Predicción del diagnóstico de diabetes a partir de perfiles clínicos de pacientes utilizando aprendizaje automático

Leydi Esperanza Pérez Leal

José Alejandro Buitrago Cárdenas

Universidad Antonio Nariño
Facultad de Ingeniería de Sistemas
Bogotá D.C., Colombia
2021

Predicción del diagnóstico de diabetes a partir de perfiles clínicos de pacientes utilizando aprendizaje automático

Leydi Esperanza Pérez Leal

José Alejandro Buitrago Cárdenas

Trabajo de grado presentado como requisito para optar al título de:
Ingeniero de Sistemas

Director:
Juan Camilo Ramírez, Ph.D.

Línea de Investigación:
Inteligencia Computacional
Grupo de Investigación:
LACSER - Laboratory for Advanced Computational Science and Engineering Research

Universidad Antonio Nariño
Facultad de Ingeniería de Sistemas
Bogotá D.C., Colombia
2021

Agradecimientos

En primer lugar queremos agradecer a Dios por permitirnos llegar hasta este punto. También expresar un profundo agradecimiento a los tutores Juan Camilo Ramirez y Jhonatan Rico, quienes con sus conocimientos y apoyo nos guiaron a través de cada una de las etapas de este proyecto para alcanzar los resultados que se buscaban.

Así mismo, queremos agradecer a la institución Universidad Antonio Nariño por brindarnos todos los recursos y herramientas que fueron necesarios para llevar a cabo el proceso de investigación. No hubiésemos podido arribar a estos resultados de no haber sido por su incondicional ayuda.

Al mismo tiempo, queremos agradecer a todos nuestros compañeros y familiares, por apoyarnos. En especial, queremos hacer mención a nuestros padres, que siempre estuvieron ahí para darnos palabras de apoyo y abrazos reconfortantes para renovar energías y poder culminar con cada uno de los objetivos propuestos.

Finalmente, queremos agradecer a todos los que aportaron un granito de arena para hacer que el presente trabajo de grado se finalizará con éxito.

Muchas gracias a todos.

Resumen

La diabetes en Colombia es una de las principales causas de muerte en la mayoría de los departamentos del país, según el Ministerio de Salud. La Organización Mundial de la Salud reconoce tres tipos principales de diabetes: tipo I, tipo II y gestacional. Una de las principales causas de mortandad por diabetes es que cuando el paciente es diagnosticado, la enfermedad ya está avanzada y por ende es difícil de tratar. Por lo tanto, es de gran importancia realizar un diagnóstico a tiempo, para que se puedan minimizar los factores que se derivan de este acontecimiento, como lo son: complicaciones graves (como: amputaciones, ataques cardíacos, daño ocular, úlcera en el pie, entre otros.); gastos monetarios (como: hospitalarios, personales, del estado); tiempo invertido, entre otros. Uno de los métodos empleados y haciendo uso de la tecnología, es la predicción del riesgo de desarrollar diabetes usando machine learning (ML), en donde se obtiene como resultado el pronóstico de la enfermedad y con ello, prevenir los resultados fatales y reducción de gastos financieros. Este proceso ya se ha venido realizando con el paso del tiempo y se encuentran varios estudios en donde se intenta predecir el diagnóstico de la diabetes utilizando aprendizaje automático. Este trabajo de grado consiste en llevar a la práctica modelos de aprendizaje automático para la predicción de diagnósticos de la diabetes a partir de perfiles clínicos de los pacientes. Se ha utilizado un conjunto de datos que ha sido recopilado previamente por el Hospital de Sylhet, Bangladesh, donde se realizó análisis de datos con varios algoritmos. Sin embargo, haciendo uso de otras técnicas de aprendizaje u otros métodos de preprocesamiento que no se han utilizado aún, se pueden obtener diferentes resultados y porque no mejores que los anteriores, lo cual se encontrará en el contexto de la investigación.

Finalmente, teniendo en cuenta las métricas *Recall, Precision, F-measure* y ROC AUC para la investigación realizada previamente y el presente trabajo de grado se tienen los siguientes resultados: para la investigación ROC AUC no se tuvo en cuenta, las demás métricas *Recall, Precision, F-measure* tuvieron un resultado promedio de 0.974 % para el modelo de bosques aleatorios; para el trabajo de grado los resultados fueron diferentes: para *Recall* y ROC AUC el mejor resultado fue bosques aleatorios con 0.98 % y 0.99 % respectivamente, para *Precision* el mejor resultado fue máquina vectores de soporte con 0.977 % y para *F-measure* el mejor resultado fue redes neuronales con 0.99 %. De forma general, los resultados obtenidos en el estudio mencionado anteriormente, indica que el mejor modelo fue: bosques aleatorios con un porcentaje de 97 %, seguido de árboles de decisión con 95 %, continua regresión logística con 92 % y por último bayes con 87 %, por otro lado, para el presente trabajo de grado, por medio de los resultados indica que el mejor modelo fue: 99 % para el modelo de bosques aleatorios, seguido con 96 % para el modelo de redes neuronales y continua 98 % para el modelo de máquinas de vectores de soporte.

Por lo anterior, se deduce que: tanto para el estudio como para el presente trabajo de grado el modelo con el mejor resultado es bosques aleatorios. Los resultados se pueden observar en

el contexto del presente proyecto de grado con las respectivas métricas obtenidas.

Palabras clave: Diabetes, Conjunto de datos, Algoritmo, Algoritmo de aprendizaje supervisado, Análisis predictivo, Algoritmos de clasificación.

Abstract

Diabetes in Colombia is one of the leading causes of death in most of the country's departments, according to the Ministry of Health. The World Health Organization recognizes three main types of diabetes: type I, type II and gestational. One of the main causes of death from diabetes is that by the time the patient is diagnosed, the disease is already advanced and therefore difficult to treat. Therefore, it is of great importance to make a diagnosis in time, so that the factors that derive from this event can be minimized, such as: serious complications (such as: amputations, heart attacks, eye damage, foot ulcer, among others); monetary expenses (such as: hospital, personal, state); time invested, among others. One of the methods used and making use of technology, is the prediction of the risk of developing diabetes using machine learning (ML), which results in the prognosis of the disease and thus, preventing fatal outcomes and reducing financial expenses. This process has already been done over time and there are several studies where they try to predict the diagnosis of diabetes using machine learning. This degree work consists of putting into practice machine learning models for the prediction of diabetes diagnoses from clinical profiles of patients. A dataset has been used which has been previously collected by Sylhet Hospital, Bangladesh, where data analysis was performed using various algorithms. However, by making use of other learning techniques or other preprocessing methods that have not been used yet, different results can be obtained and why not better than the previous ones, which will be found in the context of the research.

Finally, taking into account the metrics Recall, Precision, F-measure and ROC AUC for the research previously carried out and the present degree work, the following results are obtained: for the research ROC AUC was not taken into account, the other metrics Recall, Precision, F-measure had an average result of 0. 974 % for the random forest model; for the undergraduate work the results were different: for Recall and ROC AUC the best result was random forest with 0.98 % and 0.99 % respectively, for Precision the best result was support vector machine with 0.977 % and for F-measure the best result was neural networks with 0.99 %. In general, the results obtained in the above mentioned study indicate that the best model was: random forests with a percentage of 97 %, followed by decision trees with 95 %, continuous logistic regression with 92 % and finally Bayes with 87 %, on the other hand, for the present degree work, the results indicate that the best model was: 99 % for the random forests model, followed by 96 % for the neural networks model and continuous 98 % for the support vector machine model.

From the above, it is deduced that: both for the study and for the present degree project, the model with the best result is random forests. The results can be observed in the context of this degree project with the respective metrics obtained.

Keywords: Diabetes, Dataset, Algorithm, Supervised learning algorithm, Predictive analytics, Classification algorithms, Algorithm, Classification algorithms.

Contenido

Lista de Figuras

Lista de Tablas

Introducción

Según la Organización Panamericana de la Salud, la diabetes es una enfermedad metabólica crónica, se caracteriza por presentar alto grado de glucosa en la sangre. Es asociada con la deficiencia de la producción y/o acción de la insulina (22). Se remonta desde antes de la era cristiana. En un manuscrito del siglo XV AC descubierto por Ebers en Egipto, donde se mencionaban varios síntomas que parecen corresponder a la diabetes (4). Se considera que a nivel mundial hay 451 millones de adultos que tienen Diabetes y se aproxima que para el 2045 se incremente a 693 millones (5). En Colombia, según el Ministerio de Salud (Minsalud), se reportaron para el año pasado 1.294.940 personas que fueron diagnosticadas con la enfermedad y tiene mayor prevalencia en los departamentos de Cundinamarca (Bogotá), Antioquia y Valle del Cauca (23).

En la actualidad, existen varias clasificaciones, pero la Organización Mundial de la Salud (OMS) reconoce tres tipos de diabetes (tipo I, tipo II y gestacional); diabetes tipo I: enfermedad autoinmune, la cual destruye las células que producen insulina del páncreas; diabetes tipo II: no insulinodependiente, resistencia a la insulina; gestacional: usualmente se desarrolla en la segunda mitad del embarazo con intolerancia a la glucosa. Algunos de sus síntomas pueden ser: vómito, respiración acelerada, visión borrosa, mala cicatrización en las heridas, efectos sensoriales en manos y pies, aunque algunas personas pueden ser asintomáticos. Puede traer complicaciones como: temblores, mareos, sudoraciones, dolor de cabeza, palidez, aumento de sed, hambre, respiración acelerada, resequedad en la boca (6). También presenta complicaciones a largo plazo o complicaciones graves como daño ocular hasta ceguera, ataques cardíacos, accidentes cerebrovasculares, úlcera en el pie, amputación de extremidad afectada, daño renal, entre otro (1).

