando el algoritmo ANN.

6. GBM

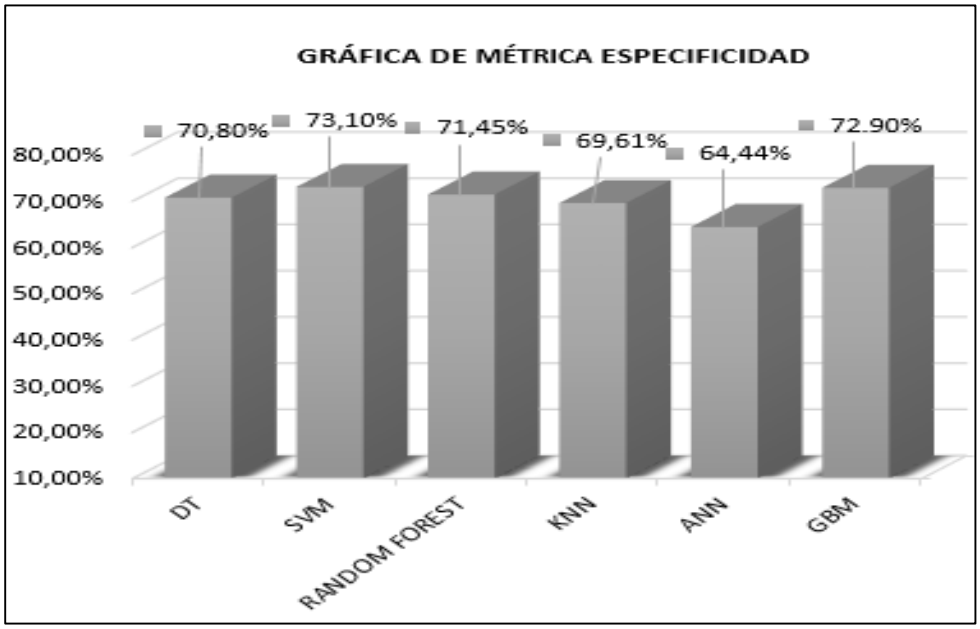
Tabla N° 33 Cálculo de la especificidad con el algoritmo GBM

Clases	$Especificidad = \frac{TN}{TN + FP} * 100$	Resultado
1	(35/(35+19)) * 100%	77.66%
2	(81/(81+19)) * 100%	78.00%
Total		72.90%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una especificidad de 78.66%, empleando el algoritmo GBM.

Figura N° 13: Resultados según la métrica de especificidad



Fuente: Elaboración propia.

Hipótesis Específica 3: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con exactitud la diabetes.

1. Árbol de decisión (DT)

Tabla N° 34 Cálculo de la exactitud con el algoritmo DT

Clases	$Exactitud = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$	Resultado
1	(88+21/(88+21+35+16)) * 100%	73.68%
2	(21+88/(21+88+10+26)) * 100%	74.00%
Total		74.68%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una exactitud de 74.68%, empleando el algoritmo DT.

2. Máquina de vectores de soporte (SVM)

Tabla N° 35 Cálculo de la exactitud con el algoritmo DT

Clases	$Exactitud = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$	Resultado
1	(74+41/(74+41+25+14)) * 100%	74.68%
2	(41+74/(41+74+14+25)) * 100%	74.67%
Total		74.67%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una exactitud de 74.67%, empleando el algoritmo SVM.

3. Random Forest (RF)

Tabla N° 36 Cálculo de la exactitud con el algoritmo RF

Clases	$Exactitud = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$	Resultado
1	(98+24/(98+24+22+10)) * 100%	79.10%
2	(24+98/(24+98+10+22)) * 100%	78.90%
Total		79.22%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una exactitud de 79.22%, empleando el algoritmo RF.

4. K-vecino más cercano (K-NN)

Tabla N° 37 Cálculo de la exactitud con el algoritmo KNN

Clases	$Exactitud = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$	Resultado
1	(84+30/(84+30+27+13)) * 100%	73.50%
2	(30+84/(30+84+13+27)) * 100%	74.00%
Total		74.02%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una exactitud de 74.02%, empleando el algoritmo KNN.

5. Redes Neuronales (ANN)

Tabla N° 38 Cálculo de la exactitud con el algoritmo ANN

Clases	$Exactitud = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$	Resultado
1	(61+37/(61+37+18+38) * 100%	63.50%
2	(37+61/(37+61++38+18)) * 100%	63.00%

Total	63.63%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una exactitud de 63.63%, empleando el algoritmo ANN.

6. GBM

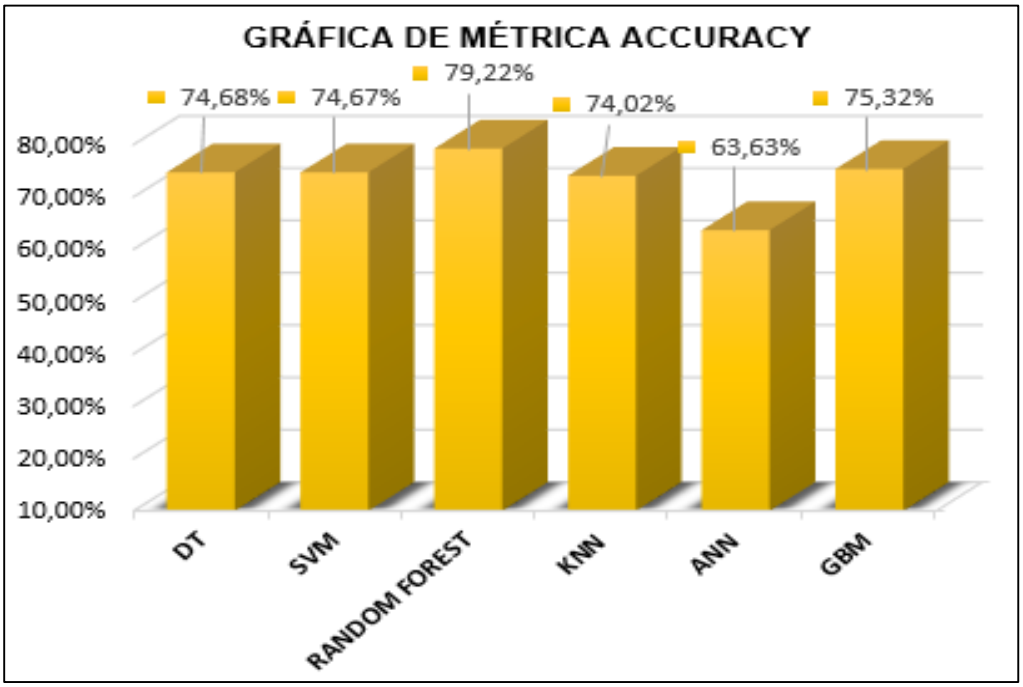
Tabla N° 39 Cálculo de la exactitud con el algoritmo GBM

Clases	$Exactitud = \frac{TP + TN}{TP + TN + FP + FN} * 100$	Resultado
1	(81+35/(81+35+19+19) * 100%	75.32%
2	(35+81/(35+81+19+19)) * 100%	75.32%
Total		75.32%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con una exactitud de 75.32%, empleando el algoritmo GBM.

Figura N° 14: Resultados según la métrica de ACCURACY



Fuente: Elaboración propia.

Hipótesis Específica 4: El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con accuracy la diabetes.

1. Árbol de decisión (DT)

Tabla N° 40 Cálculo del F1 Score con el algoritmo DT

Clases	$F1 - Score = 2\,x\,\frac{Precisión\ x\ Recall}{Precisión + Recall} * 100\%$	Resultado
1	2 * ((75.54* 90.00) / (75.54+ 90.00) * 100	82.13%
2	2 * ((69.80* 48.50) / (69.80+ 48.50) * 100	55.50%
Total		71.55%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con F1-score de 71.55%, empleando el algoritmo DT.

2. Máquina de vectores de soporte (SVM)

Tabla N° 41 Cálculo del F1 Score con el algoritmo SVM

Clases	$F1 - Score = 2\,x\,\frac{Precisión\ x\ Recall}{Precisión + Recall} * 100\%$	Resultado
1	2 * ((74.74* 84.09) / (74.74+ 84.09) * 100	67.76%
2	2 * ((74.54* 62.12) / (74.54+ 62.12) * 100	79.13%
Total		73.45%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con F1-score de 73.45%, empleando el algoritmo SVM.

3. Random Forest (RF)

Tabla N° 42 Cálculo del F1 Score con el algoritmo RF

Clases	$F1 - Score = 2\ x\ \frac{Precisión\ x\ Recall}{Precisión + Recall} * 100\%$	Resultado
1	2 * ((81.66* 90.74) / (81.66+ 90.74) * 100	85.96%
2	2 * ((70.58* 52.17) / (70.58+ 52.17) * 100	59,99%
Total		72.98%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con F1-score de 72.98%, empleando el algoritmo RF.

4. K-vecino más cercano (K-NN)

Tabla N° 43 Cálculo de F1 Score con el algoritmo KNN

Clases	$F1 - Score = 2\ x\ \frac{Precisión\ x\ Recall}{Precisión + Recall} * 100\%$	Resultado
1	2 * ((75.60* 86.59) / (75.60+ 86.59) * 100	80.72%
2	2 * ((69.76* 52.63) / (69.76+ 52.63) * 100	59.99%
Total		70.38%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con F1-score de 70.38%, empleando el algoritmo KNN.

5. Redes Neuronales (ANN)

Tabla N° 44 Cálculo de F1 score con el algoritmo ANN

Clases	$F1-Score=\ 2\ x\ \frac{Precisión\ x\ Recall}{Precisión\ +Recall} * 100$	Resultado
1	2 * ((77.21* 61.61) / (77.21+ 61.61) * 100	68.30%
2	2 * ((49.50* 67.27) / (49.50+ 67.27) * 100	57.03%
Total		62.73%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con F1-score de 62.73%, empleando el algoritmo ANN.

6. GBM

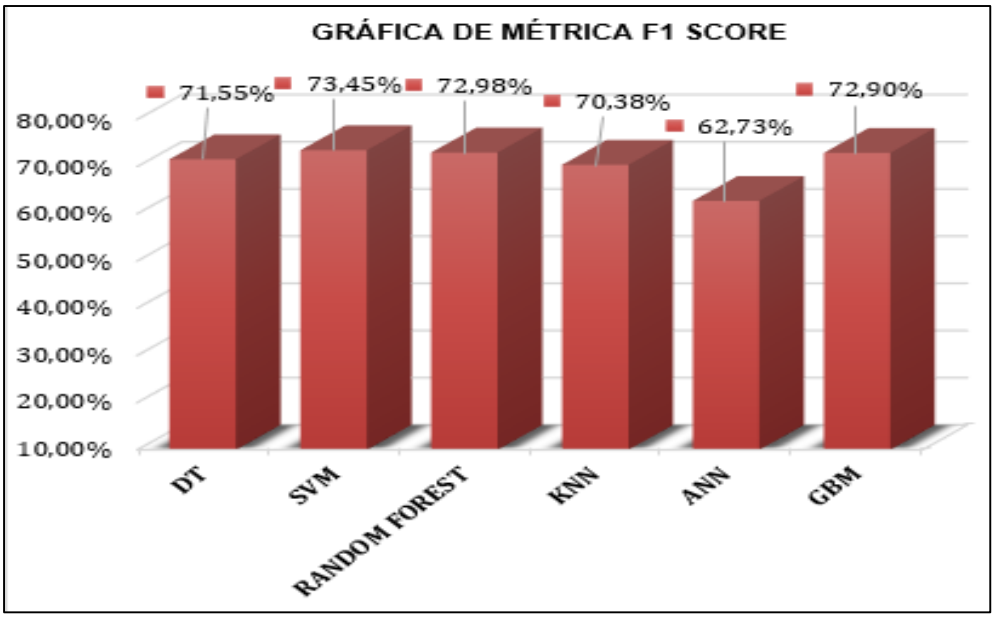
Tabla N° 45 Cálculo de la F1 Score con el algoritmo GBM

Clases	$F1-Score=\ 2\ x\ \frac{Precisión\ x\ Recall}{Precisión\ +Recall} * 100$	Resultado
1	2 * ((81.00* 81.00) / (81.00+ 81.00) * 100	77.10%
2	2 * ((64.81* 64.81) / (64.81+ 64.81) * 100	78.00%
Total		72.90%

Fuente: Elaboración propia.

El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predecirá la diabetes con F1-score de 72.90%, empleando el algoritmo GBM.

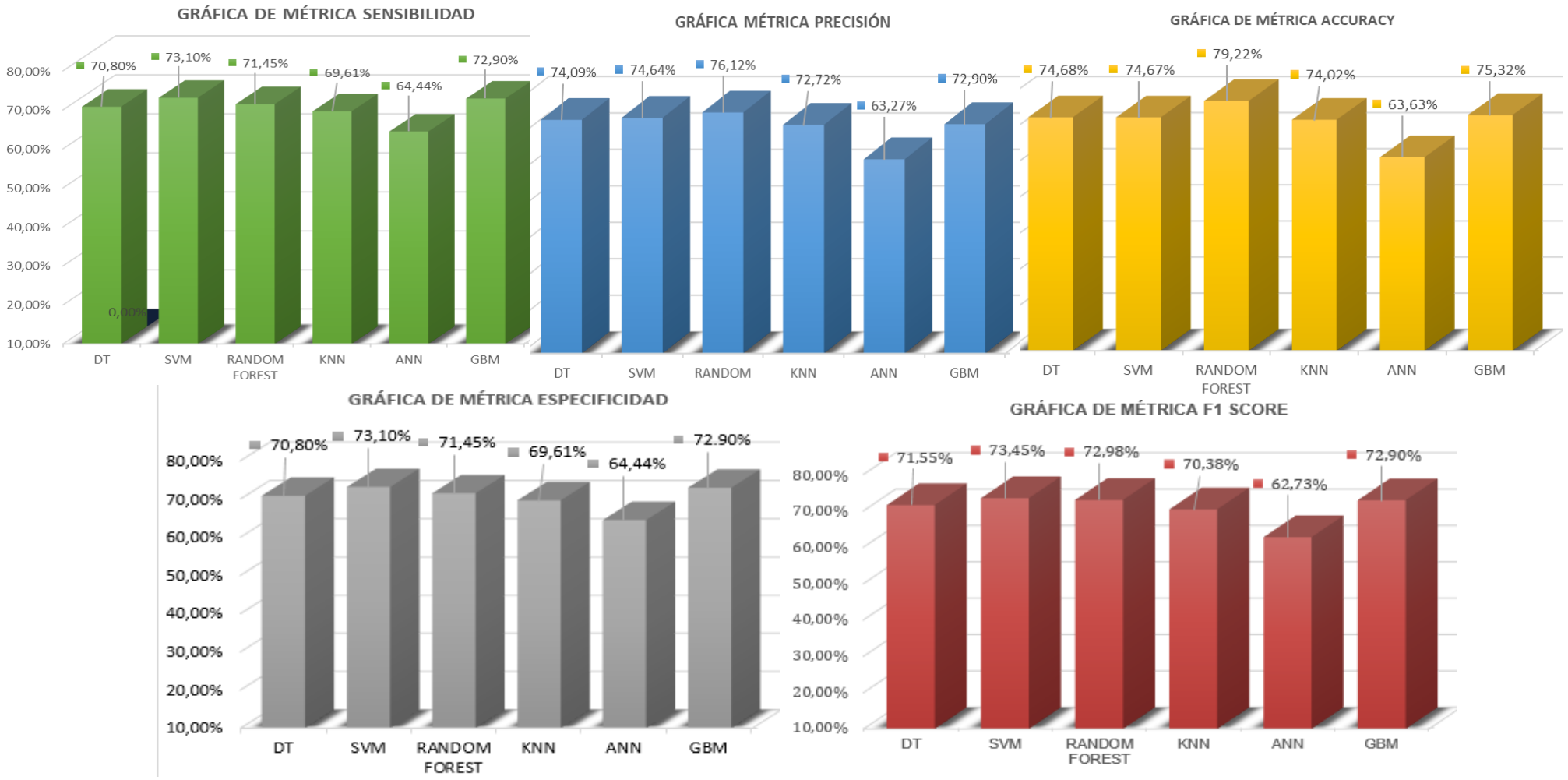
Figura N° 15: Resultado general según la métrica de F1-score



Fuente: Elaboración propia.

Hipótesis Especifica 5: La aplicación de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con F1-score la diabetes.

Figura N° 16: Resultados según las métricas



Fuente: Elaboración propia.

Interpretación: En el presente grafico se puede observar que los algoritmos con mejores resultados respecto todos los indicadores o métricas fueron el algoritmo Random Forest (RF) con un 79.22% de exactitud, posteriormente el algoritmo Gradient Boosting Machine (GBM) con un 75.32 % de exactitud y por último el algoritmo árbol de decisiones (DT), con un 74.09% en cuanto a la precisión.

V. DISCUCIÓN

En el presente capítulo, se observan las discusiones que se basan en los resultados obtenidos durante la presente investigación en función a las evidencias obtenidas en el referido proceso, para posteriormente efectuar la comparación e interpretación de los resultados; así como su importancia en la relevancia de la presente investigación.

La detección y clasificación de la diabetes es un problema para la ciencia médica, hay muchos algoritmos de Machine Learning que se utilizan para abordar este problema. En la presente investigación consideró como objetivo principal Aplicar un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes, basado en variables como el número de embarazos, presión arterial diastólica, grosor de los pliegues de la piel, nivel de insulina, índice de masa corporal, antecedentes genéticos de diabetes, edad y resultado de diabetes) (sí/no). utilizando la metodología Knowledge Discovery in Databases (KDD), método por el cual se realiza el análisis de datos a través de modelos de aprendizaje automático para la predicción mediante algoritmos, por lo que Autores como Aldair Aquino et al. (2018), Silvia Haedo et al. (2016), Gordillo et al. (2020), García Dionicio et al. (2021), utilizaron esta metodología como referencia para la extracción de conocimiento y minería de datos.

Luego de revisar las Técnicas de Machine Learning y los trabajos relacionados con las comparaciones efectuadas a través de distintos algoritmos, en base los datos contenidos en la base de datos de los indios pima, se eligieron los algoritmos más relacionados con las fuentes de información para realizar el contraste en base a los objetivos determinados para la presente investigación.