Los tratamientos recomendados para la enfermedad son: se debe tener un control de la glucosa en la sangre, dieta saludable, actividad física, no consumir tabaco y alcohol; se debe tener un control farmacológico de acuerdo a lo indicado por el médico y llevar un control médico con exámenes muy rigurosos (24). Para prevenir la enfermedad se tiene en cuenta 3 niveles de prevención: prevención primaria, acciones antes de presentar manifestaciones clínicas, como promoción a la lactancia materna, identificar la población en riesgo como familiares de pacientes diagnosticados; prevención secundaria, buen control de la enfermedad en pacientes portadores de intolerancia a la glucosa y pacientes ya diagnosticados, prevención terciaria, tomar control en pacientes con complicaciones crónicas, para detener o retardar la progresión (25).

Actualmente, la tecnología es un apoyo en todos los sectores del mundo, el sector salud no es

la excepción, por ende, el uso del aprendizaje automático es de gran ayuda para un diagnóstico anticipado y preciso de la enfermedad, esto sería una ayuda muy valiosa para el paciente, donde podría ahorrar tiempo y dinero. Se conocen estudios realizados anteriormente como: la tesis doctoral denominada propuesta de algoritmos de predicción de glucosa en pacientes diabéticos, en donde realizan monitorización continua de glucosa, insulina administrada y la ingesta de carbohidratos, dando como resultado poder valorar el impacto sobre el control glucémico (20); otro estudio es el clasificador bayesiano, este clasificador permite clasificar eventos discretos y limitados en un número de clases por medio de un entrenamiento, la detección de personas con la enfermedad tiene un promedio entre 81.53 % y 95.38 % (21).

En el presente trabajo de grado, se hace uso de un conjunto de datos público que fue recopilado por el hospital de Sylhet, Bangladesh. Este conjunto de datos fue utilizado en una investigación anterior para predecir la enfermedad usando varias técnicas de aprendizaje automático, en donde se usaron algoritmos como: bayes, árboles de decisión, regresión logística y bosques aleatorios (1); sin embargo, usando ese mismo conjunto de datos y haciendo uso de otras técnicas de aprendizaje u otros métodos de preprocesamiento que no se han utilizado aún, se pueden obtener resultados interesantes o incluso mejores en el presente trabajo de grado que los obtenidos en la investigación realizada anteriormente. Este proyecto de grado se basa en el entrenamiento de modelos de aprendizaje usando el conjunto de datos recopilado por el hospital de Sylhet, con la diferencia que se usaron otros métodos de preprocesamiento de datos y otras técnicas de aprendizaje que se relacionan en el contenido de la investigación.

1. Planteamiento del problema

1.1. Descripción del problema

Actualmente la diabetes en Colombia es una de las principales causas de muerte en los 32 departamentos del país excepto en Vaupés, según el Ministerio de Salud (Minsalud) (1). Una de las causas para que esta enfermedad presente alta tasa de mortalidad, es que el diagnóstico de la misma se da cuando la enfermedad está muy avanzada, debido a que muchos de los pacientes son asintomáticos a largo plazo (2). El diagnóstico precoz de esta enfermedad en los pacientes solo puede ser posible mediante una evaluación adecuada de los síntomas más frecuentes y signos comunes (2), es viable decir, que el diagnóstico anticipado de la diabetes es de gran importancia para su control y aún más para los pacientes diagnosticados a tiempo, ya que no tendrían que sufrir ningún contratiempo o que el avance de la enfermedad tenga desenlaces fatales.

Este trabajo de grado consiste en la implementación de modelos de aprendizaje automático para la predicción de diagnósticos de esta enfermedad a partir de perfiles clínicos de los pacientes. Los modelos de aprendizaje automático conocidos hasta ahora para predecir el diagnóstico de la diabetes logran grandes resultados, como por ejemplo: el clasificador bayesiano, este clasificador permite clasificar eventos discretos y limitados en un número de clases por medio de un entrenamiento; la detección de personas con la enfermedad tiene un promedio entre 81.53 % y 95.38 %, (21); otro ejemplo es la investigación predicción de probabilidad de diabetes en la etapa inicial con técnicas de minería de datos, en donde se usan los algoritmos de bayes, árboles de decisión, regresión logística y bosques aleatorios y se tiene como resultado 97.4 % de precisión para bosques aleatorios. Aunque ya hay investigaciones que intentan predecir el riesgo de desarrollar diabetes utilizando aprendizaje automático, ninguno de estos resultados tiene una certeza del 100 %, por lo cual existe la posibilidad de obtener mejores resultados (1). Como, por ejemplo, con este trabajo de grado.

El conjunto de datos que se va a utilizar en este trabajo de grado ya ha sido empleado para entrenar modelos utilizando varias técnicas de aprendizaje y parametrizaciones de las mismas. La presente investigación contempla utilizar el mismo conjunto de datos, pero con otras técnicas de aprendizaje y teniendo en cuenta otras parametrizaciones para buscar modelos más precisos que los que se conocen hasta ahora, lo que conlleva a desarrollar modelos predictivos usando técnicas que aún no se han implementado con el fin de obtener nuevos resultados. Los nuevos modelos que se pretenden implementar en este trabajo de grado tendrían la posibilidad de contribuir con un diagnóstico para la prevención de esta afección

y ayudar no solo a los pacientes, si no a los médicos y en general al sector de la salud.

1.2. Formulación del problema

Teniendo en cuenta los antecedentes registrados por la Organización Mundial de la Salud (OMS), en donde contemplan que esta enfermedad ha afectado a 422 millones de personas en todo el mundo en 2018 (1), y que alrededor del 50 % de todas las personas que sufren de diabetes no están diagnosticadas debido a su fase asintomática a largo plazo (2), surge la necesidad de crear modelos de aprendizaje automático precisos para la detección temprana de la diabetes. Teniendo en cuenta que ya existen estudios que intentan predecir el riesgo de desarrollar diabetes utilizando aprendizaje automático; sin embargo, ninguno de estos resultados tiene una certeza del 100 % y por lo tanto existe la posibilidad de obtener mejores resultados. Como consecuencia de lo anterior la pregunta que orienta la presente investigación es:
¿Mediante el aprendizaje automático, se pueden emplear nuevas técnicas, que permitan diseñar modelos predictivos para el diagnóstico temprano de la diabetes?

1.3. Justificación

La importancia de esta investigación radica en la contribución de predecir el riesgo de desarrollar diabetes, lo cual es de gran relevancia, porque permitirá evaluar varios factores que se tienen alrededor de esta, generando resultados positivos a todos los involucrados e interesados en el tema.
Desde el punto de vista económico es de gran consideración un diagnóstico anticipado ya que beneficiaría no solo al paciente en gastos como lo podrían llegar a ser: el transporte, medicamentos de apoyo para los tratamientos, dispositivos para medir el nivel de glucosa, exámenes generales, entre otros; sino también al sistema, en sentido de que reduce el gasto nacional bruto, dando un ejemplo por complicaciones graves (cirugías, amputaciones, hospitalizaciones por ataques, entre otros); medicamentos, pruebas de tolerancia, inyecciones, entre otros, los cuales en gran medida y diversos pacientes, considerando la población de un país alcanzan un alto valor monetario para el gobierno.
Es fundamental para la sociedad afectada con esta enfermedad, ya que un diagnóstico a tiempo puede ser de gran y vital importancia para el paciente; sin embargo, el tema no deja de ser relevante para instituciones médicas, personal médico, instituciones educativas, y los demás miembros que se puedan interesar en el tema, como el sistema de salud.
Desde el punto de vista tecnológico es de gran relevancia, por varios factores positivos que ha traído a lo largo del tiempo, como: su rapidez y efectividad en los resultados, precisión en el

diagnóstico, menor intervención en el cuerpo, mejor servicio al poder realizarlo masivamente, entrega de resultados en línea, entre otros servicios que se han alcanzado con la tecnología. Como profesionales, el desarrollo del proyecto contribuye con la búsqueda de medidas preventivas contra la diabetes al explorar alternativas no consideradas hasta ahora. El hecho de poder contribuir a la sociedad anticipando un diagnóstico como profesionales es de plena satisfacción personal, no solo por el hecho de ayudar a las personas, sino también, por contribuir con el mejoramiento de la salud de los pacientes y, por ende, de la vida como tal. Por consiguiente, encontrar modelos mediante aprendizaje automático con un acertado porcentaje de si una persona es positiva o no para diabetes en una fase temprana de su enfermedad, será de gran ayuda en primera medida para el paciente y segundo para el sector de la salud, dadas las descripciones anteriormente hechas.

1.4. Objetivos

1.4.1. Objetivo General

Determinar a partir de las métricas de desempeño más comúnmente utilizadas el mejor modelo de aprendizaje automático para la predicción de diagnósticos de diabetes a partir de los perfiles clínicos recopilados en la investigación (1).

1.4.2. Objetivo Específicos

- Diseñar varios modelos de aprendizaje para predecir la diabetes utilizando el conjunto de datos público, recopilado por el hospital de Sylhet, Bangladesh, utilizando varias técnicas de aprendizaje automático (1).
- Evaluar cada modelo, utilizando varias métricas de desempeño, incluyendo la curva ROC y el puntaje F1.
- Identificar el modelo que muestre el mejor desempeño, de acuerdo con lo realizado en el análisis de los datos obtenidos en cada prueba de los modelos de aprendizaje y compararlo con las investigaciones previas que utilizaron el mismo conjunto de datos.