Con relación al objetivo específico 1, se aplicó un análisis comparativo entre modelos de aprendizaje automático (ML) para predecir la diabetes utilizando 6 algoritmos; DT, RF, SVM, ANN, KNN y GBM, basándose en sus mejores hiperparámetros, obteniendo como mejor resultado en función a la sensibilidad el modelo SVM y GBM, el SVM con tuvo 73.10 % y GBM con un 72.90 %, por lo que se puede considerar que es bueno para la predicción de la diabetes, de igual forma tiene similitud con la investigaciones del Autores como; Leon Kopitar, Leona Cilar y Gregorio Stiglic et al. (2022), donde en su estudio para la detección temprana

de la diabetes utilizaron modelos de predicción ML, alcanzando buenos resultados el modelo SVM con 0.72% y Light GBM Con un resultado de 0,76 %.

Se puede señalar que los siguientes autores (Zhao et al. 2023), (Li, Zheng y Songbai et al. 2022) (Rajput y Khedgikar 2022), no consideraron esta métrica en sus investigaciones, ya que se orientaron en similares medidas de accuracy y Rcall.

Con relación al objetivo específico 2, se aplicó un análisis comparativo entre modelos de aprendizaje automático (ML) para predecir la diabetes utilizando; DT, RF, SVM, ANN, KNN y GBM, basándose en sus mejores hiperparámetros, obteniendo como mejor resultado en función a la precisión el modelo RF y SVM, el modelo RF tuvo un 76.12 % y el modelo SVM 74.64 % por lo que se puede validar que es bueno para la predicción de la diabetes, por lo que también tiene similitud con la investigación del Autor; Amin Mansoori et al. (2023), .que utilizó modelos RL, DT, RF, SVM, según los índices de rendimiento el modelo RF proporcionó una mejor precisión de 80.43% en la predicción de la diabetes.

Se puede señalar que, los siguientes autores (Herleen Kaur y Vinita Kumai et al. 2018), Toktam Sahranavard et al. (2023) Sara Saffar Soflae et al. (2023), consideraron otras métricas en sus investigaciones, ya que se orientaron en similares mediciones de especificidad.

Con relación al objetivo específico 3, se aplicó un análisis comparativo entre modelos de aprendizaje automático (ML) para predecir la diabetes utilizando; DT, RF, SVM, ANN, KNN y GBM, basándose en sus mejores hiperparámetros, obteniendo como mejor resultado en función a la especificidad el modelo GBM y SVM, el modelo SVM tuvo un 74.64 % y SVM un 73.10%, por lo que se puede validar que es bueno para la predicción de la diabetes, por lo que también tiene similitud con la investigación del Autor; Gamboa Cruzado y Augusto Hidalgo et al. (2023), que mostro como técnica más utilizada a los modelos SVM y KNN.

Se puede señalar que, los siguientes autores (Herleen Kaur y Vinita Kumai et al. 2018), Toktam Sahranavard et al. (2023) Sara Saffar Soflae et al. (2023),

consideraron otras métricas en sus investigaciones, ya que se orientaron en similares mediciones de especificidad.

Con relación al objetivo específico 4, se aplicó un análisis comparativo entre modelos de aprendizaje automático (ML) para predecir la diabetes utilizando; DT, RF, SVM, ANN, KNN y GBM, basándose en sus mejores hiperparámetros, obteniendo como mejor resultado en función a la exactitud el modelo RF y SVM, el modelo RF tuvo un 79.22 % y GBM un 75.32%, por lo que se puede validar que es bueno para la predicción de la diabetes, por lo que también tiene similitud con la investigación de Autores como; Raiput et al. (2023) y Amin Mansoori et al. (2023), que utilizaron modelos RL, DT, RF, SGB y Neive Bayes, según los resultados el modelo Random Forest obtuvo una exactitud de 0.79% en la predicción de la diabetes.

Se puede señalar que, los siguientes autores Rabia Emhamed y Mamlook et al. 2022), Toktam Sahranavard et al. (2023), Robert Sawyeret et al. (2023), consideraron otras métricas en sus investigaciones, ya que se orientaron en similares mediciones de especificidad de modelos de Machine Learning, F1 Score y AUC.

Con relación al objetivo específico 5, se aplicó un análisis comparativo entre modelos de aprendizaje automático (ML) para predecir la diabetes utilizando; DT, RF, SVM, ANN, KNN y GBM, basándose en sus mejores hiperparámetros, obteniendo como mejor resultado en función al F1 Score, el modelo GBM y SVM, el modelo GBM tuvo un 72.90 % y SVM un 73.45%, por lo que se puede validar que es bueno para la predicción de la diabetes, por lo que también tiene similitud con la investigación de los siguientes Autores; Leon Kopitar et al. (2023), Savitesh Kushwaha et al. (2022), Jing Li et al. (2020), utilizaron modelos SVM con un 0,859% y XGBoost con un 0,881%, Rondom Forest con 0.842% y Light GBM con un resultado de 0,846 % en función al F1 Score. Se puede señalar que, los siguientes autores Rachana Srivastava et al. (2022), Arun Kumar Aggarwal et al. (2023), Rajput y Khedgikar et al. (2023), consideraron otras métricas en sus investigaciones, ya que se orientaron en similares mediciones de precisión, AUC y sensibilidad.

VI. CONCLUSIÓN

Basándonos en los resultados obtenidos en la investigación presentamos las siguientes conclusiones:

Primero: El estudio se basó en la comparación de técnicas de Machine Learning para la prevención de la diabetes a través de seis tipos de modelos K-NN, DT, ANN, GBM, SVM y Random Forest, utilizando comandos de Python para descubrir cuál es el mejor algoritmo para el conjunto de datos, durante el desarrollo de los modelos Machine Learning, se utilizó Grid search para encontrar los hiperparámetros óptimos. Los modelos se entrenaron y se probaron usando el conjunto de datos de prueba y entrenamiento, también se utilizaron cinco métricas de rendimiento para la predicción de la diabetes, precisión, sensibilidad, especificidad, exactitud y F1 Score, así como también se utilizó La metodología Knowledge Discovery in Databases (KDD), que fue aplicada en el estudio permitiendo el descubrimiento de conocimientos útiles a partir de datos.

Segundo: Asimismo, aplicando la comparación entre los diferentes algoritmos mencionados anteriormente, los modelos de machine learning que resultaron con mayor porcentaje con respecto a la métrica especificidad fue SVM y GBM alcanzando conseguir un 73.10% y 72.90%, este resultado superó a los algoritmos como DT, ANN, KNN y Random Forest, asimismo, esta métrica no fue de mucha utilidad en otras investigaciones. Es preciso mencionar que los porcentajes pueden variar dependiendo de la cantidad de datos estudiados.

Tercero: En ese sentido, aplicando la comparación entre los diferentes algoritmos mencionados anteriormente, los modelos de machine learning que resultaron con mayor porcentaje en cuanto a la métrica de precisión fue Random Forest y SVM destacando resultados favorables, logrando conseguir un 76.12% y 74.64%, este resultado superó a los resultados de otros algoritmos como DT, ANN, KNN y Light GBM. La exclusión de esta métrica por algunos autores destaca la importancia de evaluar múltiples aspectos para obtener una comprensión detallada del rendimiento del modelo por lo que es considerada muy útil. Cabe mencionar que los

porcentajes pueden variar dependiendo de la cantidad de datos estudiados.

Cuarto: En ese sentido, aplicando la comparación entre los diferentes algoritmos mencionados anteriormente, los modelos de machine learning que resultaron con mayor porcentaje en cuanto a la métrica Especificidad fue el GBM y SVM destacando resultados favorables logrando conseguir un 78.66% y 73.10%, este resultado superó a los resultados de otros algoritmos como DT, ANN, KNN y Random Forest, Es preciso señalar que esta métrica no fue considerada en su mayoría por otros autores, lo que a través de los resultado implica su importancia para los modelos predictivos. Cabe señalar que los porcentajes pueden variar dependiendo de la cantidad de datos estudiados.

Quinto: Del mismo modo, aplicando la comparación entre los diferentes algoritmos mencionados anteriormente, los modelos de machine learning que resultaron con mayor porcentaje en cuanto a la métrica Exactitud fue GBM y Random Forest destacando resultados con mayor porcentaje, logrando conseguir un 79.22% y 75.32%, este resultado superó a los resultados de otros algoritmos como DT, ANN, KNN y SVM Es preciso señalar que esta métrica es utilizado por distintos autores en sus investigaciones por la precisión en su mayoría de los modelos de Machine Learning, asimismo es preciso señalar que los porcentajes pueden variar dependiendo de la cantidad de datos estudiados.

Sexto: En ese contexto, aplicando la comparación entre los diferentes algoritmos mencionados anteriormente, los modelos de machine learning que resultaron con mayor porcentaje en cuanto a la métrica F1 Score fueron GBM y SVM destacando resultados favorables, logrando conseguir un 78.66% y 73.45%, este resultado superó a los resultados de otros algoritmos como DT, ANN, KNN y Random Forest. Es preciso señalar que esta métrica no fue considerada en su mayoría por otros autores y es preciso señalar que los porcentajes pueden variar dependiendo de la cantidad de datos estudiados.

Séptimo: Finalmente, se concluye que, de los resultados obtenidos del estudio, en base a los seis modelos, el algoritmo Random Forest (RF) obtuvo un mayor resultado de 79,22%, seguidamente Gradient Boosting Machine (GBM) obtuvo un 75,32%, en cuanto a la métrica exactitud y el árbol de decisiones (DT) obtuvo un 74.09% en cuanto a la precisión, superando a los demás modelos, siendo identificados estos 3 modelos como los más cercanos a la predicción de la diabetes. Por lo tantos, estos algoritmos son los más eficientes, lo que demuestra que el uso del aprendizaje automático (machine learning), es indispensable en el sector de salud en beneficio para la sociedad y siendo apoyo fundamental en los diagnósticos clínicos. En ese sentido, este estudio ayudará a que futuros investigadores elaboren una mejor técnica, generando de una forma simple, directa y confiable se pueda predecir la diabetes.

VII. RECOMENDACIONES

En esta parte, seguido de las conclusiones, se menciona las recomendaciones para las futuras investigaciones en el sector salud.

Primero: Se recomienda seguir con el desarrollo de modelos predictivos usando distintos algoritmos de Machine Learning para trabajos futuros, utilizando metodologías como Knowledge Discovery in Databases (KDD), considerando distintos enfoques de minería de datos, con el fin de identificar anticipadamente la predicción de la diabetes ya que estos estudios allanarán el camino para que futuros investigadores proporcionen una mejor técnica generando una forma simple, económica, directa y confiable de predecir la diabetes.

Segundo: De igual forma, se recomienda ampliar los estudios de otros algoritmos que podrían no haberse incluido en la investigación en función a la métrica de especificidad, así como también establecer los alcances y posibles limitaciones de los algoritmos, esto podría proporcionar información adicional sobre enfoques alternativos y sus ventajas o desventajas en un modelo predictivo.

Tercero: Del mismo modo, se recomienda proponer investigaciones basadas en el desarrollo de algoritmos que permitan afinar el diagnóstico de la Diabetes empleando las mismas o un mayor número de variables, con el fin de brindar mayor precisión que a su vez sea usada en beneficio no solo del diagnóstico inmediato de la enfermedad, sino del estudio de los factores que intervienen en la misma.

Cuarto: Asimismo, es recomendable que para futuras investigaciones se sugiera realizar una comparación detallada similares u diferentes a los algoritmos estudiados, con un enfoque específico en la métrica de sensibilidad, exactitud y F1 Score, esta evaluación adicional posibilitará la obtención de una perspectiva más sólida sobre enfoques alternativos en el contexto del modelo desarrollado, proporcionando así una comprensión más clara de su aplicabilidad y sus posibles restricciones sus futuros estudios, de tal forma que podrá allanar el camino para que otros investigadores proporcionen

una mejor técnica con una forma simple, económica, directa y confiable de predecir la diabetes, basada en atributos médicos.

Quinto: Asimismo, es recomendable que se pueda ampliar la investigación con diferentes bases de datos, de esta manera poder obtener diferentes resultados en función a la metodología utilizada en la presente investigación para que se pueda obtener resultados en mejora a la predicción de la diabetes.

Sexto: Se recomienda que se considere el estudio en diferentes metodologías de investigación que puedan concluir diferentes resultados en post de la mejorar y la predicción de la diabetes; así como implementar otros algoritmos que puedan darle robustez al sistema.

REFERENCIAS

ABDULKADIR, O. y DERBEW, G., 2023. Application of supervised machine learning algorithms for classification and prediction of type-2 diabetes disease status in Afar regional state, Northeastern Ethiopia 2021. *Scientific reports*, vol. 13, no. 1, ISSN 20452322. DOI 10.1038/s41598-023-34906-1.

ABDULRAHMAN, H., 2019. You Don't Know SVD (Singular Value Decomposition). [en línea]. [consulta: 16 junio 2023]. Disponible en: https://towardsdatascience.com/svd-8c2f72e264f.

ALEGRE-DÍAZ, J., HERRINGTON, W., LÓPEZ-CERVANTES, M., GNATIUC, L., RAMIREZ, R., HILL, M., BAIGENT, C., MCCARTHY, M.I., LEWINGTON, S., COLLINS, R., WHITLOCK, G., TAPIA-CONYER, R., PETO, R., KURI-MORALES, P. y EMBERSON, J.R., 2019. Diabetes and Cause-Specific Mortality in Mexico City. *New England Journal of Medicine*, vol. 375, no. 20, ISSN 0028-4793. DOI 10.1056/nejmoa1605368.

American Diabetes Association. [en línea], 2022. [consulta: 28 abril 2023]. Disponible en: https://diabetes.org/.

AMIN, M.M.S., 2021. Developing A Machine Learning Based Prognostic Model and A Supporting Web-based Application for Predicting The Possibility of Early Diabetes and Diabetic Kidney Disease. S.l.

BERGMAN, R.N., STEFANOVSKI, D. y KIM, S.P., 2019. Systems analysis and the prediction and prevention of Type 2 diabetes mellitus. *Current Opinion in Biotechnology*, vol. 28, ISSN 18790429. DOI 10.1016/j.copbio.2014.05.007.

DAMI, A., 2021. Revisión sistemática de literatura: Técnicas de aprendizaje automático (machine learning).

FEBRIAN, M.E., FERDINAN, F.X., SENDANI, G.P., SURYANIGRUM, K.M. y YUNANDA, R., 2023. Diabetes prediction using supervised machine learning. *Procedia Computer Science*, vol. 216, ISSN 18770509. DOI 10.1016/j.procs.2022.12.107.

GANDHI, R., 2018. Support Vector Machine — Introduction to Machine Learning Algorithms. [en línea]. [consulta: 15 junio 2023]. Disponible en: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

GARCÍA-LAENCINA, P.J., VERDÚ-MONEDERO, R., LARREY-RUIZ, J., MORALES-SÁNCHEZ, J. y SANCHO-GÓMEZ, J.-L., 2017. Algoritmo KNN basado en Información Mutua para Clasificación de Patrones con Valores Perdidos. [en línea]. S.l.: Disponible en: http://www.isical.ac.in/.

GONZÁLES, L., 2020. *TÉCNICAS E INSTRUMENTOS DE INVESTIGACIÓN CIENTÍFICA ENFOQUES CONSULTING EIRL* [en línea]. S.l.: s.n. ISBN 9786124844409. Disponible en: www.cienciaysociedad.org.

GONZÁLEZ, L., 2023. Aprendizaje Supervisado: Linear Regresión. [en línea]. [consulta: 15 junio 2023]. Disponible en: https://aprendeia.com/algoritmo-regresion-lineal-simple-machine-learning/.

GUEVARA ALBAN, G., VERDESOTO ARGUELLO, A. y CASTRO MOLINA, N., 2020. Metodologías de investigación educativa (descriptivas, experimentales, participativas, y de investigación-acción). *RECIMUNDO* [en línea], [consulta: 29 junio 2023]. ISSN 2588-073X. DOI 10.26820/recimundo/4. (3). julio.2020.163-173. Disponible en: http://recimundo.com/index.php/es/article/view/860.

HERNÁNDEZ SAMPIERI, R., FERNÁNDEZ COLLADO, C. y BAPTISTA LUCIO, P., 2004. METODOLOGÍA DE LA INVESTIGACIÓN. S.l.

INTERNATIONAL DIABETES FEDERATION, 2021. IDF Diabetes Atlas 10th edition. [en línea], [consulta: 18 abril 2023]. Disponible en: www.diabetesatlas.org.

KOPITAR, L., KOCBEK, P., CILAR, L., SHEIKH, A. y STIGLIC, G., 2020b. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific Reports*, vol. 10, no. 1, ISSN 20452322. DOI 10.1038/s41598-020-68771-z.

KRASTEVA, Assya, PANOV, V., KRASTEVA, Adriana, KISSELOVA, A. y KRASTEV, Z., 2020. *Oral cavity and systemic diseases - Diabetes mellitus*. febrero 2018. S.l.: s.n.

KUMAR DEWANGAN, A. y AGRAWAL, P., 2022. Classification of Diabetes Mellitus Using Machine Learning Techniques. [en línea]. S.l.: Disponible en: http://www.ics.uci.edu/~mlearn/databases/thyroid.

LI, J., XU, Z., XU, T. y LIN, S., 2022. Predicting Diabetes in Patients with Metabolic Syndrome Using Machine-Learning Model Based on Multiple Years' Data. *Diabetes, Metabolic Syndrome and Obesity*, vol. 15, ISSN 11787007. DOI 10.2147/DMSO.S381146.

LÓPEZ BRIEGA, R., 2018. Libro online de IAAR - Machine Learning. [en línea]. [consulta: 16 junio 2023]. Disponible en: https://iaarbook.github.io/machine-learning/.

MANAGEMENT SOLUTIONS ESPAÑA, 2018. Machine Learning, una pieza clave en la transformación de los modelos de negocio. S.l.