1.5. Alcance y Limitaciones

1.5.1. Alcance

Esta investigación tiene como alcance la implementación, entrenamiento y evaluación de varios modelos de aprendizaje automático (entre los cuales podemos mencionar redes neuronales y bosques aleatorios) para predecir el riesgo de desarrollar diabetes. Para lo cual, se tiene en cuenta un conjunto de datos recopilado por el Hospital de Sylhet, Bangladesh; estos

datos presentan el historial clínico de personas con diagnóstico positivo y no diagnosticado con la enfermedad para efectos de evaluar la precisión de los mismos.
Para el desarrollo de esta investigación, el tipo de modelos a utilizar se conocen como modelos de clasificación, ya que cada uno intenta clasificar cada dato (los síntomas de un paciente) en una de dos categorías: "paciente con diabetes" y "paciente sin diabetes". La investigación pretende por medio del uso de modelos de clasificación, predecir el riesgo de desarrollar la diabetes en una etapa temprana, valiéndose de técnicas de aprendizaje. Dicha estrategia tiene como propósito comprobar si con las técnicas sugeridas se puede obtener modelos más confiables que con las técnicas utilizadas en investigaciones previas, incluyendo aquella donde se recopilaron los datos.

1.5.2. Limitaciones

- El conjunto de datos a utilizar para el entrenamiento de los algoritmos, es de uso público y no fue recopilado por los participantes de esta investigación. Esto se debe a que la logística requerida para recopilar el volumen de datos suficiente para este estudio, está fuera de las posibilidades de un trabajo de grado para pregrado, por lo tanto, se opta por utilizar un conjunto de datos recopilado por otros investigadores.

- Se utilizarán algoritmos o técnicas de aprendizaje ya existentes, tales como redes neuronales y bosques aleatorios, para entrenar nuevos modelos utilizando parametrizaciones de éstos, así como estrategias de procesamientos de datos no completadas en investigaciones previas, incluyendo aquella donde se recopilaron los datos.

- No se van a realizar modificaciones y/o mantenimiento en el modelo entregado en el proyecto final.

2. Marco de referencia

2.1. Marco teórico

2.1.1. Diabetes

Es una enfermedad metabólica crónica, se caracteriza por presentar niveles elevados de glucosa en la sangre o más comúnmente denominado como azúcar en la sangre. Esta enfermedad con el tiempo puede llevar a daños graves en el corazón, ojos, nervios, riñones y vasos sanguíneos. La diabetes tipo II es la más común y es cuando el cuerpo es resistente a la insulina o no produce la necesaria. La diabetes tipo I también conocida como diabetes insulinodependientes, es cuando el páncreas produce poca o no produce la insulina por sí mismo (22).

Síntomas

La diabetes puede provocar dificultades en todas partes del cuerpo y aumentar el riesgo de muerte prematura. Entre las complicaciones que se presentan pueden ser: insuficiencia renal, amputaciones, pérdida de visión, daño en los riñones (22). Algunos síntomas pueden ser: sed excesiva, poliuria, aumento anormal de la necesidad de comer, respiración acelerada, mala cicatrización de heridas, náusea o vómito, visión borrosa y modificaciones sensoriales en manos y pies; también, es preciso indicar que algunos individuos pueden ser asintomáticos(6).

Tratamiento

El diagnóstico temprano de la diabetes es esencial en la salud del paciente, ya que mejora sus condiciones de vida con el tratamiento correspondiente y puede evitarle complicaciones fatales. Se aconseja realizar ejercicio con regularidad, comer de manera saludable, evitar hábitos como: fumar o tomar, chequear la presión arterial y los lípidos para reducir el riesgo cardiovascular; de ser necesario la medicación y exámenes periódicos de detección de daños en los ojos, riñones y pies, esto con el fin de facilitar un tratamiento temprano (22).

Perfil clínico

Son los datos o información que se relaciona con la enfermedad, describe la salud y los factores determinantes en una población dada. Toda la agrupación factible que pueda causar la enfermedad (22).

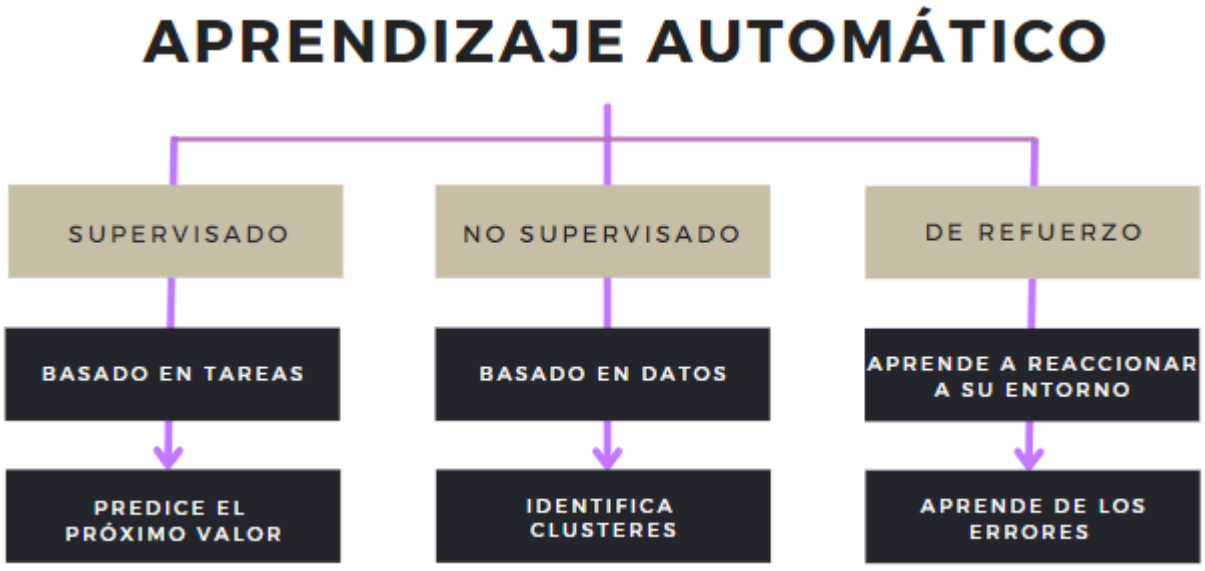
2.1.2. Aprendizaje automático (en inglés, machine learning ML)

Es un aspecto de la informática en el que las máquinas tienen la capacidad de aprender sin estar programados para ello. Es la ciencia de enseñar a las máquinas a que aprendan por sí mismas. Un gran ejemplo de ello son las sugerencias de búsqueda de Google o las sugerencias de Facebook. El aprendizaje automático hace uso de los algoritmos para aprender de patrones de datos y utilizan ese conocimiento adquirido para tomar decisiones (30).

Tipos de aprendizaje automático

Como se observa en la figura 2-1, existen tres tipos de aprendizaje automático: aprendizaje supervisado, no supervisado y de refuerzo. En el aprendizaje supervisado, los algoritmos usan datos previamente identificados para indicarle como tiene que ser categorizada la nueva información, se requiere de una persona que se encarga de retroalimentar el algoritmo; en el aprendizaje no supervisado, los algoritmos no usan datos identificados previamente para indicarle al algoritmo como va a ser la clasificación de la información, el algoritmo debe encontrar la manera de realizar la clasificación; por ende, no requiere de una persona que retro-alimente el algoritmo; en el aprendizaje por refuerzo, los algoritmos aprenden de la experiencia, es decir, se debe dar un refuerzo positivo o incentivo cada vez que aciertan (30).

Figura 2-1.: La imagen representa los tipos de aprendizaje automático y en qué se encuentran basados.



Adaptado de: “Inteligencia artificial”. Autor: LASSE ROUHIAINEN (30)

2.1.3. Clasificación binaria

Se denomina clasificación a cuando se tiene un conjunto de elementos y se arreglan de acuerdo a un juicio específico. Clasificación binaria es aquel tipo de clasificación donde los elementos de un conjunto solo tienen dos valores posibles, es decir, se clasifica el conjunto en dos categorías, un ejemplo es si un tumor es maligno o benigno, si una imagen contiene una flor o una fruta, lo cual, podría interpretarse como 1 para maligno y 0 para benigno, también 1 para fruta y 0 para la flor (28).

2.1.4. Variables Dummy

Es donde se elimina la variable de entrada que pueden ser enteros o cadenas y se agrega una nueva variable binaria para cada valor entero único en la variable, esto crea una nueva columna binaria para cada categoría (45). Por ejemplo, se tiene la siguiente tabla 2-1 de una veterinaria en donde se observa las columnas nombre dueño, nombre mascota y el tipo de mascota.

Tabla 2-1.: Creación de variable dummy a partir de la variable mascota

Nombre Dueño	Nombre mascota	Mascota
Juan Garrido	Gogo	Gato
Lili Torres	Firulais	Perro
Santiago Puentes	Snack	Tortuga
Valeria Torrado	Nemo	Pez
Gabriel Fuentes	Michi	Gato
Karen Vivas	Veloz	Tortuga

Fuente: elaboración propia

Se realiza la transformación a variables dummy de la siguiente forma: se crean las variables binarias para cada uno de los elementos de la columna mascota como se observa en la tabla 2-2, indicando 1 en la nueva matriz donde coincidan verticales con horizontales y 0 para las demás casillas, como se puede ver en la tabla. Las columnas Gato, perro, tortuga y pez son las variables dummy.

2.1.5. Redes Neuronales Artificiales (En inglés, Artificial Neural Networks, ANN)

Las redes neuronales artificiales se crearon a partir del funcionamiento de las redes de neuronas naturales que son células del sistema nervioso que sirven para emitir y trasmitir la información. Ellas comunican la información mediante impulsos nerviosos, la acción se realiza desde el cerebro hacia el resto del cuerpo y viceversa, o desde una parte del cuerpo hacia

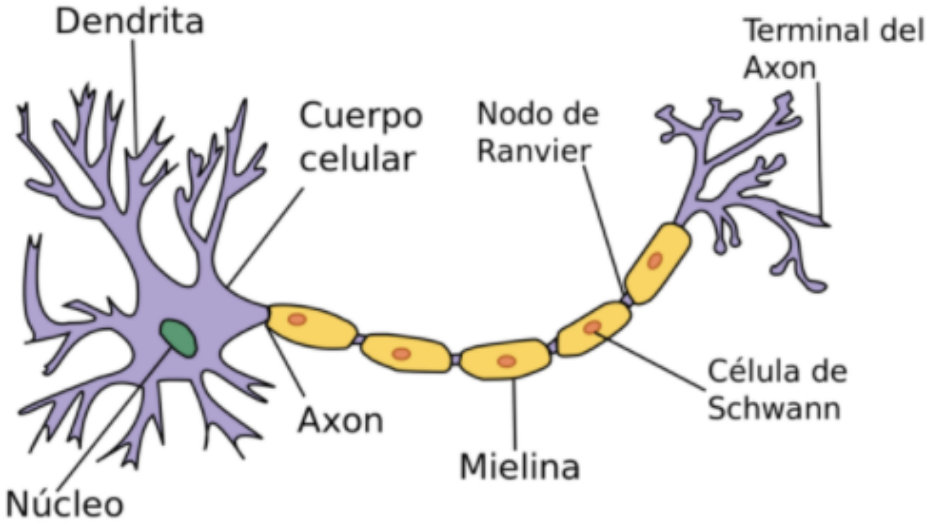
Tabla 2-2.: Creación de variables dummy a partir de la variable mascota

Nombre Dueño	Nombre mascota	Mascota	Gato	Perro	Tortuga	Pez
Juan Garrido	Gogo	Gato	1	0	0	0
Lili Torres	Firulais	Perro	0	1	0	0
Santiago Puentes	Snack	Tortuga	0	0	1	0
Valeria Torrado	Nemo	Pez	0	0	0	1
Gabriel Fuentes	Michi	Gato	1	0	0	0
Karen Vivas	Veloz	Tortuga	0	0	1	0

Fuente: elaboración propia

otras partes. Entonces, la información es recibida por la neurona a través de las dendritas, posterior el núcleo se encarga de procesar la información y luego sale a través del axón, el cual está conectado a las dendritas de otra neurona. La figura 2-2 se representan las partes de la neurona (26).

Figura 2-2.: La imagen representa las partes de la neurona, en donde la información ingresa por las dendritas y sale a través del axón a otra neurona.



Fuente: "Neuronas". Autor: Julia Máxima Uriarte (26)

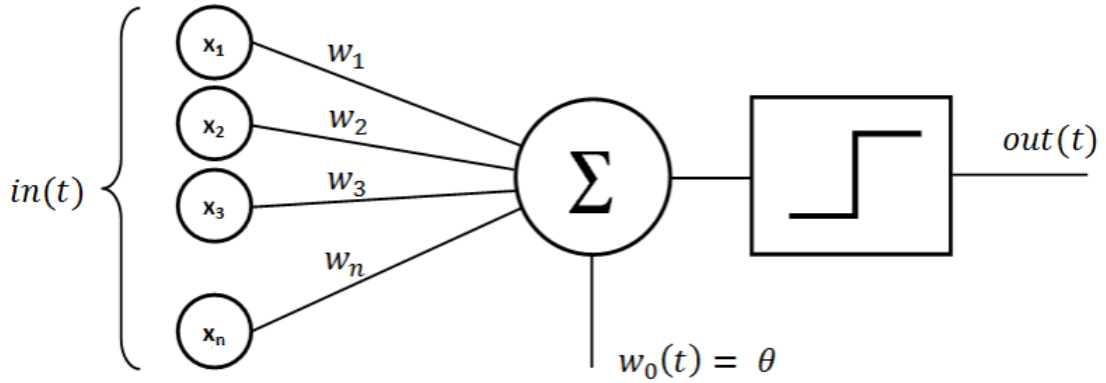
Las redes neuronales artificiales son sistemas de computación que pretenden simular la receptividad al aprendizaje humano, a través de una arquitectura semejante a la del sistema nervioso humano (27). Se usan para distinguir modelos como: figuras, pergaminos, orientación económica, entre otros. Una peculiaridad importante es que tienen la capacidad de

aprender y de mejorar su funcionamiento, por otra parte, una gran ventaja es que presenta muchas características semejantes a las del cerebro (44). Las redes neuronales se pueden encontrar dentro de los tres tipos de aprendizaje automático mencionados anteriormente.

Perceptrones Monocapa

Es la red neuronal más elemental y está compuesta por una sola neurona, como se observa en la figura 2-3. La neurona recibe varios datos de entrada denominados X_1, X_2, $X_3...X_n$, para enseguida calcular los pesos de estos valores. Los pesos son W_1, W_2, $W_3...W_n$ y se le suma un valor W_0 denominado **bias**. Los pesos y el valor W_0 son "aprendidos" o son los que se usan para el entrenamiento de la neurona. Posterior, la suma pasa por una función de activación, procesa la información y produce una salida que es el resultado final (28) (29).

Figura 2-3.: La imagen representa el perceptrón. Los datos de ingreso son representados por la X, luego calcula los pesos W y le suma el Wo, lo pasa por una función y genera una salida como resultado final.



Fuente: "Perceptrón". Autor: Scott Robinson (27)

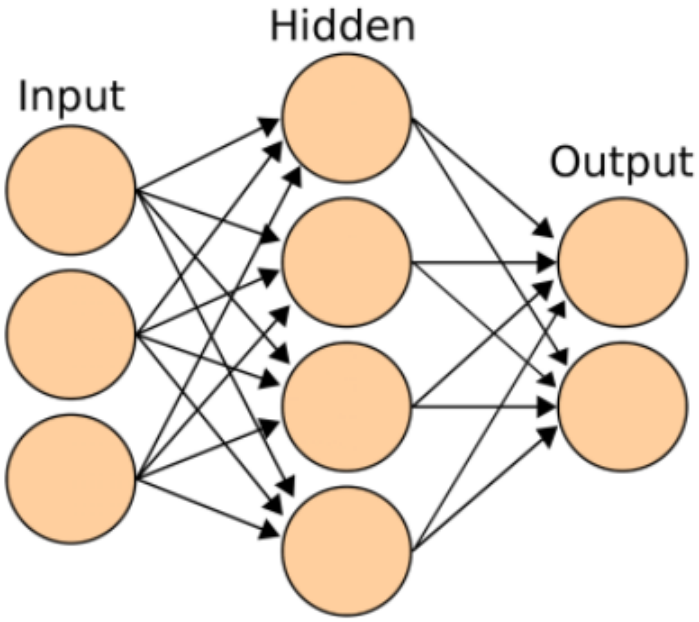
Perceptrón Multicapa

Mas comúnmente llamados redes neuronales artificiales, son múltiples neuronas conectadas en forma de red; esta presenta una capa de entrada (*input*), puede haber una o más capas ocultas (*hidden*) y una capa de salida (*output*) (28), como se observa en la figura 2-4.

Formas de conexión entre neuronas

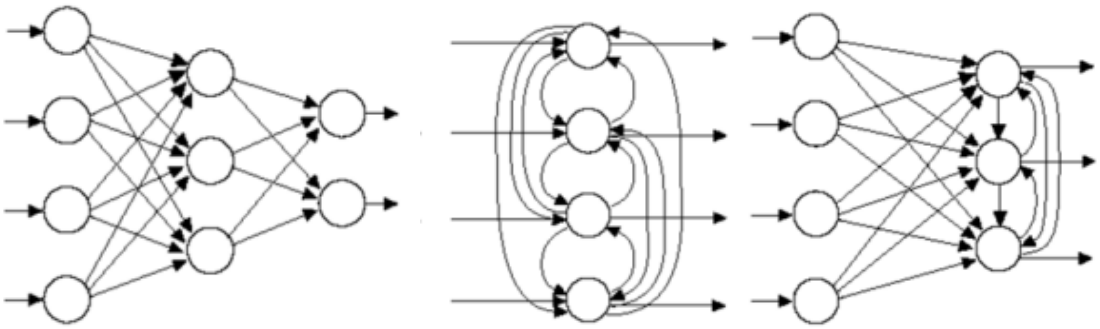
Está relacionada con la forma en que las salidas de las neuronas están dirigidas para convertirse en entradas de otras neuronas. La señal de salida de una neurona puede ser una entrada de otro elemento o incluso puede ser de sí mismo, como se observa en la figura 2-5.

Figura 2-4.: Perceptrón Multicapa: varias neuronas conectadas en red, una capa de entrada, una oculta y una capa de salida.



Fuente: "Redes multicapa". Autor: Scott Robinson (27)

Figura 2-5.: La imagen representa la forma de conexión entre neuronas, donde la imagen a representa la conexión hacía adelante, la imagen b conexión lateral y la imagen c la conexión atrás.



(a) Conexión adelante (b) Conexión lateral (c) Conexión atrás

Fuente: "Redes neuronales artificiales" (29)

Se denomina propagación hacia adelante, cuando la salida de la neurona no es una entrada del mismo nivel o del anterior, como se muestra en la figura 2-5.a ; hacia atrás, cuando las salidas pueden estar conectadas a niveles previos o al mismo nivel incluyendo la misma neurona, ver figura 2-5.c; las conexiones laterales se les denomina sistemas recurrentes, son redes de propagación hacia atrás que presentan ciclos o lazos cerrados, como se observa en la figura 2-5.b (29).