MANSOORI, A., SAHRANAVARD, T., HOSSEINI, Z.S., SOFLAEI, S.S., EMRANI, N., NAZAR, E., GHARIZADEH, M., KHORASANCHI, Z., EFFATI, S., GHAMSARY, M., FERNS, G., ESMAILY, H. y MOBARHAN, M.G., 2023. Prediction of type 2 diabetes mellitus using hematological factors based on machine learning approaches: a cohort study analysis. *Scientific Reports*, vol. 13, no. 1, ISSN 20452322. DOI 10.1038/s41598-022-27340-2.

MORALES, P., 2019. Tipos de variables y sus implicaciones en el diseño de una investigación. [en línea]. S.l.: [consulta: 30 junio 2023]. Disponible en: https://gc.scalahed.com/recursos/files/r161r/w25732w/morales.pdf.

MUJUMDAR, A. y VAIDEHI, V., 2019. Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*. S.l.: Elsevier B.V., pp. 292-299. vol. 165. DOI 10.1016/j.procs.2020.01.047.

ORDÓÑEZ, D.A. y VIZCARRA, E.R., 2018. Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics. S.l.: s.n. ISBN 0000000276399.

PINEDA-JARAMILLO, J.D., 2019. A review of machine learning (ML) algorithms used for modeling travel mode choice•. *DYNA (Colombia)*, vol. 86, no. 211, ISSN 00127353. DOI 10.15446/dyna.v86n211.79743.

RAJPUT, M.R. y KHEDGIKAR, S.S., 2022. Diabetes prediction and analysis using medical attributes: A Machine learning approach. [en línea], DOI 10.37896/JXAT14.01/314405. Disponible en: https://www.researchgate.net/publication/357648143.

ROMAN, V., 2019. Supervised Learning: Basics of Classification and Main Algorithms | by Victor Roman | Towards Data Science. [en línea]. [consulta: 15 junio 2023]. Disponible en: https://towardsdatascience.com/supervised-learning-basics-of-classification-and-main-algorithms-c16b06806cd3.

RUSSELL, R., 2018. Machine Learning Guía Paso a Paso Para Implementar Algoritmos De Machine Learning Con Python. [en línea]. S.l.: Disponible en: www.detodopython.com.

RUSSELL, S. y NORVIG, P., 2018. Artificial Intelligence a Modern Approach. [en línea]. S.l.: Disponible en: www.PlentyofeBooks.net.

SALAMANCA, I., 2021. Técnicas de aprendizaje automático aplicadas en los sistemas de predicción | Tecnología Investigación y Academia. *Tecnología Investigación y Academia*, vol. 8, no. 1,

SÁNCHEZ CARLESSI, H., ROMERO REYES, C. y MEJÍA SÁENZ, K., 2018. Manual de términos en investigación científica, tecnológica y humanística. S.l.

SANDOVAL, J.L., 2018. ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA ANÁLISIS Y PREDICCIÓN DE DATOS., vol. 11,

SHALEV-SHWARTZ, S. y BEN-DAVID, S., 2018. Understanding Machine Learning: From Theory to Algorithms. [en línea]. S.l.: Disponible en: http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning.

STAWARZ, K., KATZ, D., AYOBI, A., MARSHALL, P., YAMAGATA, T., SANTOS-RODRIGUEZ, R., FLACH, P. y O'KANE, A.A., 2023. Co-designing opportunities for Human-Centred Machine Learning in supporting Type 1 diabetes decision-making. *International Journal of Human Computer Studies*, vol. 173, ISSN 10959300. DOI 10.1016/j.ijhcs.2023.103003.

TEODORO, N. y NIETO, E., 2018. TIPOS DE INVESTIGACIÓN. [en línea]. S.l.: [consulta: 29 junio 2023]. Disponible en: https://d1wqtxts1xzle7.cloudfront.net/99846223/250080756-libre.pdf?1678813555=&response-content-disposition=inline%3B+filename%3DTipos_de_Investigacion.pdf&Expires=1688176550&Signature=YQwQkB0u9~cZzzLwzhQ70o07wfThfR0Yme4k~xSQpRkifg5hDPvdj8TovoF3MBVCMN8uvFZaN205O0HUzQb5mvAjM~SOf85K8a1On2hTMtdju4swHl37Y4byN9KvoRIMuNuAN9cwUSnhOp1MHftQ9ncqoim5SAgEU-xt39QtJ0IdrGrOdCnpS0XsvN5K2U22n69HYFfhGvmCp6u9VtEgNrMp5vQPCdyfPBTE6FFQYJf6doPysVSIPvDPLhjoQvd11LvbDzIDvC3hfq0plxL4QF4yKjx-j6WQ3T8NT1LZyBXZUMu66C6kmIK~u68KhCu55HdqqYREnTxfKdxdoqVdqw&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

VEGA-MALAGÓN, G., ÁVILA-MORALES, J., VEGA-MALAGÓN, A.J., CAMACHO-CALDERÓN, N., BECERRIL-SANTOS, A. y LEO-AMADOR, G., 2014. PARADIGMAS EN LA INVESTIGACIÓN. ENFOQUE CUANTITATIVO Y CUALITATIVO. [en línea]. S.l.: [consulta: 29 junio 2023]. Disponible en: https://core.ac.uk/reader/236413540#related-papers.

ZAPETA HERNÁNDEZ, A., GALINDO ROSALES, G.A., JUAN SANTIAGO, H.J. y MARTÍNEZ LEE, M., 2022. Métricas de rendimiento para evaluar el aprendizaje automático en la clasificación de imágenes petroleras utilizando redes neuronales convolucionales. *Ciencia Latina Revista Científica Multidisciplinar*, vol. 6, no. 5, ISSN 2707-2207. DOI 10.37811/cl_rcm.v6i5.3420.

ZHAO, Q., ZHU, J., SHEN, X., LIN, C., ZHANG, Y., LIANG, Y., CAO, B., LI, J., LIU, X., RAO, W. y WANG, C., 2023. Chinese diabetes datasets for data-driven machine learning. *Scientific Data*, vol. 10, no. 1, ISSN 20524463. DOI 10.1038/s41597-023-01940-7.

ANEXOS

ANEXO Nº 1

Tabla N° 46 TABLA DE OPERACIONALIZACIÓN DE VARIABLES

VARIABLES	DEFINICIÓN CONCEPTUAL	DEFINICIÓN OPERACIONAL	Dimensiones (sub variables)	INDICADORES	ESCALA DE MEDICIÓN
Independiente: Técnicas de Maching learning	Las técnicas son aquellas que se enfocan en predecir una respuesta cualitativa o cualitativa mediante el análisis de datos y el reconocimiento de patrones (Salamanca 2021)**.** El machine learning, común mente abreviado como (ML), es un tipo de Inteligencia artificial (IA) que identifica patrones de entrada de datos y contiene algoritmos que evolucionan con el tiempo (Dami 2021)**.**	Es el que se usa para comprender métodos y obtener un modelamiento de conocimiento mediante el aprendizaje. Machine Learning está inmerso en muchas áreas, las que mayormente destacan son las siguientes, transporte autónomo, seguridad informática, genética y su finalidad es reconocer o pronosticar patrones complejos fundados en data. - Algoritmos - Técnicas - Grandes	- Algoritmos - Técnicas - Grandes cantidades de información - Entrenamiento de algoritmo		De Razón

VARIABLES	DEFINICIÓN CONCEPTUAL	DEFINICIÓN OPERACIONAL	Dimensiones (sub variables)	INDICADORES	ESCALA DE MEDICIÓN
Dependiente: Predicción de la diabetes	La predicción de la diabetes se define a la necesidad de predecir correctamente el desarrollo de esta enfermedad para permitir la intervención y, por lo tanto, retrasar la progresión de la enfermedad y el trastorno metabólico resultante en el paciente. (Bergman, Stefanovski y Kim, 2018)**.**	La predicción de la diabetes se medirá a través de las métricas de precisión, especificidad y sensibilidad, las mismas que serán obtenidas a partir de las herramientas y técnicas consideradas en el Machine Learning.	Métricas de Evaluación de Modelo	Precisión (TP/(TP+FP)) * 100 Sensibilidad (TP/(TP+FN)) * 100 Especificidad (TN/(TN+FP)) * 100 Exactitud (TP+TN) / (TP+TN+FP+FN) F1 Score 2 * ((Precisión * Sensibilidad) / (Precisión + Sensibilidad))	Razón

Fuente: Elaboración propia.

ANEXO N°2:

Tabla N° 47 MATRIZ DE CONSISTENCIA

PROBLEMA	OBJETIVO	HIPÓTESIS	VARIABLES	METODOLOGÍA	JUSTIFICACIÓN
PROBLEMA GENERAL:	**OBJETIVO GENERAL:**	**HIPÓTESIS GENERAL:**	**VARIABLE INDEPENDIENTE:**		
¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo permitirá predecir la diabetes?	Aplicar un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes.	La aplicación de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice la diabetes.	Técnicas de Maching learning	**Tipo de investigación:** Aplicada **Diseño de Investigación:** Preexperimental **Población:** Todas las personas con la enfermedad de diabetes Mujeres de al menos 21 años de herencia indígena pima, la cual consta de 768 registros **La Muestra:** Tomando en cuenta un nivel de confianza del 95% y un margen de error del 5% la muestra es: 257 registros.	El proyecto baja su justificación en que, luego de comparar las distintas técnicas de machine learning a través de algoritmos y métodos, se podrá establecer y determinar cuál es el más adecuado para el manejo de la información acerca de los factores que determinan en el diagnóstico de la diabetes, de tal forma que se podrá establecer una identificación temprana evitando de esta manera, los problemas colaterales que esta ocasiona dicha enfermedad.

PROBLEMA	OBJETIVO	HIPÓTESIS	VARIABLES	METODOLOGÍA	JUSTIFICACIÓN
PROBLEMAS ESPECÍFICOS:	**OBJETIVO ESPECÍFICO:**	**HIPÓTESIS ESPECÍFICA:**	**VARIABLE DEPENDIENTE**	**Tipo de diseño:**	Social: dado que, debido a que dicha enfermedad está relacionada al deterioro de varios órganos, su predicción evitará que el tratamiento sea una preocupación, así como reducir las atenciones médicas relacionadas con esta enfermedad, se fomentará la conciencia social para la reducción de los casos. **Económico**: de justifica en que nuestro trabajo, facilitará el reconocimiento de los factores sintomatológicos
1. ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función al Precisión permitirá predecir la diabetes?	1. Utilizar análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes en función a la precisión.	1. El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con precisión a la diabetes.	Predicción de la diabetes	Técnicas e Instrumentos de recolección de Datos: Encuestas Fichas de observación	
2. ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función a la Especificidad permitirá predecir la diabetes?	2. Usar análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes en función a la sensibilidad.	2. El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con sensibilidad la diabetes.			
3. ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función a la	3. Emplear un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la	3. El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice			

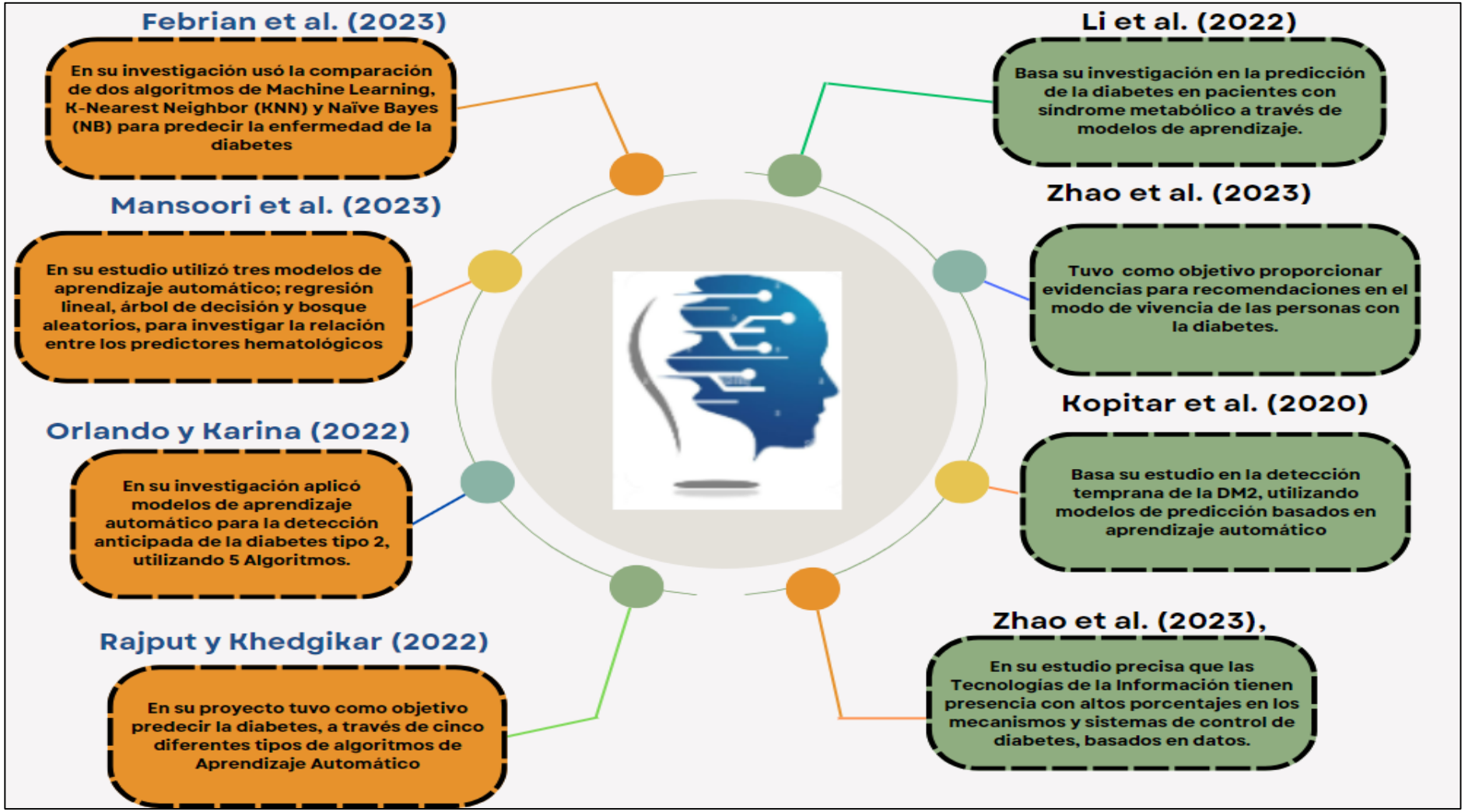
Sensibilidad permitirá predecir la diabetes?	diabetes en función a la especificidad	con especificidad la diabetes.			de la diabetes melitos, de tal forma que se pueda evitar los largos y costosos tratamientos que acarrea el padecimiento de esta enfermedad. También, se justifica tecnológicamente por la implementación de un sistema predictivo que contribuirá con las investigaciones futuras para el desarrollo de nuevas técnicas de desarrollo.
4. ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función a la Exactitud permitirá predecir la diabetes?	4. Emplear un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes en función a Exactitud	4. El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con Exactitud la diabetes.			
5. ¿Cómo un análisis comparativo de técnicas de Machine Learning sobre el método de muestreo en función a F1 Score permitirá predecir la diabetes?	5. Emplear un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo para predecir la diabetes en función a F1 Score.	5. El empleo de un análisis comparativo con técnicas de Machine Learning sobre el método de muestreo predice con F1 Score la diabetes.			

Tabla 47. MATRIZ DE CONSISTENCIA

Fuente: Elaboración propia.

Anexo 3:

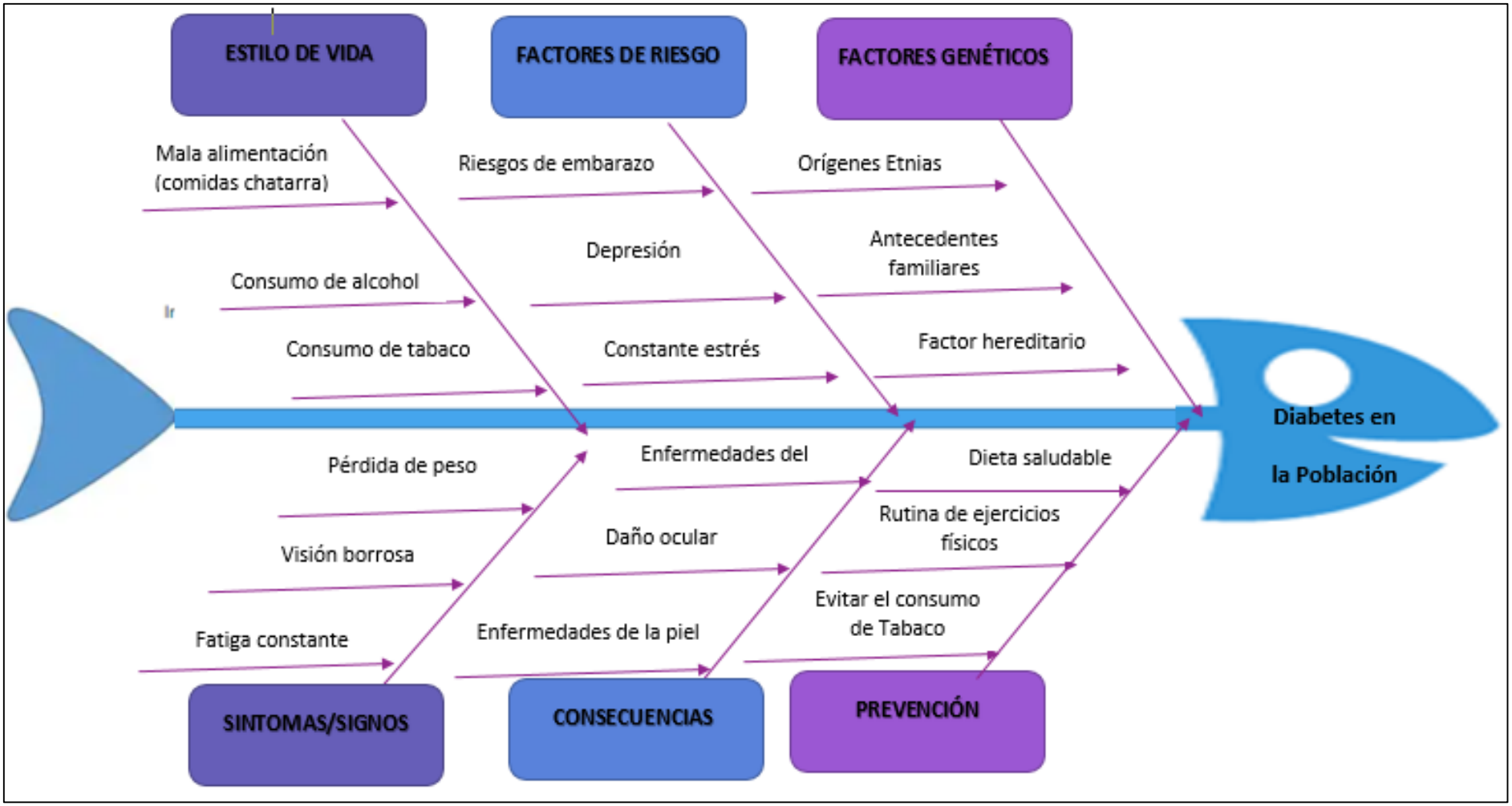
FIGURA Nª 17: MAPA CONCEPTUAL DE ANTECEDENTES



Fuente: Elaboración propia.