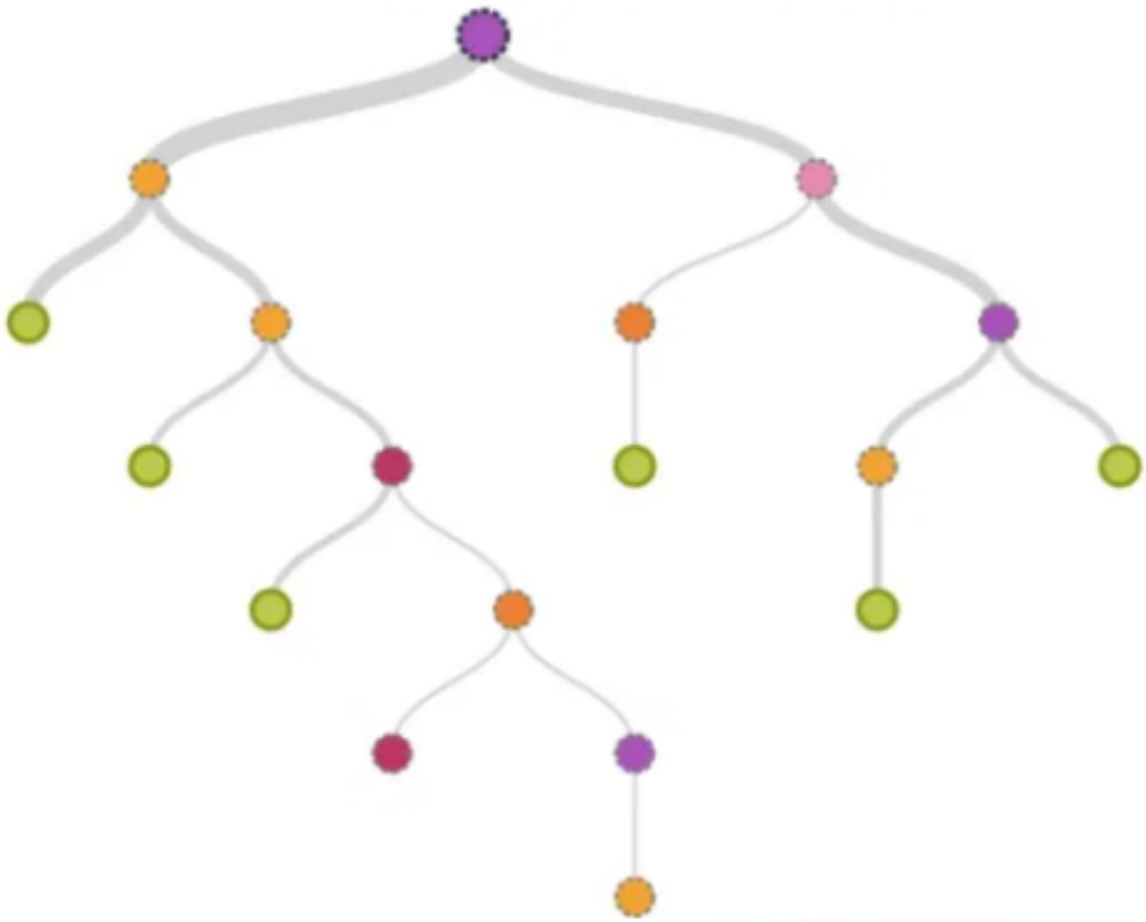
2.1.6. Bosques aleatorios (en inglés, Random Forest, RF)

Los árboles de decisión son pantallas gráficas de soluciones factibles a una decisión basadas en ciertas circunstancias; el primero nodo se le conoce como raíz (root) y luego se divide el resto de atributos de entrada en dos ramas planteando una condición que puede ser verdadera o falsa (ver figura 2-6). Entonces, una vez planteada la definición de árboles de decisión podemos decir que, bosques aleatorios son algoritmos de aprendizaje supervisado que, a partir de un conjunto de datos de entrenamiento, genera una cantidad de árboles de decisión, donde cada uno de estos árboles fue entrenado con un conjunto aleatorio de muestras, como se observa en la figura 2-7. Asimismo, en la división en vez de tener en cuenta las características para una buena división, se considera un subconjunto de estas. Scikit-Learn es la biblioteca de aprendizaje automático más completa de Python, en la implementación de la librería de Scikit-Learn, se calcula el valor medio de las predicciones probabilísticas. Como resultado de esa aleatoriedad el rumbo del modelo aumenta, sin embargo, con la combinación de los aprendices, reduce la varianza y nivela el efecto que genera un mejor resultado, el cual es más difícil de interpretar que el de un único árbol de decisión, lo que se considera como una desventaja. Scikit-Learn implementa este algoritmo para clasificación y regresión (28).

2.1.7. Máquinas de vectores de soporte (en inglés, Support Vector Machines, SVM)

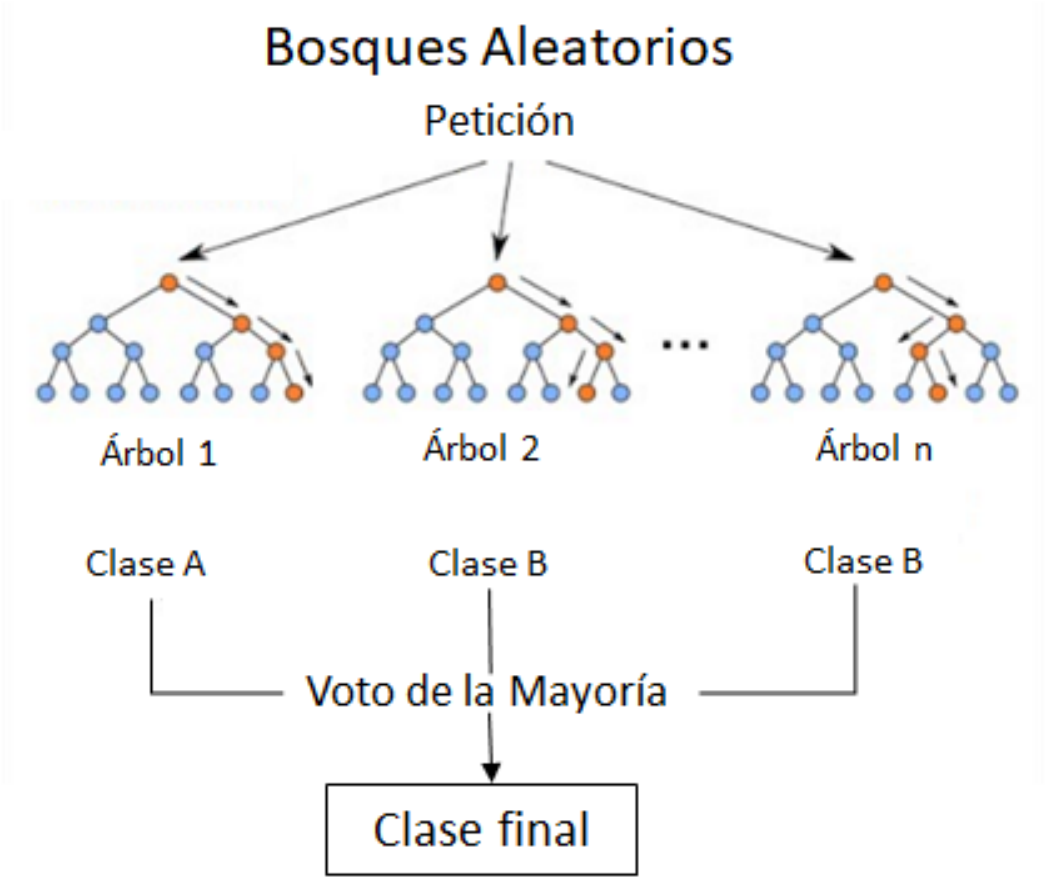
Son algoritmos que pertenecen al tipo aprendizaje supervisado, fueron desarrollados por los laboratorios de AT&T, en donde, se tiene un conjunto de puntos en un plano que pertenecen a dos clases diferentes, el objetivo es encontrar una recta (hiperplano), que pueda separar las dos clases de puntos, como se observa en la figura 2-8 presenta tres líneas que separan los puntos, pero solo la línea H3 es la mejor opción, ya que, maximiza la distancia mínima a los puntos negros y blancos, optimizando la capacidad de generalización del algoritmo. Los puntos más cerca a la recta seleccionada, para este caso H3, son los que determinan la posición de la recta y se le conoce como vectores de soporte (support vectors) y la suma de las distancias que separan a estos puntos del hiperplano de máximo margen se denomina margen (28).

Figura 2-6.: La imagen representa un algoritmo de árbol decisión en Python



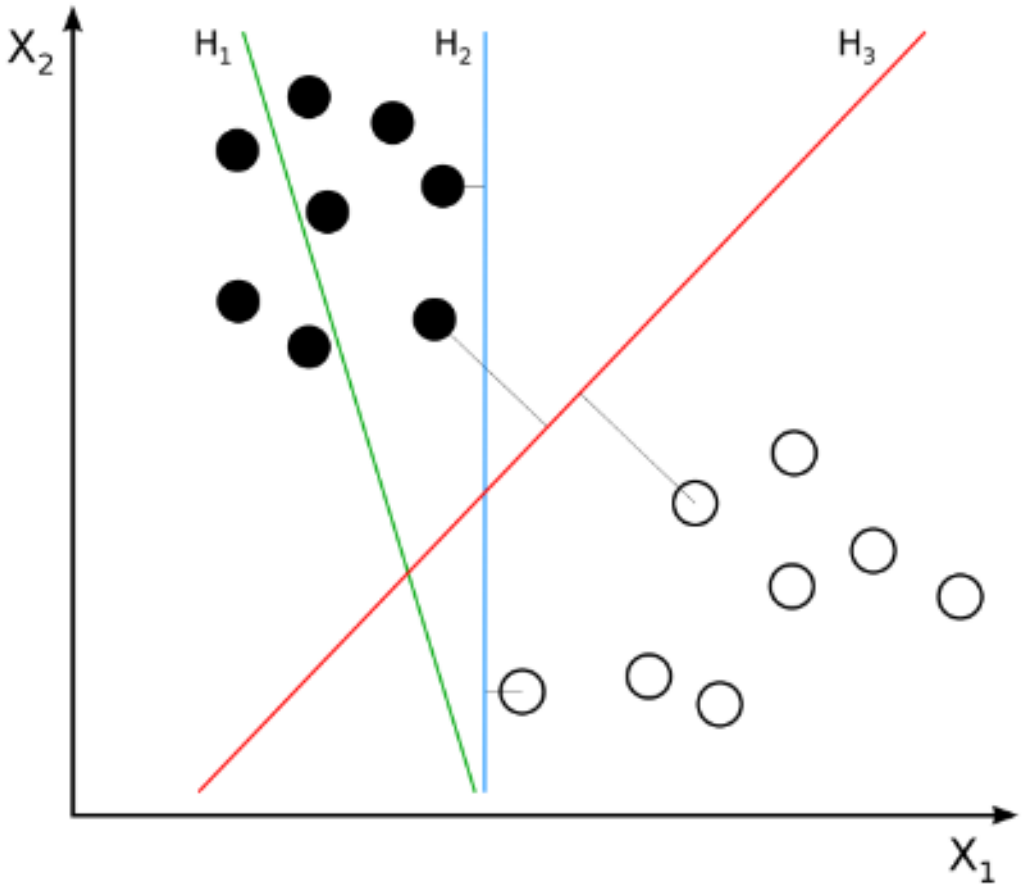
Fuente: árbol de decisión. Autor: Rahul Saxena (43)

Figura 2-7.: La imagen representa bosques aleatorios donde se observa un conjunto de árboles de decisión.