ANEXO N° 4

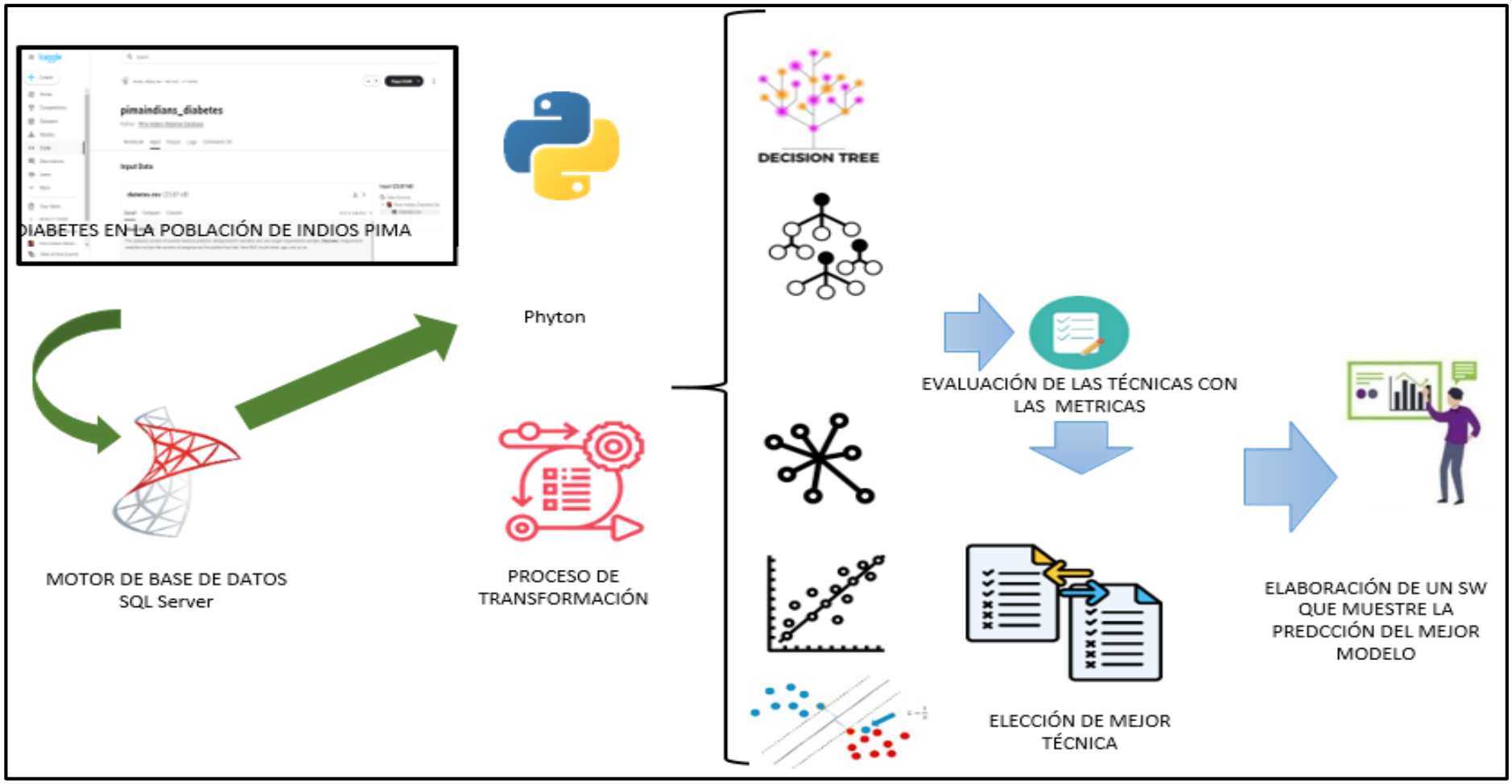
FIGURA N° 18: DIAGRAMA DE HISHIKAWA



Fuente: Elaboración propia.

ANEXO N° 5

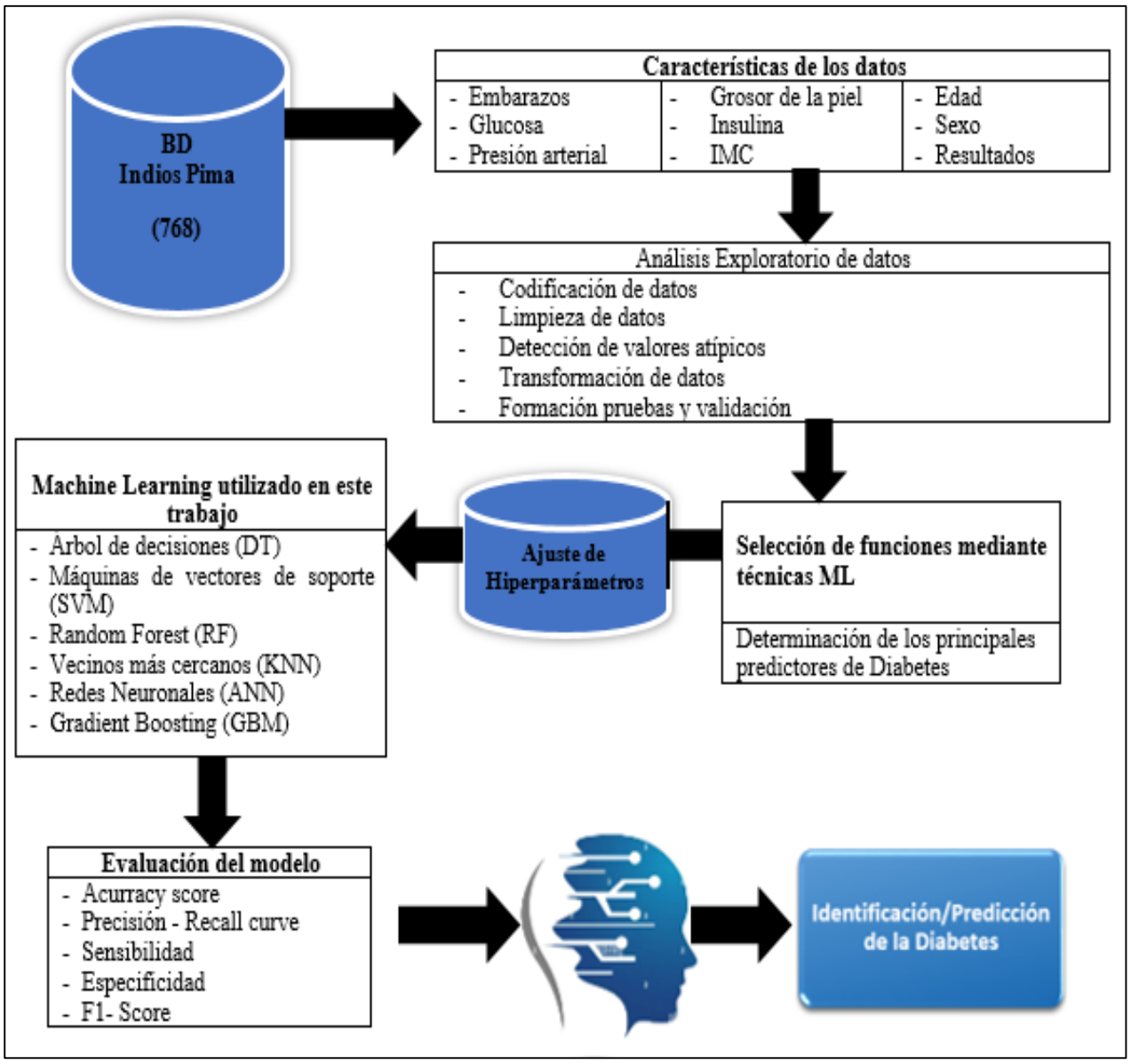
FIGURA Nº 19: PROTOTIPO DEL SISTEMA



Fuente: Elaboración propia.

ANEXO N° 6

FIGURA N° 20: PROCESO DEL MODELO DE DATOS



Fuente: Elaboración propia.

ANEXO Nº 7: INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO DT

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Post Test
Investigadores	- Chira Bohórquez Piero - Rivera Munive Kevin
Fecha de inicio	11/09/2023
Algoritmo	DT

MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	**Positivo (P)**
Real	**Negativo**	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	**Indicador**	**Medida**	**Fórmula**	**Precisión**
1	Exactitud	Razón	(TP+TN)/(TP+TN+FP+FN)	74.68%
2	Precisión	Razón	(TP/(TP+FP)) * 100	74.09%
3	Sensibilidad	Razón	(TP/(TP+FN)) * 100	70.80%
4	Especificidad	Razón	(TN/(TN+FP)) * 100	70.80%
5	F1 Score	Razón	2 * ((Precisión * Sensibilidad) / (Precisión + Sensibilidad))	71.55%

Fuente: Elaboración propia.

ANEXO Nº 8: INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO RANDOM FOREST

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Post Test
Investigadores	- Chira Bohorquez Piero - Rivera Munive Kevin
Fecha de inicio	11/09/2023
Algoritmo	Random Forest

MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	**Positivo (P)**
Real	**Negativo**	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	**Indicador**	**Medida**	**Fórmula**	**Precisión**
1	Exactitud	Razón	(TP+TN)/(TP+TN+FP+FN)	79.22%
2	Precisión	Razón	(TP/(TP+FP)) * 100	76.12%
3	Sensibilidad	Razón	(TP/(TP+FN)) * 100	71.45%
4	Especificidad	Razón	(TN/(TN+FP)) * 100	71.45%
5	F1 Score	Razón	2 * ((Precisión * Sensibilidad) / (Precisión + Sensibilidad))	72.98%

Fuente: Elaboración propia.

ANEXO Nº 9: INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO K-NEAREST NEIGHBOOR (KNN)

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Post Test
Investigadores	- Chira Bohorquez Piero - Rivera Munive Kevin
Fecha de inicio	11/09/2023
Algoritmo	KNN

MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	**Positivo (P)**
Real	**Negativo**	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	**Indicador**	**Medida**	**Fórmula**	**Precisión**
1	Exactitud	Razón	(TP+TN)/(TP+TN+FP+FN)	74.02%
2	Precisión	Razón	(TP/(TP+FP)) * 100	72.72%
3	Sensibilidad	Razón	(TP/(TP+FN)) * 100	69.61%
4	Especificidad	Razón	(TN/(TN+FP)) * 100	69.02%
5	F1 Score	Razón	2 * ((Precisión * Sensibilidad) / (Precisión + Sensibilidad))	70.38%

Fuente: Elaboración propia.

Anexo Nº 10: INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO GBM

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Post Test
Investigadores	- Chira Bohorquez Piero - Rivera Munive Kevin
Fecha de inicio	11/09/2023
Algoritmo	GBM

MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	**Positivo (P)**
Real	**Negativo**	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	**Indicador**	**Medida**	**Fórmula**	**Precisión**
1	Exactitud	Razón	(TP+TN)/(TP+TN+FP+FN)	75.32%
2	Precisión	Razón	(TP/(TP+FP)) * 100	72.90%
3	Sensibilidad	Razón	(TP/(TP+FN)) * 100	72.90%
4	Especificidad	Razón	(TN/(TN+FP)) * 100	78.66%
5	F1 Score	Razón	2 * ((Precisión * Sensibilidad) / (Precisión + Sensibilidad))	78.66%

Fuente: Elaboración propia.

ANEXO Nº 11: INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO SUPPORT VECTOR MACHINES SVM

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Post Test
Investigadores	- Chira Bohorquez Piero - Rivera Munive Kevin
Fecha de inicio	11/09/2023
Algoritmo	Support Vector Machines

MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	**Positivo (P)**
Real	**Negativo**	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	**Indicador**	**Medida**	**Fórmula**	**Precisión**
1	Exactitud	Razón	(TP+TN)/(TP+TN+FP+FN)	74.67%
2	Precisión	Razón	(TP/(TP+FP)) * 100	74.67%
3	Sensibilidad	Razón	(TP/(TP+FN)) * 100	74.67%
4	Especificidad	Razón	(TN/(TN+FP)) * 100	74.67%
5	F1 Score	Razón	2 * ((Precisión * Sensibilidad) / (Precisión + Sensibilidad))	74.67%

Fuente: Elaboración propia.

ANEXO Nº 12: INTRUMENTO DE FICHA DE REGISTRO PARA EL ALGORITMO ANN

RESUMEN DE EVALUACIÓN	
Tipo de Prueba	Post Test
Investigadores	- Chira Bohorquez Piero - Rivera Munive Kevin
Fecha de inicio	11/09/2023
Algoritmo	ANN

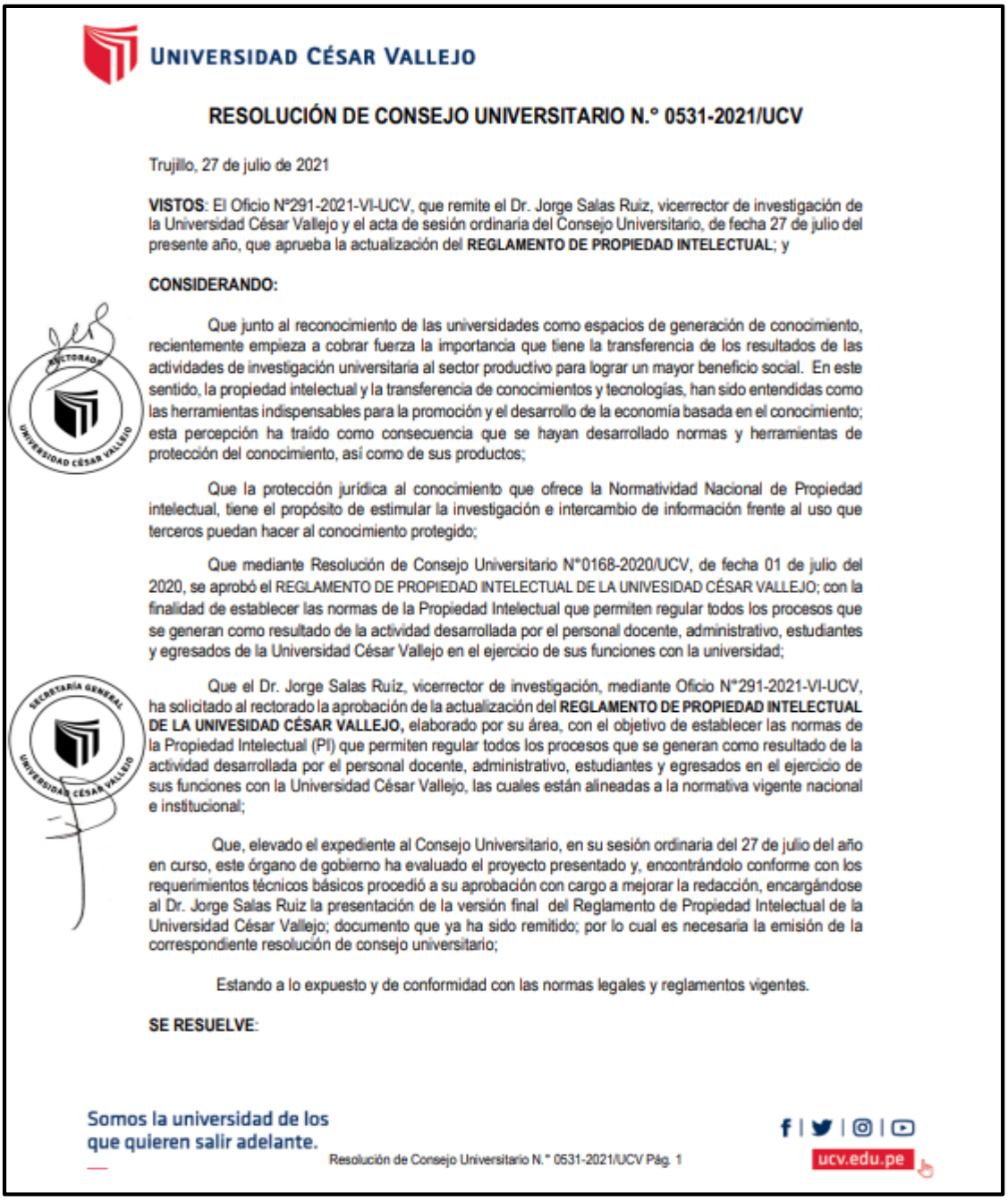
MATRIZ DE CONFUSIÓN			
		Estimación por el modelo	
		Negativo (N)	**Positivo (P)**
Real	**Negativo**	TN	FP
	Positivo	FN	TP

MÉTRICAS POR EVALUAR				
Ítem	**Indicador**	**Medida**	**Fórmula**	**Precisión**
1	Exactitud	Razón	(TP+TN)/(TP+TN+FP+FN)	63.63%
2	Precisión	Razón	(TP/(TP+FP)) * 100	63.27%
3	Sensibilidad	Razón	(TP/(TP+FN)) * 100	64.44%
4	Especificidad	Razón	(TN/(TN+FP)) * 100	74.67%
5	F1 Score	Razón	2 * ((Precisión * Sensibilidad) / (Precisión + Sensibilidad))	72.73%

Fuente: Elaboración propia.

ANEXO N° 13:

FIGURA 21: RESOLUCIÓN DEL CONSEJO UNIVERSITARIO



UNIVERSIDAD CÉSAR VALLEJO

RESOLUCIÓN DE CONSEJO UNIVERSITARIO N.° 0531-2021/UCV

Trujillo, 27 de julio de 2021

VISTOS: El Oficio N°291-2021-VI-UCV, que remite el Dr. Jorge Salas Ruiz, vicerrector de investigación de la Universidad César Vallejo y el acta de sesión ordinaria del Consejo Universitario, de fecha 27 de julio del presente año, que aprueba la actualización del **REGLAMENTO DE PROPIEDAD INTELECTUAL**; y

CONSIDERANDO:

Que junto al reconocimiento de las universidades como espacios de generación de conocimiento, recientemente empieza a cobrar fuerza la importancia que tiene la transferencia de los resultados de las actividades de investigación universitaria al sector productivo para lograr un mayor beneficio social. En este sentido, la propiedad intelectual y la transferencia de conocimientos y tecnologías, han sido entendidas como las herramientas indispensables para la promoción y el desarrollo de la economía basada en el conocimiento; esta percepción ha traído como consecuencia que se hayan desarrollado normas y herramientas de protección del conocimiento, así como de sus productos;

Que la protección jurídica al conocimiento que ofrece la Normatividad Nacional de Propiedad intelectual, tiene el propósito de estimular la investigación e intercambio de información frente al uso que terceros puedan hacer al conocimiento protegido;

Que mediante Resolución de Consejo Universitario N°0168-2020/UCV, de fecha 01 de julio del 2020, se aprobó el REGLAMENTO DE PROPIEDAD INTELECTUAL DE LA UNIVESIDAD CÉSAR VALLEJO; con la finalidad de establecer las normas de la Propiedad Intelectual que permiten regular todos los procesos que se generan como resultado de la actividad desarrollada por el personal docente, administrativo, estudiantes y egresados de la Universidad César Vallejo en el ejercicio de sus funciones con la universidad;

Que el Dr. Jorge Salas Ruiz, vicerrector de investigación, mediante Oficio N°291-2021-VI-UCV, ha solicitado al rectorado la aprobación de la actualización del **REGLAMENTO DE PROPIEDAD INTELECTUAL DE LA UNIVESIDAD CÉSAR VALLEJO,** elaborado por su área, con el objetivo de establecer las normas de la Propiedad Intelectual (PI) que permiten regular todos los procesos que se generan como resultado de la actividad desarrollada por el personal docente, administrativo, estudiantes y egresados en el ejercicio de sus funciones con la Universidad César Vallejo, las cuales están alineadas a la normativa vigente nacional e institucional;

Que, elevado el expediente al Consejo Universitario, en su sesión ordinaria del 27 de julio del año en curso, este órgano de gobierno ha evaluado el proyecto presentado y, encontrándolo conforme con los requerimientos técnicos básicos procedió a su aprobación con cargo a mejorar la redacción, encargándose al Dr. Jorge Salas Ruiz la presentación de la versión final del Reglamento de Propiedad Intelectual de la Universidad César Vallejo; documento que ya ha sido remitido; por lo cual es necesaria la emisión de la correspondiente resolución de consejo universitario;

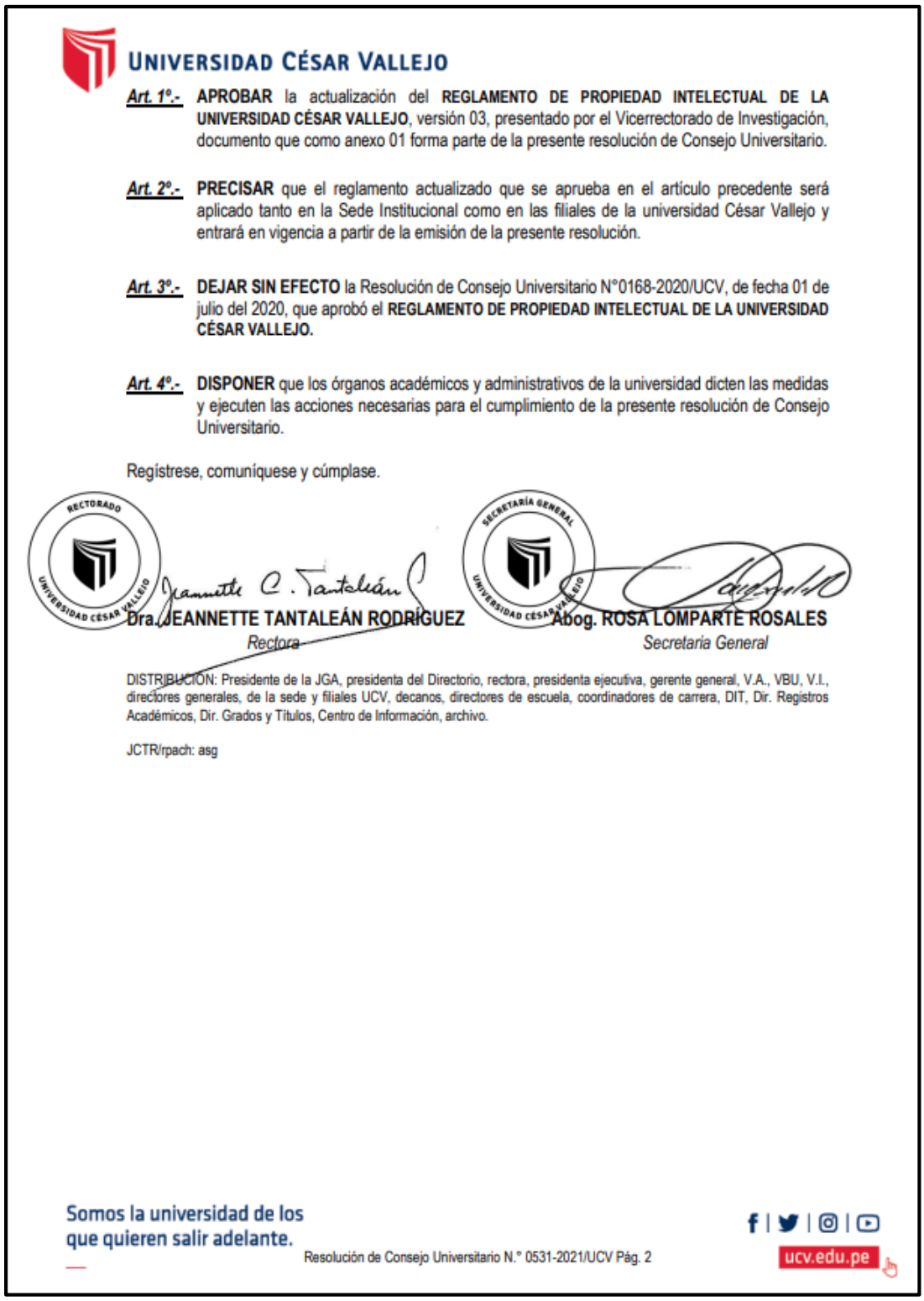
Estando a lo expuesto y de conformidad con las normas legales y reglamentos vigentes.

SE RESUELVE:

Somos la universidad de los que quieren salir adelante.

Resolución de Consejo Universitario N.° 0531-2021/UCV Pág. 1

ucv.edu.pe



UNIVERSIDAD CÉSAR VALLEJO

Art. 1°.- **APROBAR** la actualización del **REGLAMENTO DE PROPIEDAD INTELECTUAL DE LA UNIVERSIDAD CÉSAR VALLEJO**, versión 03, presentado por el Vicerrectorado de Investigación, documento que como anexo 01 forma parte de la presente resolución de Consejo Universitario.

Art. 2°.- **PRECISAR** que el reglamento actualizado que se aprueba en el artículo precedente será aplicado tanto en la Sede Institucional como en las filiales de la universidad César Vallejo y entrará en vigencia a partir de la emisión de la presente resolución.

Art. 3°.- **DEJAR SIN EFECTO** la Resolución de Consejo Universitario N°0168-2020/UCV, de fecha 01 de julio del 2020, que aprobó el **REGLAMENTO DE PROPIEDAD INTELECTUAL DE LA UNIVERSIDAD CÉSAR VALLEJO.**

Art. 4°.- **DISPONER** que los órganos académicos y administrativos de la universidad dicten las medidas y ejecuten las acciones necesarias para el cumplimiento de la presente resolución de Consejo Universitario.

Regístrese, comuníquese y cúmplase.

RECTORADO UNIVERSIDAD CÉSAR VALLEJO

Dra. JEANNETTE TANTALEÁN RODRÍGUEZ
Rectora

SECRETARÍA GENERAL UNIVERSIDAD CÉSAR VALLEJO

Abog. ROSA LOMPARTE ROSALES
Secretaria General

DISTRIBUCIÓN: Presidente de la JGA, presidenta del Directorio, rectora, presidenta ejecutiva, gerente general, V.A., VBU, V.I., directores generales, de la sede y filiales UCV, decanos, directores de escuela, coordinadores de carrera, DIT, Dir. Registros Académicos, Dir. Grados y Títulos, Centro de Información, archivo.

JCTR/rpach: asg

Somos la universidad de los que quieren salir adelante.

Resolución de Consejo Universitario N.° 0531-2021/UCV Pág. 2

ucv.edu.pe

Fuente: Elaboración propia.

ANEXO Nº 14: VALIDACIÓN DEL INSTRUMENTO

I. DATOS GENERALES:

Apellidos y Nombres del experto: Daza Vergaray Alfredo
Título y/o grado: Dr. Ingeniería de Sistemas
Fecha: 02/12/23
Instrumento: Cuestionario
Autor: Chira Bohorquez Piero Alejandro
Rivera Munive Kevin
Título de la investigación:
Análisis comparativo de técnicas de Machine Learning sobre el método de muestreo para la predicción de diabetes

II. ASPECTOS DE VALIACIÓN:

INDICADORES	CRITERIOS	DEFICIENTE 0-20%	REGULAR 21-50%	BUENO 51-70%	MUY BUENO 71-80%	EXCELENTE 81-100%
1.Claridad	Esta formulado con el lenguaje apropiado				X	
2.Objetividad	Esta expresado en conducta observable				X	
3.Actualidad	Es adecuado al avance de la ciencia.				X	
4.Organización	Existe una organización lógica.				X	
5.Suficiencia	Comprende los aspectos de cantidad y calidad				X	

6.Intencionalidad	Es adecuado para valorar aspectos del sistema metodológico y científico.				X	
7.Consistencia	Está basado en aspectos teóricos, científicos acordes a la tecnología educativa.				X	
8.Coherencia	Entre los índices, indicadores, dimensiones.				X	
9.Metodología	Responde al propósito del trabajo bajo los objetivos a lograr.				X	
10.Pertinencia	El instrumento es adecuado al tipo de investigación.				X	
Promedio de Validación					80%	

III. Promedio de Valoración: 80%

IV. Observaciones:

ANEXO Nº 15:

FIGURA Nª 22 INNOVACIÓN Y APORTE TECNOLÓGICO

Antecedentes	Sistema intelligente	Técnicas usadas (algoritmos)	Validación cruzada	Lenguaje/plataforma de programación o Biblioteca	Métricas	Metodología de desarrollo
1 (Febrian et al. 2023)	No	K-Nearest Neighbor (KNN) Naïve Bayes (NB)	No especifica	Python Pima Indians Diabetes Database	Accuracy, Prediction Recall	No especifica
2 (Mansoori et al. 2023)	No	Regresión Logística (LR) Árbol de decisiones (DT) Bosque Aleatorio (RF)	Si	No especifica	Precisión, F1-Score, Sensibilidad, Especificidad, Accuracy	SMOTE
3 (Russell y Norvig 2016)	No	Naive Bayes (NB), neural networks (BP), Redes Neuronales (ANN)	No especifica	Matlab	Precisión, F1-Score	No especifica

4 (Orlando y Karina 2022)	No	K-vecino más cercano (K-NN) Bernoulli Naïve Bayes (BNB) Árbol de Decisión (DT), Regresión Logística (LR) Máquina de Vectores de Soporte (SVM)	No especifica	Python	Precisión, F1-Score, Sensibilidad, Especificidad, Accuracy recuperación	No especifica
5 Según (Amín 2021),	No	Regresión Logística (LR) Redes Neuronales (ANN) Decisión tree (DT) Neighbor(KNN)	No especifica	Python	Precisión, Recall y F1- Score	No especifica
6 (Tusell 2016),	No	Redes Neuronales (ANN) Decisión tree (DT) Neighbor(KNN) Naive Bayes (NB),	No especifica	Java	F1-Score, Precisión, Accuracy y Recall	No especifica

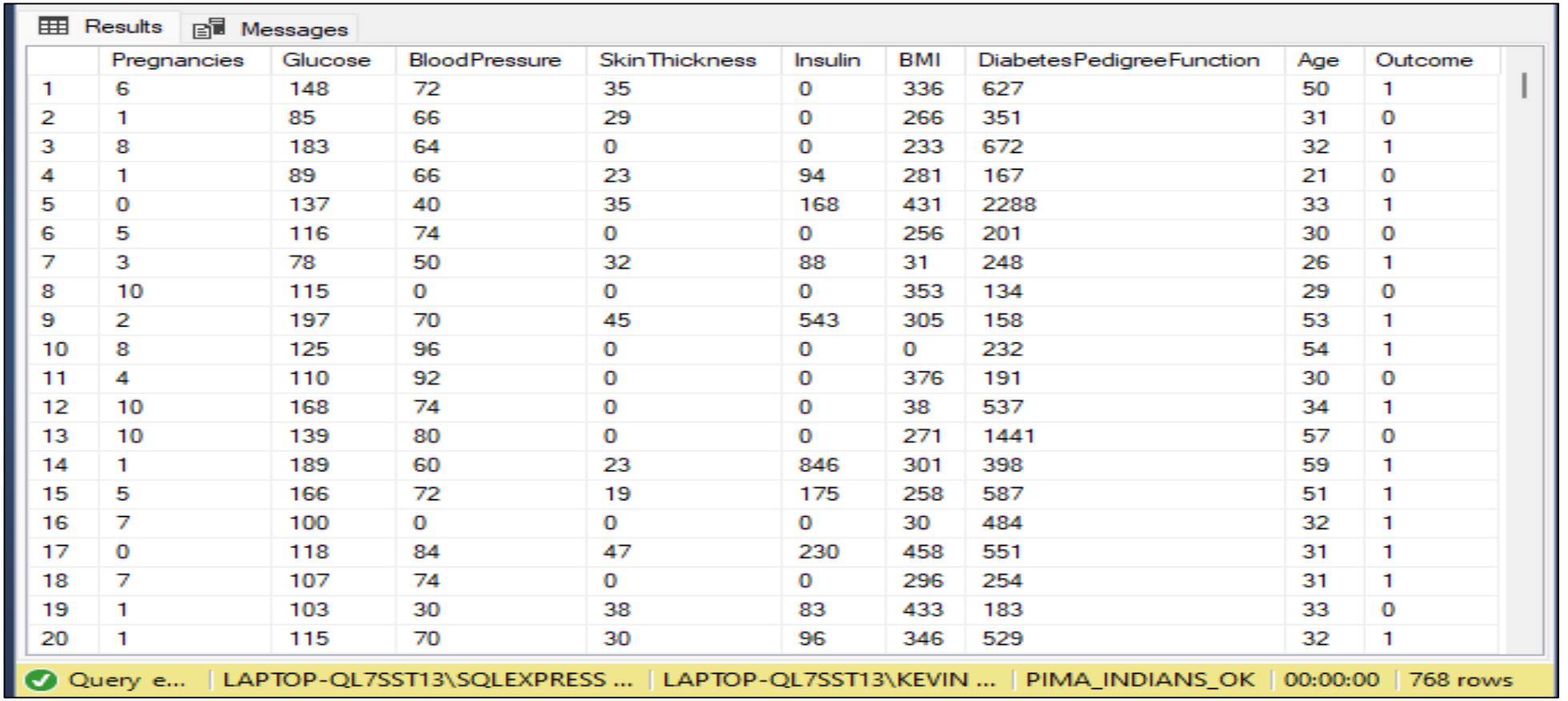
7 (Gandhi 2018)	No	Regresión Logística (LR) Bosque Aleatorio (RF) Random Forest	No especifica	Python	Accuracy, Precisión, Recall F1-Score	No especifica
8 Según (García-2017),	No	K-Vecinos más cercanos (KNN) Redes Neuronales ANN	Si	Python	Precisión, Recall, F1-Score	No especifica
9 Según (Russell y Norvig 2010)	No	Regresión Logística (LR) Bosque Aleatorio (RF) Random Forest	Si	SPSS statistics y Modeler	Precisión, Recall, F1-Score	No especifica
10 (Abdulrahman 2019)	No	Naive Bayes (NB) Redes Neuronales (ANN) K-Vecinos más cercanos (KNN)	No especifica	Python	Accuracy, Precisión, Recall y F1-Score	No especifica
11		Naive Bayes (NB) Redes Neuronales (ANN)				No especifica

(López Briega 2018)	No	K-Vecinos más cercanos (KNN)	No especifica	Python	Accuracy, Precisión, Recall y F1-Score	
12 (Rajput y Khedgika 2022)	No	K-Vecinos más cercanos (KNN) Naive beyes (NB) Bosque Aleatorio (RF) Máquina de Vectores de Soporte (SVM) Decisión tree (DT)	Si	No especifica	Exactitud Precisión Recuperación	No especifica
13 (Li et al. 2022)	No	Regresión Lineal (RL) Árbol de decisiones (DT) Decisión tree (DT) Bosque aleatorio y Xgboost	No especifica	Python 3.8	Accuracy, Precisión, Recall y F1-Score	No especifica
14 (Kopitar et al. 2020)	No	Glmnet Random forest XGBoost LightGBM,	Si	Python	Accuracy, Precisión, Recall y F1-Score	Método de regresión multivariada

15 Según (Li et al. 2022)	No	Regresión Logística (RL) Árbol de decisiones (DT) Bosque aleatorio y Xgboost	Si	Python	Accuracy, precisión, recall y F1-Score	No especifica
16 (Zhao et al. 2023)	No	Redes Neuronales (ANN) Decisión tree (DT) Neighbor(KNN) Naive Bayes (NB), ,	si	No especifica	Accuracy Rendimiento Precisión	No especifica
El propósito de esta investigación es realizar la compasión de modelos de aprendizaje automático (ML) para ver que algoritmo se adapta más a la predicción de la diabetes utilizando distintas métricas, tomando en cuenta que en los anteriores trabajos detallados anteriormente no han implementado un software para representar la información de sus resultados por tal motivo, este trabajo se considera de carácter innovador y original. Asimismo se muestran autores que realizaron similares trabajo y genera un aporte innovador tecnológico.						