Fuente: medium.com. Autor: Yuri Puma (31)

Figura 2-8.: Máquinas de vectores de soporte: genera una recta en la mitad de los puntos que maximiza la distancia mínima a los puntos negros y blancos.



Fuente: “SVM”. Autor: Daniel Burrueco. (28)

2.1.8. Python

Python es un lenguaje de programación con estructuras de datos de alto nivel eficientes y orientado a objetos, es potente y fácil de aprender. Tiene una sintaxis elegante y es de tipado dinámico, un lenguaje ideal para scripting y desarrollo rápido de aplicaciones en muchas áreas, para la mayoría de plataformas. El intérprete de Python con su gran cantidad de librerías se encuentran disponibles libremente para la mayoría de plataformas desde la página oficial de Python, https://www.python.org/, se pueden distribuir libremente y en el mismo sitio se pueden encontrar distribuciones y direcciones a muchos módulos de Python, programas, herramientas y adicionalmente documentación (32).

2.1.9. Scikit-Learn

Es una biblioteca de aprendizaje automático y de código abierto de propósito general en Python. Proporciona eficientes, rápidas y sencillas implementaciones de algoritmos de vanguardia, presenta amplia cobertura de los métodos de aprendizaje automático; además, tiene un fuerte apoyo de la comunidad para la documentación, el seguimiento de errores y el control de calidad. Por último, la biblioteca impone una unificación de datos de entrada/salida y presenta un procedimiento de ajuste de modelo fijo, lo que hace que sea más fácil el cambio de un método a otro (33) (34).

Herramientas científicas básicas

Numpy: proporciona en ndarray tipo de datos en Python, en donde ndarray se refiere a un arreglo de N dimensiones donde N es cualquier número. Maneja la persistencia eficiente de una matriz para entrada y salida; además, proporciona operaciones básicas como el producto escalar (33).
Scipy: proporciona funciones matemáticas de alto nivel para muchos dominios, en los que se pueden incluir álgebra lineal, procesamiento de señales y optimización. Para lograr un mejor rendimiento scipy se encuentra vinculada con bibliotecas compiladoras como BLAS, Arpack y MKL (33).
Matplotlib: es una biblioteca para generación de gráficos, está integrada en la pila científica de Python. Matplotlib ofrece cifras con calidad de publicación y lo hace en diferentes formatos (33).

Conceptos asociados a las métricas de rendimiento

Conceptos de Scikit learn
En scikit-learn, los objetos y algoritmos aceptan datos de entrada en forma de matrices bidimensionales de tamaño X, lo que la hace independiente y genérico del dominio. Sin embargo,

los objetos de Scikit-learn comparten un conjunto uniforme de métodos que dependen de su propósito: los estimadores, los predictores y los transformadores (33).

Estimador: pueden ajustar modelos a partir de datos, es decir, expone un método *fit* para aprender los parámetros del modelo a partir de los datos de entrenamiento. Las tareas de aprendizaje automático, como la selección de características o la reducción de dimensionalidad; también, se proporcionan como estimadores (33).

Predictor: es un estadístico con propiedades especificas, con un método *predict* tomando una matriz con entrada X y genera nuevos datos para cada muestra. Es decir, hace predicciones sobre nuevos datos (33).

Transformador: convierten datos de una representación a otra. Para modificar o filtrar datos antes de alimentar a un algoritmo de aprendizaje, algunos estimadores, implementan un método transform. Se les conoce como transformadores a los algoritmos de preprocesamiento, selección de características y reducción de dimensionalidad. También existe un método para invertir la trasformación denominado inverse-transform (33).

Métricas de rendimiento

Se encargan de valorar el rendimiento de un modelo de aprendizaje automático. Su objetivo es estimar la precisión de la generalización de un modelo sobre los datos futuros; estas juegan un papel muy importante en problemas de clasificación ya que facilitan la elección del mejor algoritmo (35) (38).

Tabla de confusión: es una tabla de frecuencias donde las filas pertenecen a la clase predicha y las columnas son la clase real, representando el número de predicciones de cada clase mutuamente excluyente (35).

Tabla 2-3.: Matriz de confusión para variables dicotómicas o una variable con dos valores posibles

		Clasificador 1		
Clasificador 2		Clase 1: Positivo	Clase 2: Negativo	Total
	Clase 1: Positivo	f_{11}= Verdadero positivo	f_{10} = Falso negativo	F_1T
	Clase 2: Negativo	f_{01} = Falso positivo	f_{00} = Verdadero negativo	F_0T
		F_1T	f_0T	N

Fuente: Borja (2020) (35)

La tabla 2-3 ejemplifica el caso de clasificación binaria, donde:

- Positivo y Negativo: es una forma de clasificación binaria, por ejemplo, con los datos de los pacientes se puede construir un modelo que prediga si un paciente tiene una enfermedad (Positivo) o no la tiene (Negativo).
- Verdaderos Positivos (f_{11}): cuando el clasificador 2 es 1 (Verdadero) y el clasificador 1 es también 1 (Verdadero)

- Verdaderos Negativos (f_{00}): cuando el clasificador 2 es 0 (Falso) y el pronosticado también es 0 (Falso).
- Falsos Positivos (f_{01}): cuando el clasificador 2 es 0 (Falso) y el pronosticado es 1 (True).
- Falsos Negativos (f_{10}): Cuando el clasificador 2 es 1 (Verdadero) y el valor predicho es 0 (Falso) (39).
- (f_{11} y f_{00}): positivo y negativo corresponden a las instancias clasificadas en cada clase, donde clase hace referencia a la clasificación positivo o negativo.
- F y f: representa las frecuencias marginales correspondientes a: F_1T - F_0T es igual a la cantidad total de elementos en cada clase.
- f_1T – f_0T: es el total de instancias clasificadas por el algoritmo como positivo y negativo.
- N: representa el tamaño muestral utilizado para la clasificación (35).

Accuracy: es la métrica más utilizada debido a su comodidad de cálculo y conocimiento para evaluar la validez general del algoritmo (35).
Recall o sensibilidad: permite conocer los correspondientes de casos positivos que fueron correctamente clasificados. Un modelo perfecto de recall es igual a 1 para cada clase (35).
La medida F1-score: une las métricas accuracy con recall, es directamente proporcional al aumento de las dos medidas, por lo cual valores altos en la métrica F1-score comprueba que el algoritmo de clasificación predice de manera ideal la clase positiva (35).
La curva ROC: es una gráfica bidimensional de la sensibilidad versus (1 – especificidad) para cada clase. La medida es la AUC que corresponde al área bajo la curva ROC los valores están entre 0 y 1, un caso aleatorio donde la AUC igual a 0,5 (35).
Índice kappa: es una métrica utilizada para el análisis de concordancia entre dos observadores humanos (35).
La tabla 2-4 presenta las fórmulas que se usan para el cálculo de cada métrica en un intervalo de (0,1).

2.1.10. Método de selección de variables

También conocido como reducción de dimensionalidad, antes que nada, vamos a ver que es dimensionalidad. La dimensionalidad son las características o variables de entrada que presenta un conjunto de datos. La tabla 2-5 muestra una parte del conjunto de datos de una veterinaria que contiene tres características. Entonces, el número de dimensiones es tres. Por ejemplo, para demostrar el primer punto de datos en el espacio de tres dimensiones, usamos la notación p1 (Juan Garrido, Gogo, Gato) (40).

Tabla 2-4.: Métricas para la evaluación del rendimiento en clasificadores

Métrica	Fórmula	Descripción
Accuracy	$\frac{f_{11}+f_{00}}{N}$	Proporción de clasificaciones predichas de manera correcta sobre el total de instancias.
Recall (sensibilidad)	$\frac{f_{11}}{f_{11}+f_{01}}$	Proporción de casos positivos bien clasificados.
Especificidad	$\frac{f_{00}}{f_{00}+f_{10}}$	Proporción de casos negativos bien clasificados.
1-especificidad	$\frac{f_{10}}{f_{10}+f_{00}}$	Proporción de casos positivos mal clasificados (error Tipo I).
F1 – score	$2 * \frac{accuracy*recall}{accuracy+recall}$	Media armónica de las métricas accuracy y recall.
Índice kappa	$k = \frac{Po-Pe}{1-Pe}$	Po = proporción de accuracy observado. Por lo tanto Po = accuracy.
Pe	$\frac{F_{1T}*f_{1T}+F_{0T}*f_{0T}}{N^2}$	Proporción de accuracy esperado por puro azar.
AUC	$\frac{recall-(1-especificidad)+1}{2}$	Probabilidad de clasificar correctamente una clase positiva al azar más que una negativa escogida al azar.