Fuente: Elaboración propia.

FIGURA N° 23: EXPORTACIÓN A SQL SERVER.



Results | Messages

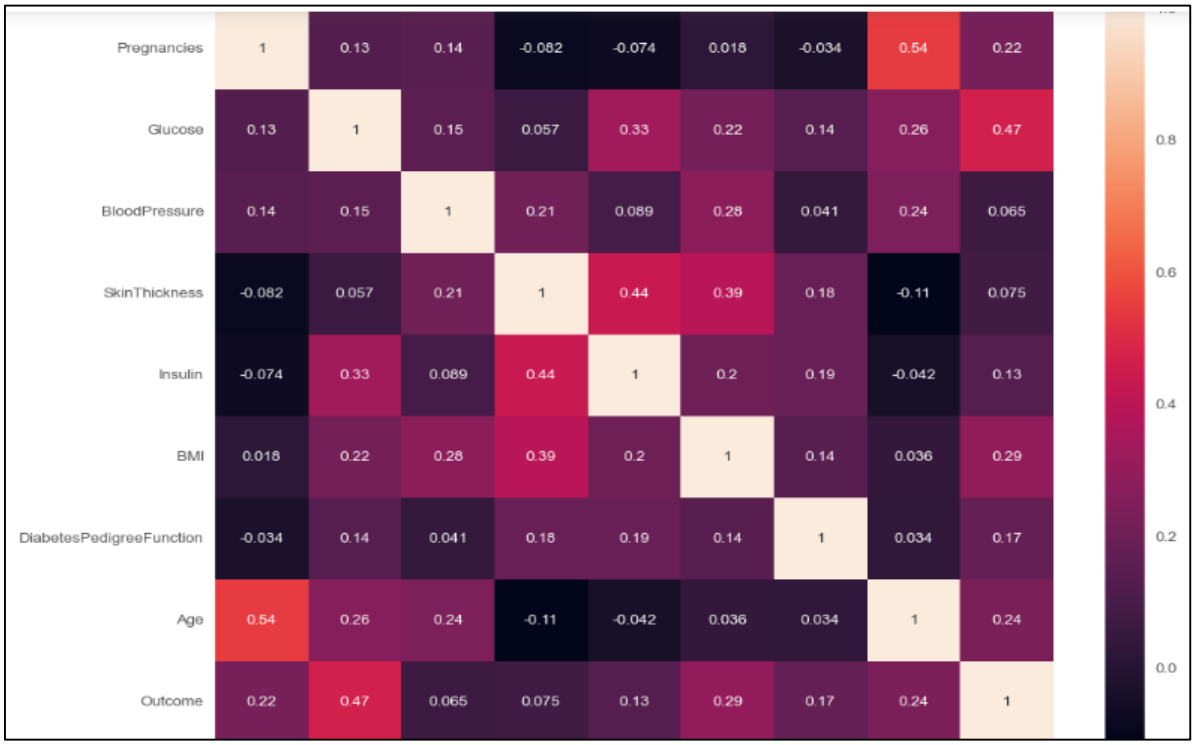
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	336	627	50	1
2	1	85	66	29	0	266	351	31	0
3	8	183	64	0	0	233	672	32	1
4	1	89	66	23	94	281	167	21	0
5	0	137	40	35	168	431	2288	33	1
6	5	116	74	0	0	256	201	30	0
7	3	78	50	32	88	31	248	26	1
8	10	115	0	0	0	353	134	29	0
9	2	197	70	45	543	305	158	53	1
10	8	125	96	0	0	0	232	54	1
11	4	110	92	0	0	376	191	30	0
12	10	168	74	0	0	38	537	34	1
13	10	139	80	0	0	271	1441	57	0
14	1	189	60	23	846	301	398	59	1
15	5	166	72	19	175	258	587	51	1
16	7	100	0	0	0	30	484	32	1
17	0	118	84	47	230	458	551	31	1
18	7	107	74	0	0	296	254	31	1
19	1	103	30	38	83	433	183	33	0
20	1	115	70	30	96	346	529	32	1

Query e... | LAPTOP-QL7SST13\SQLEXPRESS ... | LAPTOP-QL7SST13\KEVIN ... | PIMA_INDIANS_OK | 00:00:00 | 768 rows

Fuente: Motor SQL-server.

Se efectuó la carga de la base de datos indios pima de manera satisfactoria en el motor base datos SQL, con todas las variables definidas para la posterior aplicación de técnicas machine learning.

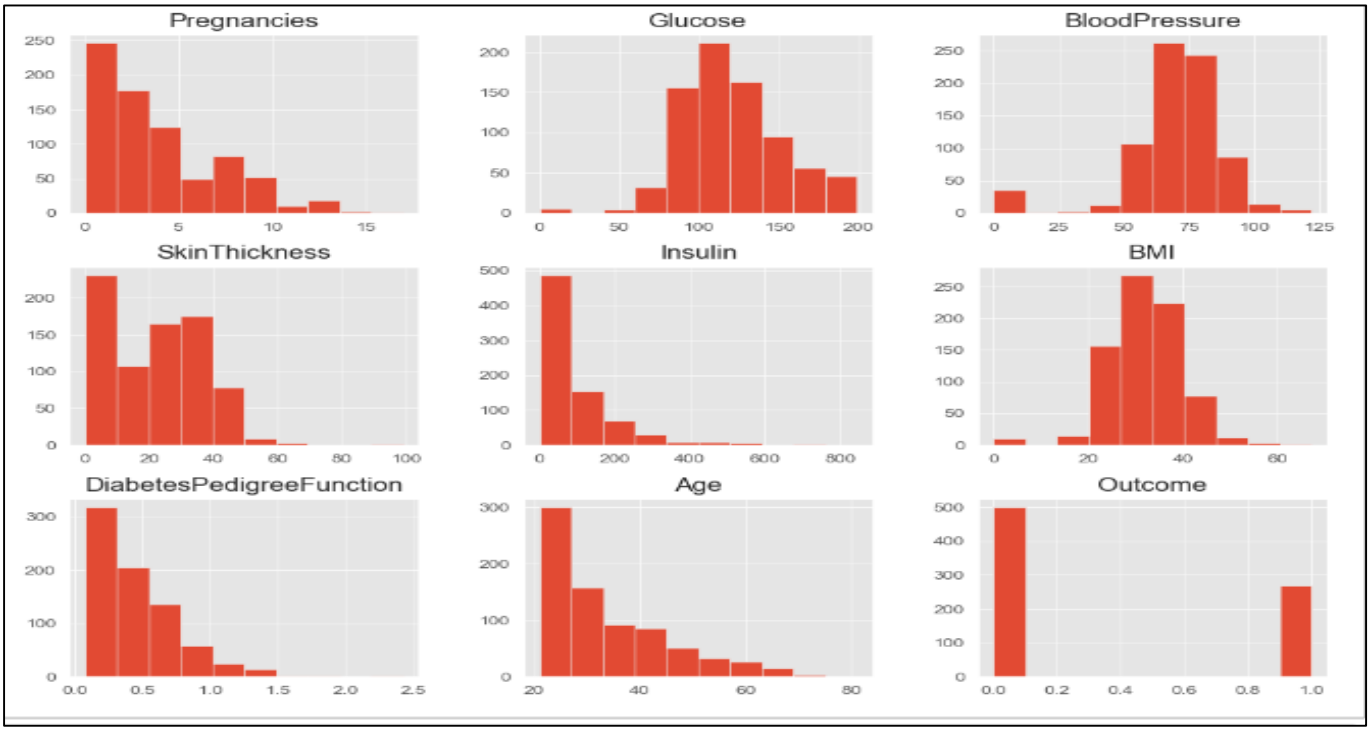
FIGURA N° 24: MATRIZ DE CORRELACIÓN DE VARIABLES



Fuente: Elaboración propia.

Posteriormente, se efectuó la conexión entre el motor de base de datos y en la plataforma Jupyter, para efectuar el análisis exploratorio de los datos, obteniendo la matriz de correlación de los datos contenidos, de los cuales destacan algunos valores que superan la correlación promedio, la cual equivale a 0.5. Tal como es el caso entre edad y embarazos, donde se evidencia que el número de embarazos aumenta a medida que avanza la edad y se detiene a partir de cierta edad. Asimismo, se encuentra una correlación significativa entre la glucosa y la insulina, donde un aumento en los niveles de glucosa se asocia con una mayor probabilidad de diagnóstico de diabetes. Para la glucosa y la diabetes: cuanto mayor sea el nivel de glucosa, mayor será la cantidad de insulina necesaria para regularlo. Además, existe una relación entre el IMC y la grasa corporal: cuanto mayor es el IMC, mayor es el porcentaje de grasa del paciente.

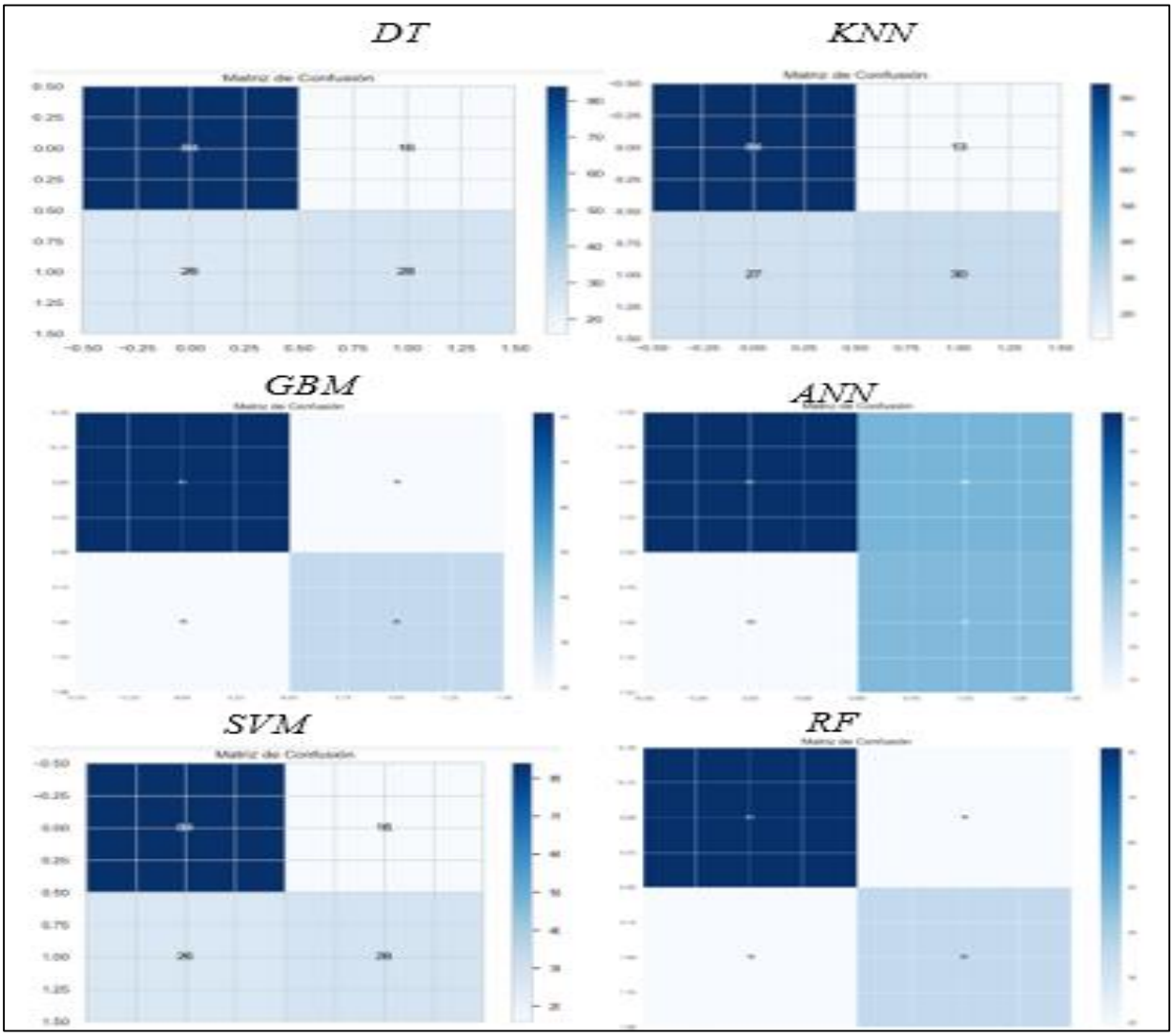
FIGURA N° 25: GRÁFICO HISTOGRAMAS



Fuente: Elaboración propia.

Mediante la consulta realizada a la base de datos a través del sistema Júpiter, se generó el gráfico de histogramas donde muestra la distribución de las variables; número de embarazos, presión arterial diastólica, grosor de los pliegues de la piel, nivel de insulina, índice de masa corporal, antecedentes genéticos de diabetes, edad y resultado de diabetes) (sí/no), de forma gráfica para la toma de decisiones en torno a la predicción de la diabetes.

FIGURA N° 26: MATRIZ DE CONFUSIÓN PARA ML



Fuente: Elaboración propia.

En este grafico se muestra las matrices de confusión para los 6 algoritmos, técnicas de machine learning; Árbol de decisiones (DT), la máquina de vectores de soporte (SVM), Gradient Boosting Machine (GBM), K-vecino más cercano (K-NN), Redes Neuronales (ANN) y Random Forest (RF), a fin de identificar posteriormente la predicción basada en el índice porcentual de predicciones acertadas para la predicción de la diabetes.

Minería de datos

Selección de Variables

Para realizar la selección se incluyeron todas las variables, previamente se realizó la importación de las librerías para la conexión a la base de datos y de los 06 algoritmos de selección de variables de la herramienta Jupyter, realizando la importación de las Librerías correspondientes para los algoritmos y métricas definidas.

Figura N° 27 Conexión a la base de datos (SQL)

```
In [2]: import pandas as pd
        import pyodbc
        #procedimiento para la conexion con el servidor
        conexion = pyodbc.connect('Driver={SQL Server};'
                                  'Server=LAPTOP-QL7SST13\SQLEXPRESS;'
                                  'Database=PIMA_INDIANS_OK;'
                                  'Trusted_Connection=yes;'
                                  )

In [21]: cursor = conexion.cursor()

         consulta_pima=pd.read_sql_query('select Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFuncti
         print (consulta_pima)
         print (type(consulta_pima))
```

Fuente: Elaboración propia.

Figura N° 28: Importación de librerías

```
In [3]: #importacion de librerias

        import pyodbc
        import pandas as pd
        import numpy as np
        import statsmodels.api as sm
        from sqlalchemy import create_engine
        import hist
        import scipy.cluster.hierarchy as sch
        from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
        import plotly.express as px
        import plotly.graph_objects as go
        from sklearn.tree import plot_tree
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.metrics import confusion_matrix
        from sklearn.metrics import roc_curve, auc
        from sklearn.model_selection import learning_curve
        from sklearn.metrics import precision_score, recall_score
        from sklearn.metrics import confusion_matrix, precision_recall_curve, roc_curve, auc
        from sklearn.metrics import RocCurveDisplay
        from sklearn.preprocessing import label_binarize
        from sklearn.metrics import roc_curve, auc
        from itertools import cycle
        from sklearn.metrics import silhouette_samples, silhouette_score
        from matplotlib.ticker import PercentFormatter
        sns.set(style='whitegrid')
```

Fuente: Elaboración propia.

Figura N° 29: Importación de librerías de algoritmos

```
In [3]: #importacion de librerias

        import pyodbc
        import pandas as pd
        import numpy as np
        import statsmodels.api as sm
        from sqlalchemy import create_engine
        import hist
        import scipy.cluster.hierarchy as sch
        from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
        import plotly.express as px
        import plotly.graph_objects as go
        from sklearn.tree import plot_tree
```

Fuente: Elaboración propia.