Fuente: Borja (2020) (35)

Tabla 2-5.: Creación de variable dummy a partir de la variable mascota. Tabla de referencia.

	Nombre Dueño	Nombre mascota	Mascota
1	Juan Garrido	Gogo	Gato
2	Lili Torres	Firulais	Perro
3	Santiago Puentes	Snack	Tortuga
4	Valeria Torrado	Nemo	Pez
5	Gabriel Fuentes	Michi	Gato
6	Karen Vivas	Veloz	Tortuga

Fuente: Elaboración propia

Una vez entendido el concepto de dimensionalidad se puede definir la reducción de la dimensionalidad, se trata de reducir la cantidad de características en un conjunto de datos, lo que se hace es que los algoritmos de selección de variables proyectan datos de dimensiones elevadas en un espacio de disminución de dimensiones y conservan la máxima cantidad de información posible. Al reducir la dimensionalidad de los datos, se mejora el rendimiento de los algoritmos de aprendizaje automático (40).

Análisis de componentes principales (en inglés, *Principal Component Analysis (PCA)*)
Es una técnica de reducción de dimensionalidad, el algoritmo busca encontrar una representación de baja dimensión de los datos, una vez encontrada retiene la mayor cantidad de variación posible. Es decir, busca la correlación entre características, si esta es muy alta entre un subconjunto de las características, PCA intentará combinar esas características y representar estos datos con un número menor de características (40).

Eliminación recursiva de características (en inglés, *Recursive Feature Elimination (RFE)*
Es una técnica de reducción de dimensionalidad, es el proceso recursivo que ordena las variables teniendo en cuenta una medida de significancia para cada variable, que es dada por un clasificador. Con cada repetición que se realice, se mide la importancia de las variables y la menos importante se suprime. Cuando se requiere de realizar el proceso más rápido se remueve un grupo de variables (41).

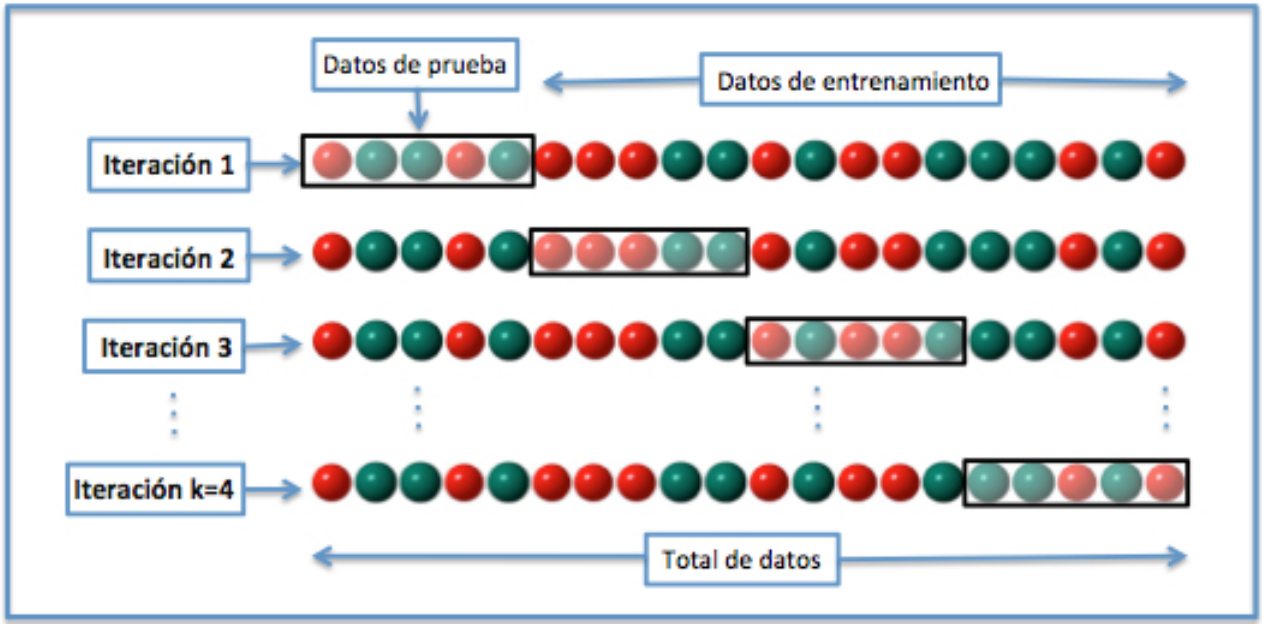
2.1.11. Validación cruzada

Los datos son divididos en dos grupos: un bloque de entrenamiento y otro de validación; al realizar este proceso se presenta un problema y es que se le están escondiendo datos al modelo en el grupo de validación y si no se tienen los datos suficientes al hacer este proceso se agrava aún más el problema. Por este motivo se crea la validación cruzada, ya que resuelve el inconveniente al aumentar la complejidad del cálculo, en donde divide el bloque inicial de entrenamiento en dos subbloques (A y B), se entrenan los dos bloques de la siguiente forma: uno se entrena con los datos de A y se confirma con B y otro se entrena con B y se valida con A, se combinan para llegar a una predicción final (28).
Validación cruzada de k iteraciones
En la validación cruzada de K iteraciones los datos se dividen en K bloques. Un bloque para prueba y los demás para entrenamiento. Se repite durante k iteraciones con cada uno de los bloques de prueba y sus resultados se combinan (ver figura 2-9). K suele tomar valores como: 2, 3, 5 o 10 (28).

2.2. Antecedentes o Estado del Arte

Dentro de la literatura se encuentran numerosos resultados en cuanto a la predicción de la diabetes, haciendo uso de diferentes algoritmos y obteniendo diferentes resultados de acuerdo a los parámetros tomados en cuenta para la predicción. Para encontrar o crear algo realmente robusto, que tenga la capacidad predictiva, se debe depender de un programa de análisis predictivo que incluyan los almacenes de datos, algoritmos de análisis e informes que proporcionen decisiones óptimas (13).

Figura 2-9.: Validación cruzada de K iteraciones donde K=4.



Fuente: "Validación cruzada". Autor: Daniel Burrueco.(28)

En primera instancia se realizó un estudio de la predicción de diabetes de tipo 2 en adultos jóvenes en presencia de componentes heredo familiares utilizando el "*FINnish Diabetes Risk Score*" (FINDRISC). El objetivo del estudio fue predecir el riesgo y determinar los factores de riesgo ambientales para desarrollar diabetes tipo 2 en adultos jóvenes con componente heredo-familiar. La metodología utilizada, fue un estudio descriptivo, en donde se realizó examen físico, interrogatorio y laboratorio a 40 pacientes no diabéticos de ambos sexos, entre 20 y 45 años con historia familiar para diabetes tipo 2. La finalidad del estudio fue que el 73 % de los sujetos con historial familiar con diabetes tienen entre 4 % a 33 % de riesgo para desarrollar esa enfermedad en los próximos 10 años; esta probabilidad pudiera reducirse al modificar los hábitos de alimentación y actividad física (18). El estudio mencionado se diferencia con el presente trabajo de grado en que no se tienen en cuenta características heredo familiares, sin embargo, cabe la posibilidad de que sí tengan familiares diagnosticados, hay pacientes que ya tienen la enfermedad y otros que no presentan diabetes, no se va a usar el *FINnish Diabetes Risk Score* (FINDRISC).

Del mismo modo, un segundo estudio, que hace la validación del FINDRISC (*FINnish Diabetes Risk SCore*) para la predicción del riesgo de diabetes tipo 2 en una población del sur de España. Un estudio con vista futura, de base poblacional desarrollado en la población de Pizarra (Málaga). Durante la primera fase del estudio que se realizó en los años 1997-

1998 en el cual había un total de 1.051 personas con edades entre 18-65 años que fueron seleccionados de forma aleatoria del censo municipal de la localidad. Entre los años 2003 y 2004 los participantes en el primer estudio fueron nuevamente evaluados. Un total de 824 individuos (78,4 %) completaron esta segunda fase del estudio. En las dos fases del estudio se administró una sobrecarga oral de glucosa a todos los participantes que fueron diagnosticados con diabetes. Se evaluó la capacidad del FINDRISC para detectar la diabetes tipo 2 no diagnosticada (primera fase: estudio transversal) y en la predicción de la incidencia de diabetes tipo 2 (segunda fase: estudio de cohortes). Finalmente, los resultados que mostró el test, fueron buenos resultados tanto para detectar diabetes tipo 2 no diagnosticada. La mejor predicción de riesgo de diabetes tipo 2 incidente se encontró en los sujetos con glucemia en ayunas $>$ 100 mg/dl y un FINDRISC mayor o igual 9 (odds ratio [OR]: 19,37; intervalo de confianza del 95 % [IC 95 %]: 8,86-42,34; p $<$ 0,0001). Las conclusiones, los resultados del estudio muestran que el FINDRISC puede ser una herramienta útil para detectar sujetos con alto riesgo de diabetes en esta población (19). El estudio mencionado se diferencia con el presente trabajo de grado en que los pacientes ya presentan la enfermedad, no se le va a suministrar glucosa para su ingesta, no se va a usar el *FINnish Diabetes Risk Score* (FINDRISC).