Figura N° 30: Importación de librerías de para métricas

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import roc_curve, auc
from sklearn.model_selection import learning_curve
from sklearn.metrics import precision_score, recall_score
from sklearn.metrics import confusion_matrix, precision_recall_curve, roc_curve, auc
from sklearn.metrics import RocCurveDisplay
from sklearn.preprocessing import label_binarize
from sklearn.metrics import roc_curve, auc
from itertools import cycle
from sklearn.metrics import silhouette_samples, silhouette_score
from matplotlib.ticker import PercentFormatter
sns.set(style='whitegrid')
```

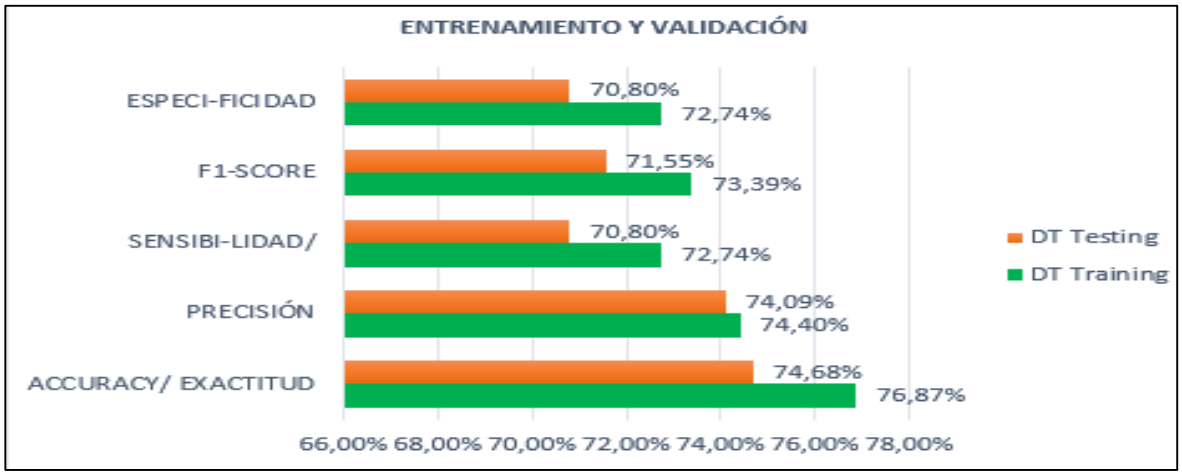
Fuente: Elaboración propia.

Figura N° 31: Resultados obtenidos para DT

```
RESULTADOS DEL CONJUNTO DE ENTRENAMIENTO
ACCURACY:  0.7817589576547231
PRECISION:  0.8050526315789474
RECALL:  0.7040270346381301
F1-SCORE:  0.7201991484499342
ESPECIFICIDAD:  0.7040270346381301
**********REPORTE DE ENTRENAMIENTO**********
              precision    recall  f1-score   support

       False       0.77      0.96      0.85       402
        True       0.84      0.45      0.59       212

    accuracy                           0.78       614
   macro avg       0.81      0.70      0.72       614
weighted avg       0.79      0.78      0.76       614

Matriz de confusión
[[384  18]
 [116  96]]
-------------------------------------------------
RESULTADOS DEL CONJUNTO DE VALIDACIÓN
ACCURACY:  0.7077922077922078
PRECISION:  0.6964332546551272
RECALL:  0.6364795918367347
F1-SCORE:  0.6395693555936963
ESPECIFICIDAD:  0.6364795918367347
**********REPORTE DE VAIDACIÓN**********
              precision    recall  f1-score   support

       False       0.72      0.90      0.80        98
        True       0.68      0.38      0.48        56

    accuracy                           0.71       154
   macro avg       0.70      0.64      0.64       154
weighted avg       0.70      0.71      0.68       154

Matriz de confusión
[[88 10]
 [35 21]]
```

Fuente: Elaboración propia.

Figura N° 32: Gráfica de Comparación entre Entrenamiento y Validación



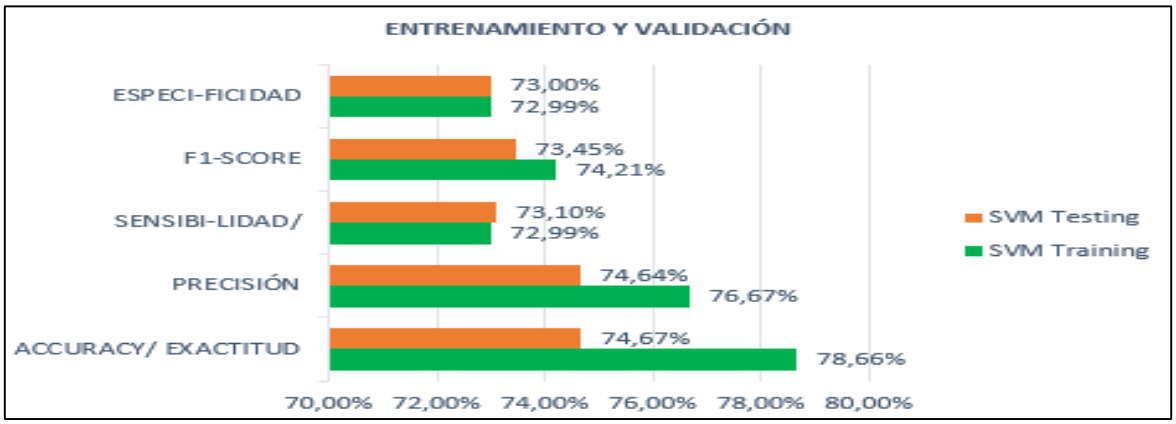
Fuente: Elaboración propia.

Figura N° 33: Resultados obtenidos para SVM

```
RESULTADOS DEL CONJUNTO DE ENTRENAMIENTO
ACCURACY:  0.7866449511400652
PRECISION:  0.7667772373134121
RECALL:  0.7299937518023647
F1-SCORE:  0.7421747534225938
ESPECIFICIDAD:  0.7299937518023647
***********REPORTE DE ENTRENAMIENTO***********
              precision    recall  f1-score   support

         0.0       0.81      0.90      0.85       412
         1.0       0.73      0.56      0.64       202

    accuracy                           0.79       614
   macro avg       0.77      0.73      0.74       614
weighted avg       0.78      0.79      0.78       614

Matriz de confusión
[[369  43]
 [ 88 114]]
------------------------------------------------
RESULTADOS DEL CONJUNTO DE VALIDACIÓN
ACCURACY:  0.7467532467532467
PRECISION:  0.7464646464646465
RECALL:  0.7310606060606061
F1-SCORE:  0.7345649003403014
ESPECIFICIDAD:  0.7310606060606061
***********REPORTE DE VALIDACIÓN***********
              precision    recall  f1-score   support

         0.0       0.75      0.84      0.79        88
         1.0       0.75      0.62      0.68        66

    accuracy                           0.75       154
   macro avg       0.75      0.73      0.73       154
weighted avg       0.75      0.75      0.74       154

Matriz de confusión
[[74 14]
 [25 41]]
```

Fuente: Elaboración propia.

Figura N° 34: Gráfica de Comparación entre Entrenamiento y Validación



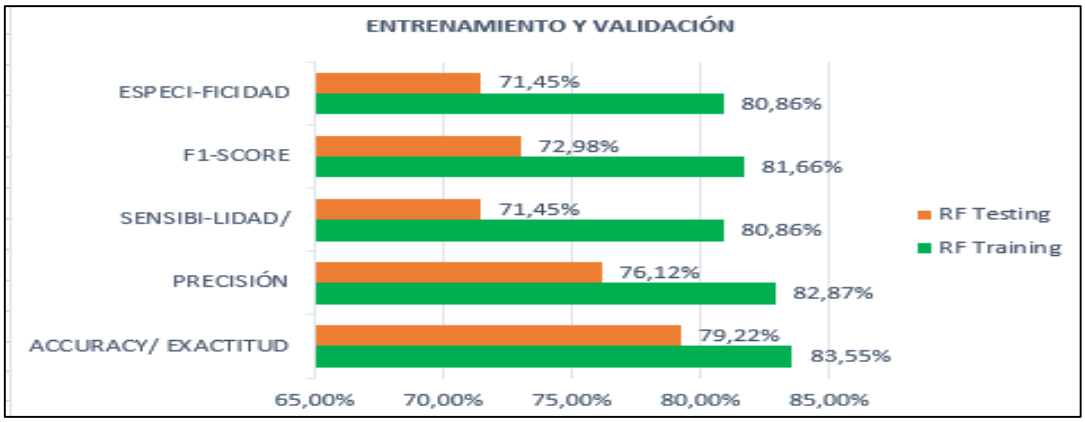
Fuente: Elaboración propia.

Figura N° 35: Resultados obtenidos para Random Forest

```
RESULTADOS DEL CONJUNTO DE ENTRENAMIENTO
ACCURACY:  0.8355048859934854
PRECISION:  0.8287558900924057
RECALL:  0.8086619783048354
F1-SCORE:  0.8166280785482588
ESPECIFICIDAD:  0.8086619783048354
**********REPORTE DE ENTRENAMIENTO**********
              precision    recall  f1-score   support

         0.0       0.85      0.91      0.88       392
         1.0       0.81      0.71      0.76       222

    accuracy                           0.84       614
   macro avg       0.83      0.81      0.82       614
weighted avg       0.83      0.84      0.83       614

Matriz de confusión
[[355  37]
 [ 64 158]]
-------------------------------------------------
RESULTADOS DEL CONJUNTO DE VALIDACIÓN
ACCURACY:  0.7922077922077922
PRECISION:  0.7612745098039215
RECALL:  0.7145732689210951
F1-SCORE:  0.7298245614035088
ESPECIFICIDAD:  0.7145732689210951
**********REPORTE DE VALIDACIÓN**********
              precision    recall  f1-score   support

         0.0       0.82      0.91      0.86       108
         1.0       0.71      0.52      0.60        46

    accuracy                           0.79       154
   macro avg       0.76      0.71      0.73       154
weighted avg       0.78      0.79      0.78       154

Matriz de confusión
[[98 10]
 [22 24]]
```

Fuente: Elaboración propia.

Figura N° 36: Gráfica de Comparación entre Entrenamiento y Validación



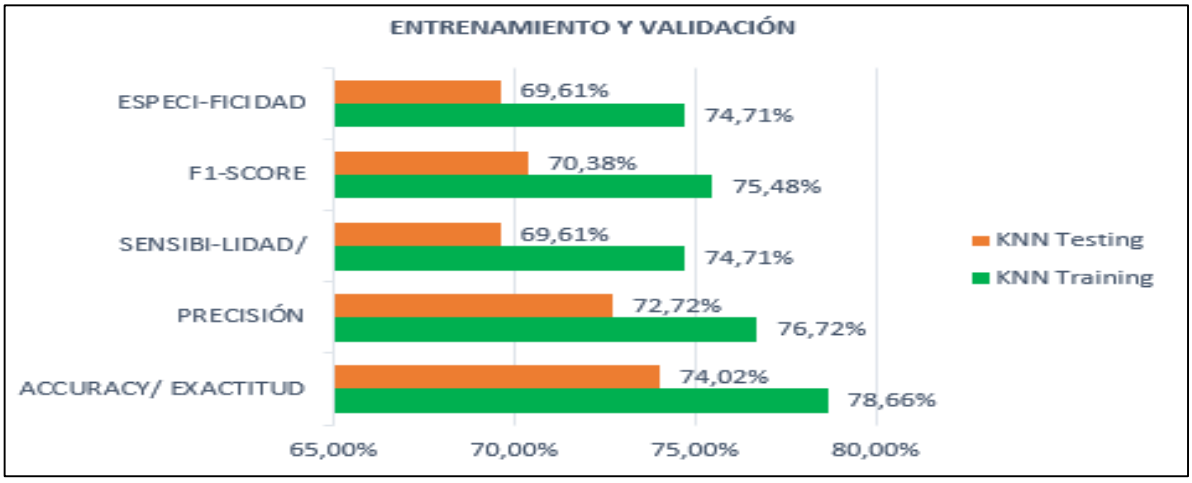
Fuente: Elaboración propia.

Figura N° 37: Resultados obtenidos para KNN

```
RESULTADOS DEL CONJUNTO DE ENTRENAMIENTO
ACCURACY:  0.7866449511400652
PRECISION:  0.7672975172975173
RECALL:  0.747151106041184
F1-SCORE:  0.7548902195608782
ESPECIFICIDAD:  0.747151106041184
***********REPORTE DE ENTRENAMIENTO***********
              precision    recall  f1-score   support

         0.0       0.81      0.87      0.84       403
         1.0       0.72      0.62      0.67       211

    accuracy                           0.79       614
   macro avg       0.77      0.75      0.75       614
weighted avg       0.78      0.79      0.78       614

Matriz de confusión
[[352  51]
 [ 80 131]]
------------------------------------------------
RESULTADOS DEL CONJUNTO DE VALIDACIÓN
ACCURACY:  0.7402597402597403
PRECISION:  0.727215587680704
RECALL:  0.6961475854584915
F1-SCORE:  0.7038461538461538
ESPECIFICIDAD:  0.6961475854584915
***********REPORTE DE VALIDACIÓN***********
              precision    recall  f1-score   support

         0.0       0.76      0.87      0.81        97
         1.0       0.70      0.53      0.60        57

    accuracy                           0.74       154
   macro avg       0.73      0.70      0.70       154
weighted avg       0.73      0.74      0.73       154

Matriz de confusión
[[84 13]
 [27 30]]
```

Fuente: Elaboración propia.

Figura N° 38: Gráfica de Comparación entre Entrenamiento y Validación



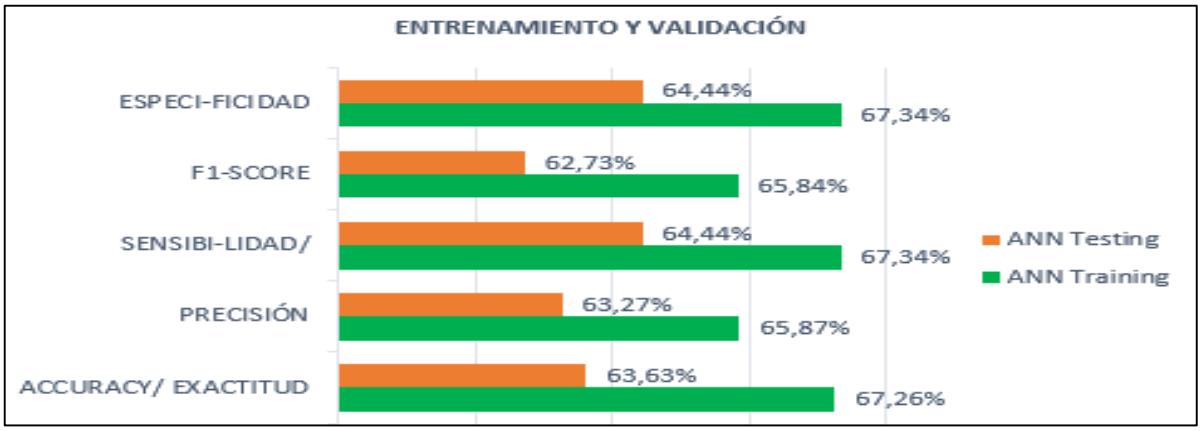
Fuente: Elaboración propia.

Figura N° 39: Resultados obtenidos para ANN

```
print(confusion_matrix(y_val, y_pred_val))

RESULTADOS DEL CONJUNTO DE ENTRENAMIENTO
ACCURACY:  0.6726384364820847
PRECISION:  0.6587985593002315
RECALL:  0.6734396403357803
F1-SCORE:  0.6584839403272538
ESPECIFICIDAD:  0.6734396403357803
**********REPORTE DE ENTRENAMIENTO**********
              precision    recall  f1-score   support

         0.0       0.80      0.67      0.73       401
         1.0       0.52      0.68      0.59       213

    accuracy                           0.67       614
   macro avg       0.66      0.67      0.66       614
weighted avg       0.70      0.67      0.68       614

Matriz de confusión
[[269 132]
 [ 69 144]]
--------------------------------------------------
RESULTADOS DEL CONJUNTO DE VALIDACIÓN
ACCURACY:  0.6363636363636364
PRECISION:  0.6327426160337553
RECALL:  0.6444444444444444
F1-SCORE:  0.6273120138288678
ESPECIFICIDAD:  0.6444444444444444
**********REPORTE DE VAIDACIÓN**********
              precision    recall  f1-score   support

         0.0       0.77      0.62      0.69        99
         1.0       0.49      0.67      0.57        55

    accuracy                           0.64       154
   macro avg       0.63      0.64      0.63       154
weighted avg       0.67      0.64      0.64       154

Matriz de confusión
[[61 38]
 [18 37]]
```

Fuente: Elaboración propia.

Figura Nª 40 Gráfica de Comparación entre Entrenamiento y Validación



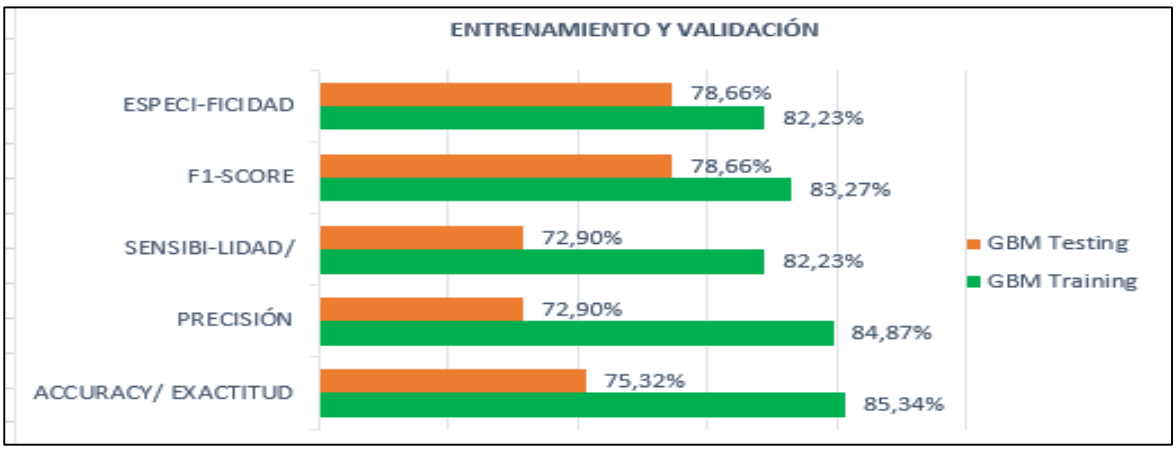
Fuente: Elaboración propia.

Figura N° 41: Resultados obtenidos para GBM

```
RESULTADOS DEL CONJUNTO DE ENTRENAMIENTO:
Accuracy: 0.8534201954397395
Precision: 0.8487108190091002
Recall: 0.8223130841121495
F1-score: 0.8327178058969547
Especificidad: 0.8223130841121495
-----Reporte de Entrenamiento:-----
              precision    recall  f1-score   support

         0.0       0.86      0.93      0.89       400
         1.0       0.84      0.72      0.77       214

    accuracy                           0.85       614
   macro avg       0.85      0.82      0.83       614
weighted avg       0.85      0.85      0.85       614

Matriz de confusión:
[[370  30]
 [ 60 154]]
//////////////////////////////////////////////////////
RESULTADOS DEL CONJUNTO DE VALIDACIÓN:
Accuracy: 0.7532467532467533
Precision: 0.7290740740740741
Recall: 0.7290740740740741
F1-score: 0.7290740740740741
Especificidad: 0.7290740740740741
-----Reporte de Validación:-----
              precision    recall  f1-score   support

         0.0       0.81      0.81      0.81       100
         1.0       0.65      0.65      0.65        54

    accuracy                           0.75       154
   macro avg       0.73      0.73      0.73       154
weighted avg       0.75      0.75      0.75       154

Matriz de confusión:
[[81 19]
 [19 35]]
```

Fuente: Elaboración propia.

Figura N° 42: Gráfica de Comparación entre Entrenamiento y Validación



Fuente: Elaboración propia.

Figura N° 43: Modelo entrenado - DT

```
import pickle
# Guardar el modelo entrenado en un archivo pickle
with open('decision_tree_model_trained.pkl', 'wb') as model_file:
    pickle.dump(grid_search.best_estimator_, model_file)
```

Fuente: Elaboración propia.

Figura N° 44: Modelo entrenado - SVM

```
import pickle
# Guardar el modelo entrenado en un archivo pickle
with open('support_vector_machine_trained.pkl', 'wb') as model_file:
    pickle.dump(grid_search.best_estimator_, model_file)
```

Fuente: Elaboración propia.

Figura N° 45: Modelo entrenado - GBM

```
import pickle
# Guardar el modelo entrenado en un archivo pickle
with open('gradient_boosting_machine.pkl', 'wb') as model_file:
    pickle.dump(grid_search.best_estimator_, model_file)
```

Fuente: Elaboración propia.

- **Sistema Inteligente con ML**

Figura N° 46: Librerías usadas para el sistema predictivo desde Spyder

```
import streamlit_option_menu as som
import streamlit as st
import pandas as pd
import numpy as np
import pickle

#ALGORITMOS------------------------------------------------------------
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, recall_score, f1_score, precision_score
```

Fuente: Elaboración propia.

Figura N° 47: Conexión a la base de datos desde Spyder

```
#import inspect

import pyodbc
#CONEXIÓN A BASE DE DATOS----------------------------------------------------------------
#DATOS DEL SERVIDOR
server = 'DESKTOP-105VM27'
bd= 'PIMA_INDIANS_OK'

#CONEXION AL SERVIDOR (SQL)
conexion = pyodbc.connect('Driver={SQL Server};'
                          'Server=LAPTOP-QL7SST13\SQLEXPRESS;'
                          'Database=PIMA_INDIANS_OK;'
                          'Trusted_Connection=yes;'
                         )


# CONSULTA LISTADO DE LA DATA MODELOS
consulta_pima_mod=pd.read_sql("""select Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabe
filtradogpimamod = pd.DataFrame(consulta_pima_mod)
```

Fuente: Elaboración propia.

Figura N° 48: Modelos importados

```
# Seleccionar las variables
selected_features = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin','BMI', 'DiabetesPedi

# Definir las variables
X = consulta_pima_mod[selected_features]
y = consulta_pima_mod["Outcome"]

# Dividir los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)


# Modelo DT
model_dt = pickle.load(open('C:\\Users\\KEVIN\\Desktop\\I Pyma\\ALGORITMS OK\sistema_academic\\models\\Decision_

# Modelo RF
model_rf = pickle.load(open('C:\\Users\\KEVIN\\Desktop\\I Pyma\\ALGORITMS OK\sistema_academic\\models\\Random_F

#Modelo GBM
model_gbm = pickle.load(open('C:\\Users\\KEVIN\\Desktop\\I Pyma\\ALGORITMS OK\sistema_academic\\models\\Gradien


if selected == "Datos de entrada":
    st.header('Datos de entrada')
    col1, col2, col3 = st.columns([1, 1, 1])

    def get_user_input():
        with col1:
            Preg= st.number_input('🧾 Numero de Embarazos:', 1, 12, step=1)
            st.write("")

            Gluc= st.number_input('🧾 Nivel de Glucosa:', 1, 199, step=1)
            st.write("")
```

Fuente: Elaboración propia.

Figura N° 49: Construcción del sistema

```
        user_data = {
            'Pregnancies': Preg,  # Numero de Embarazos
            'Glucose': Gluc,  # Nivel de Glucosa
            'BloodPressure': Blood,  # Presión arterial
            'SkinThickness': Skin,  #  Grosor de la piel
            'Insulin': Insu,  # Nivel de Insulina
            'BMI': Imc,  # Indice de Masa Corporal
            'DiabetesPedigreeFunction': Pedi,  #Función del pedigrí de la diabetes
            'Age': Edad,  # Edad

    }

        features = pd.DataFrame(user_data, index=[0])
        return features

  # Llamada a la función get_user_input
    user_input = get_user_input()

#-------------------------------------------------------------------------------------

    with col3:

        # Crear una lista de nombres de algoritmos
        algorithms = ['DESICION TREE', 'RANDOM FOREST', 'GRADIENT BOOSTING MACHINES']

        # Agregar un selectbox para que el usuario elija un algoritmo
        selected_algorithm = st.selectbox('SELECCIONE UN ALGORITMO:', algorithms)

        y_true = consulta_pima_mod["Outcome"]

        # Evaluar el algoritmo seleccionado
        if selected_algorithm == 'DESICION TREE':
            if st.button("EVALUAR - DT"):
```

Fuente: Elaboración propia.

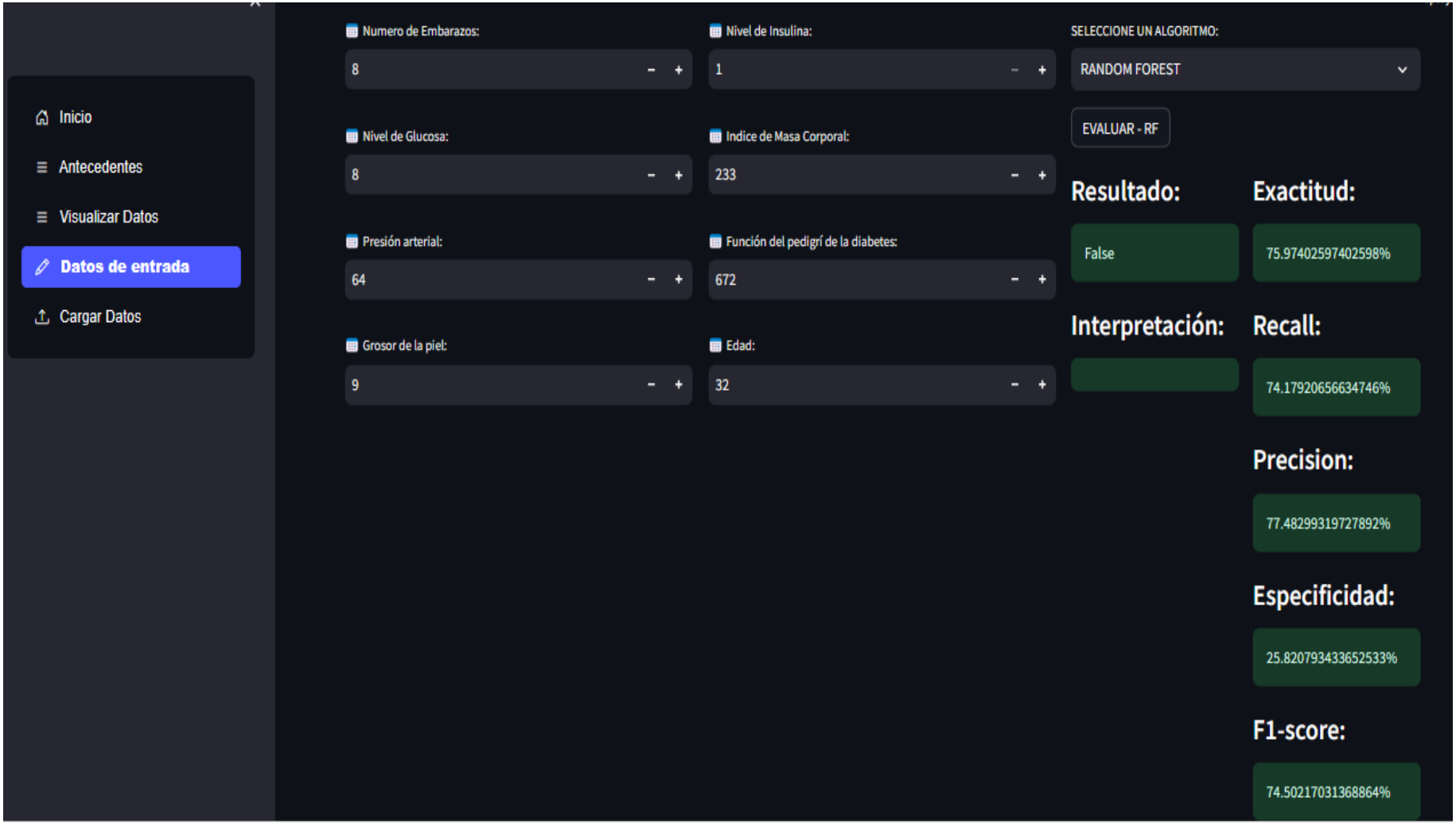
Figura N° 50: Vista del Sistema predictivo

Fuente: Elaboración propia.

Figura N° 51: Antecedentes del Sistema predictivo

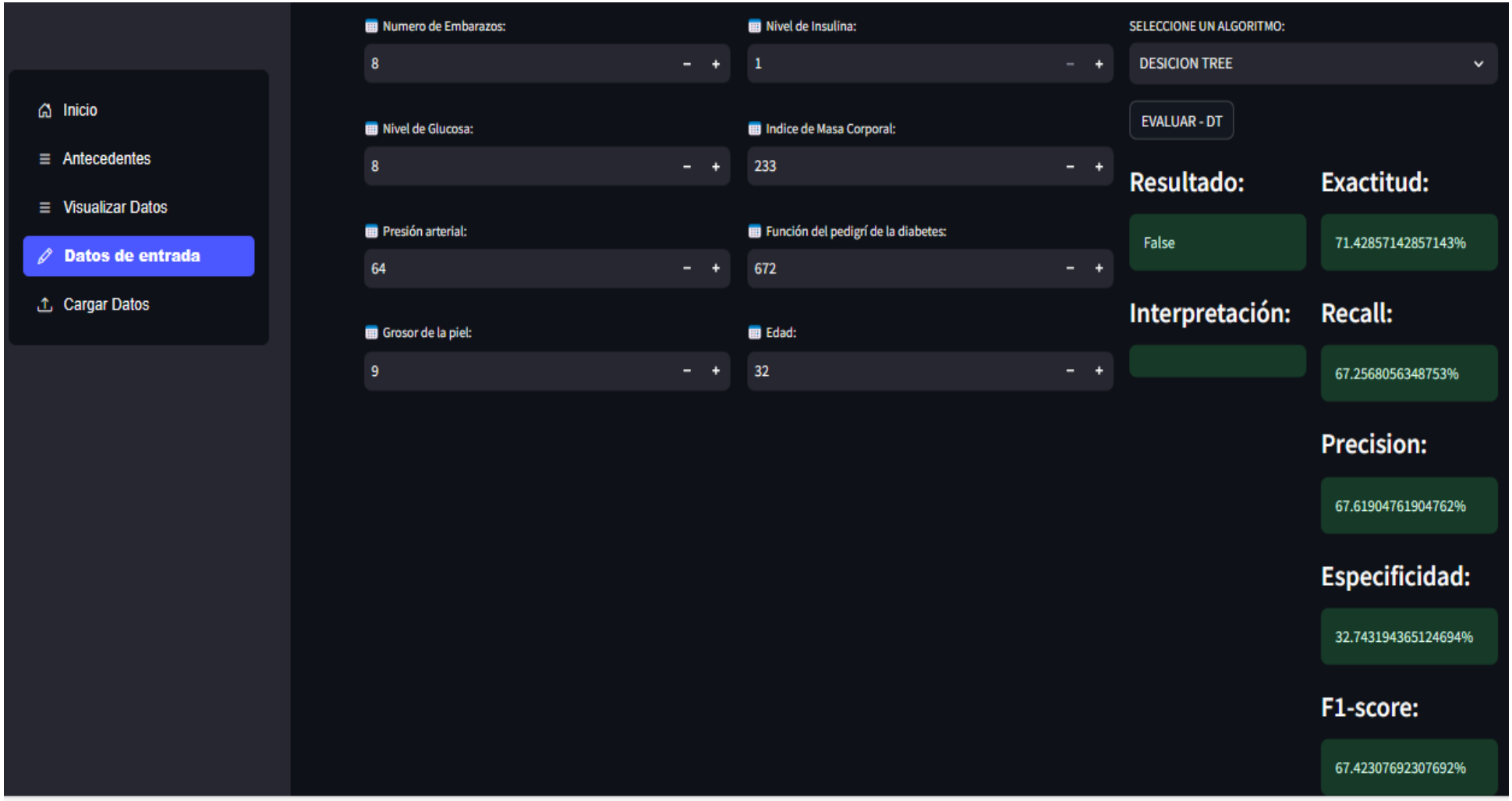
Fuente: Elaboración propia.

Figura N° 52: Resultado del Algoritmo Random Forest (RF)



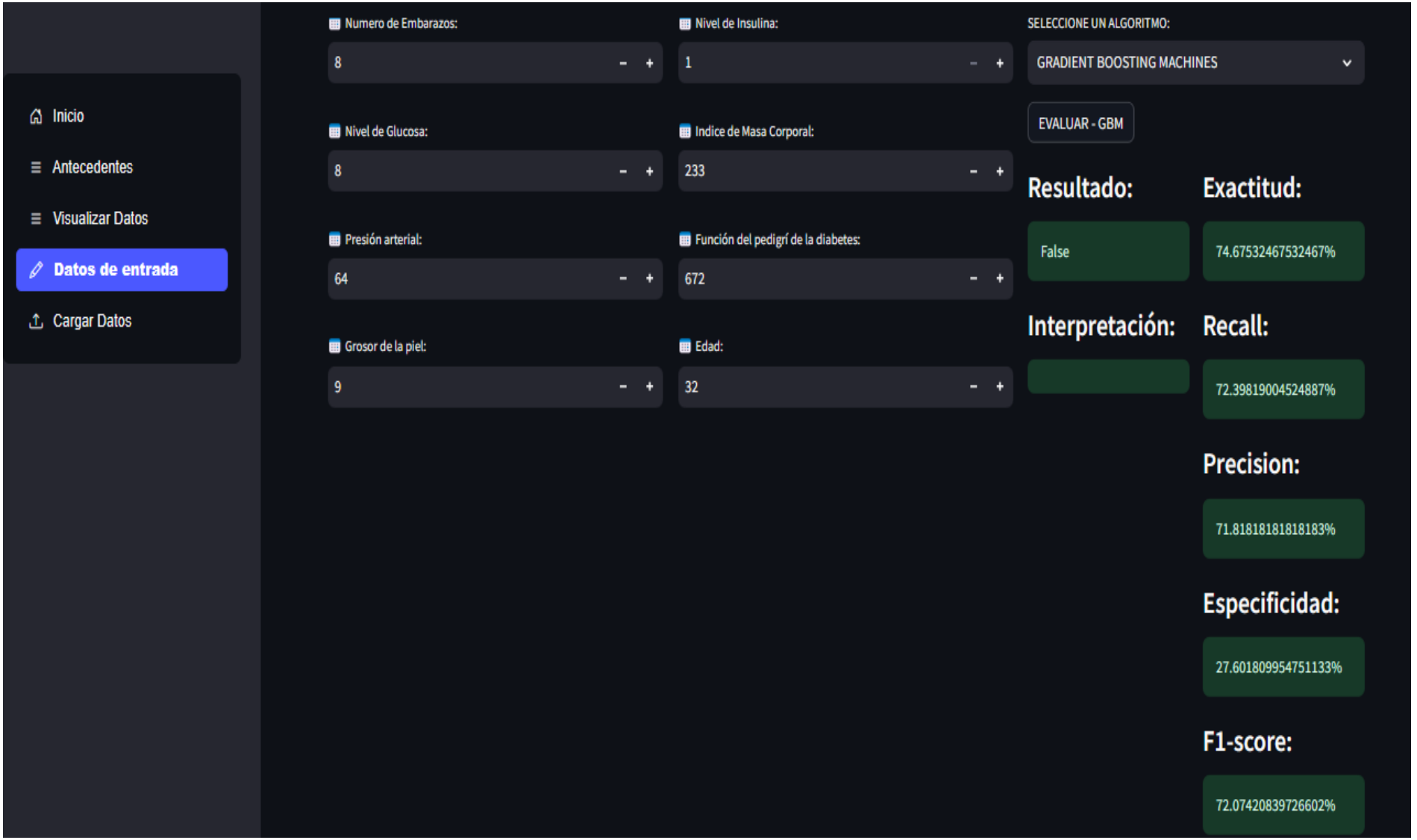
Fuente: Elaboración propia.

Figura N° 53: Resultado del Algoritmo Árbol de Decisiones (DT)



Fuente: Elaboración propia.

Figura N° 54: Resultado del Algoritmo Gradient Boosting Machine (GBM)



Fuente: Elaboración propia

ANEXO Nº 16:

Figura N° 55: Turnitin - Porcentaje de proyecto

Fuente: Elaboración propia

UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Declaratoria de Autenticidad del Asesor

Yo, DAZA VERGARAY ALFREDO, docente de la FACULTAD DE INGENIERÍA Y ARQUITECTURA de la escuela profesional de INGENIERÍA DE SISTEMAS de la UNIVERSIDAD CÉSAR VALLEJO SAC - LIMA NORTE, asesor de Tesis Completa titulada: "Análisis comparativo de técnicas de Machine Learning sobre el método de muestreo para la predicción de diabetes", cuyos autores son CHIRA BOHORQUEZ PIERO ALEJANDRO, RIVERA MUNIVE KEVIN, constato que la investigación tiene un índice de similitud de 17.00%, verificable en el reporte de originalidad del programa Turnitin, el cual ha sido realizado sin filtros, ni exclusiones.

He revisado dicho reporte y concluyo que cada una de las coincidencias detectadas no constituyen plagio. A mi leal saber y entender la Tesis Completa cumple con todas las normas para el uso de citas y referencias establecidas por la Universidad César Vallejo.

En tal sentido, asumo la responsabilidad que corresponda ante cualquier falsedad, ocultamiento u omisión tanto de los documentos como de información aportada, por lo cual me someto a lo dispuesto en las normas académicas vigentes de la Universidad César Vallejo.

LIMA, 26 de Diciembre del 2023

Apellidos y Nombres del Asesor:	**Firma**
DAZA VERGARAY ALFREDO **DNI:** 40466240 **ORCID:** 0000-0002-2259-1070	Firmado electrónicamente por: ADAZAVE el 26-12-2023 12:25:07

Código documento Trilce: TRI - 0708299

INVESTIGA UCV