Por otro lado, una tesis doctoral, se ha realizado como propuesta de algoritmos de predicción de glucosa en pacientes diabéticos, en donde, proponen diferentes algoritmos de predicción de glucemia para pacientes con diabetes, basados en una información registrada por un sistema de monitorización continua de glucosa, así como incorporando la información de la insulina administrada y la ingesta de carbohidratos en los pacientes. Los algoritmos propuestos han sido evaluados en simulación y utilizando datos de pacientes registrados en diferentes estudios clínicos. Para ello se ha desarrollado una amplia metodología, que trata de caracterizar las prestaciones de los modelos de predicción desde todos los puntos de vista, en los cuales se encuentra: precisión, retardo, ruido y capacidad de detección de situaciones de riesgo. Se han desarrollado las herramientas de simulación necesarias y se han analizado y preparado las bases de datos de pacientes. También, se ha probado uno de los algoritmos propuestos para comprobar la validez de la predicción en tiempo real en un escenario clínico. Se han desarrollado las herramientas que han permitido llevar a cabo el protocolo experimental definido, en el que el paciente consulta la predicción bajo demanda y tiene el control sobre las variables metabólicas. Como resultado, este experimento ha permitido valorar el impacto sobre el control glucémico del uso de la predicción de glucosa en pacientes diabéticos (20). El estudio mencionado se diferencia con el presente trabajo de grado en que se va hacer uso de varios modelos de aprendizaje, los datos que usaron en el estudio son los registrados por los sensores instalados en los pacientes y en la investigación es un conjunto de datos público.

Un estudio adicional propone un sistema predictivo para diagnosticar la diabetes basado en un clasificador Bayesiano. Los clasificadores bayesianos permiten clasificar eventos discre-

tos y limitados (variables independientes) en un número determinado de clases, definiendo una función estadística para cada clase. En la definición de estas funciones estadísticas toman como referencia una base de datos de entrenamiento. Con base en estas funciones definidas, el sistema podrá clasificar un nuevo conjunto de variables independientes (datos de prueba) y establecer la clase a la que pertenecen, con base en la función estadística que genera el mayor valor. En el caso de enfermedades, el sistema opera a partir de una serie de parámetros simples tomados del paciente y no exige análisis de laboratorio. La metodología utilizada estuvo compuesta por los siguientes pasos: 1) Definición de las bases de datos; 2) Estructuración matemática del clasificador bayesiano; 3) Construcción del algoritmo de clasificación; y 4) Validación del sistema. Los resultados obtenidos, el clasificador bayesiano para identificar personas con diabetes se propone como una herramienta de utilidad para la detección temprana de esta enfermedad, sin recurrir a pruebas de laboratorio. En particular, el nivel de acierto del sistema puede variar entre un 87.6 % y un 96.9 % en dependencia del número de características del paciente que sean analizadas. Por otro lado, la detección de personas con la enfermedad oscila entre el 92.3 % y 98.4 %, al tiempo que la detección de personas sin la enfermedad oscila entre el 81.53 % y 95.38 %. Como futuras líneas de investigación se plantea adaptar el clasificador bayesiano con el fin de que se pueda emplear para detectar tempranamente diversas enfermedades, aspecto que le dará más versatilidad para abordar una mayor diversidad de problemas en el área de la salud (21). Esta investigación se diferencia del presente trabajo de grado en cuanto a algoritmos, ya que no se enfocará explícitamente en clasificación bayesiano, contará con la investigación y el uso de más de un algoritmo de predicción.

Finalmente, el estudio por el cual se basa el presente trabajo de grado. Los autores hacen uso del conjunto de datos que recopiló el Hospital de Sylhet, Bangladesh, este conjunto de datos contiene la información de síntomas de pacientes del hospital. Los datos alimentan a los algoritmos de predicción, en el estudio se utilizaron los modelos: Naive Bayes, Árboles de decisión, Regresión logística y bosques aleatorios. En seguida, realizaron un preprocesamiento que fue ignorar tuplas con valores incompletos. Luego, realizaron técnicas de validación cruzada y división porcentual diez veces mayor. Entonces, se toman los síntomas de los pacientes como entrada, el mejor algoritmo construye el sistema utilizando el conjunto de datos como base de datos. El resultado obtenido en el estudio fue satisfactorio para bosques aleatorios, en donde, 97 % se clasifico correctamente usando validación cruzada y el 99 % usando división porcentual (1) .

La tabla 2-6 presenta una comparación de los estudios entre sí y con el presente trabajo de grado, en donde, ✓significa que este ítem se utilizó en el estudio y x significa que esa información o ese dato no se especifica en el estudio.

Tabla 2-6.: Tabla de comparación de estudios mencionados anteriormente vs el presente trabajo de grado. La columna con nombre trabajo de grado hace referencia a las características del presente trabajo de grado.

Comparación						
Datos	Aldana (2011) (18)	Soriguer (2012) (19)	Pérez (2014) (20)	Castrillón (2017) (21)	Islam (2020) (1)	Trabajo de grado
N° pacientes	40	1051	10	x	520	520
Diagnosticados	x	x	x	✓	✓	✓
No Diagnosticados	✓	✓	x	✓	✓	✓
Edad	20 - 45	18 - 65	12 - 40	x	20 - 65	20 - 65
Modelos	x	x	Redes neuronales	clasificador bayesiano	arboles, bosques aleatorios, bayes, regresión	vectores, bosques aleatorios, redes neuronales

Fuente: elaboración propia

2.3. Marco Legal

A continuación, se especificarán algunas leyes que aplican el desarrollo de este proyecto.

- Ley de propiedad intelectual: el artículo 61 de la Constitución Política de Colombia consagra la protección de la propiedad intelectual. En tanto que la Decisión Andina 351 de 1993, Decisión 344 de 1993 de Cartagena, la Ley 23 de 1.982 y la ley 44 de 1993, así como la ley 1455 de 2011 por medio de la cual se aprueba el Protocolo de Madrid sobre el registro internacional de marcas, conforman las principales normas sobre el derecho de autor y los derechos conexos en Colombia. Estas normas recogen los principios consagrados en los tratados internacionales más importantes en materia del derecho de autor y los derechos conexos que ha suscrito nuestro país, como el Convenio de Berna para la protección de las obras literarias y artísticas, la Convención de Roma sobre la protección de los artistas, intérpretes o ejecutantes, los productores de fonogramas y los organismos de radiodifusión, así como el tratado de la OMPI. De igual manera, el Código Penal consagra las sanciones a quienes vulneren el derecho de autor y los derechos conexos (17).

- Habeas Data: a través de la Ley 1581 de 2012 y el Decreto 1377 de 2013, se desarrolla el derecho constitucional que tienen todas las personas a conocer, suprimir, actualizar y rectificar todo tipo de datos personales recolectados, almacenados o que hayan sido

objeto de tratamiento en bases de datos en las entidades públicas y privadas. La Corte Constitucional lo definió como el derecho que otorga la facultad al titular de datos personales de exigir de las administradoras de esos datos el acceso, inclusión, exclusión, corrección, adición, actualización y certificación de los datos, así como la limitación en las posibilidades de su divulgación, publicación o cesión, de conformidad con los principios que regulan el proceso de administración de datos personales. Asimismo, ha señalado que este derecho tiene una naturaleza autónoma que lo diferencia de otras garantías con las que está en permanente relación, como los derechos a la intimidad y a la información (15).

- Derecho a la intimidad: el artículo 15 de la Constitución de 1991 establece que toda persona tiene derecho a la intimidad personal y familiar. Reza: "La correspondencia y demás formas de comunicación privada son inviolables. Sólo pueden ser interceptadas o registradas mediante orden judicial, en los casos y con las formalidades que establezca la ley"(16).

El presente trabajo respetó lo indicado por las leyes mencionadas anteriormente.

3. Metodología

3.1. Hipótesis

Aplicando otras técnicas de aprendizaje y teniendo en cuenta otras parametrizaciones, se podrán obtener modelos más precisos que los que se conocen hasta ahora y se pueden obtener mejores resultados en las métricas ROC y F1.

3.2. Metodología de la investigación

3.2.1. Recopilación de datos

El conjunto de datos a utilizar para el entrenamiento de los algoritmos, es de uso público y no fue recopilado por los participantes de esta investigación. Se optó por utilizar un conjunto de datos recopilado por otros investigadores y de uso público.
El conjunto de datos incluye información de los síntomas relacionados con la diabetes, incluidos síntomas que pueden causar la enfermedad. Datos sobre 520 personas. El conjunto de datos fue creado al diligenciar un cuestionario con personas que recientemente fueron diagnosticadas con diabetes o que aún no son diabéticas pero tienen algunos o muchos de los síntomas. El cuestionario fue realizado a los pacientes directamente por Sylhet Diabetes Hospital de Sylhet, Bangladesh.

3.2.2. Preprocesamiento de datos

El preprocesamiento que se ha utilizado en este trabajo de grado, fue realizar el cambio de variables categóricas por variables dummy, la idea es crear una columna para cada valor distinto que exista en la característica que se está codificando y, para cada registro, marcar con un 1 la columna a la que pertenezca dicho registro y dejar las demás con 0. En la tabla 3-1 se muestra un fragmento de conjunto de datos que se uso para la investigación y en la figura 3-1 se visualiza el nuevo archivo generado con las variables dummy.
Una vez realizado el proceso de creación de variables dummy, el conjunto de datos queda con muchas características o variables nuevas; de tener 17 variables, el conjunto de datos paso a tener 32 variables. Con tantas variables el resultado del modelo puede no ser el esperado, entonces se procede con la reducción de dimensionalidad en donde se pretende llegar a tener la mínima cantidad de variables. Se ejecuta el código que se encuentra en la carpeta del

Tabla 3-1.: Fragmento del conjunto de datos de la presente investigación

Age	Gender	Polyuria	Polydipsia	weakness	Obesity	class
40	Male	No	Yes	No	Yes	Positive
58	Male	No	No	No	Yes	Positive
41	Male	Yes	No	No	Yes	Positive
60	Male	Yes	Yes	Yes	Yes	Positive
50	Female	No	Yes	Yes	Yes	Positive
38	Female	No	No	No	No	Positive

Fuente: elaboración propia

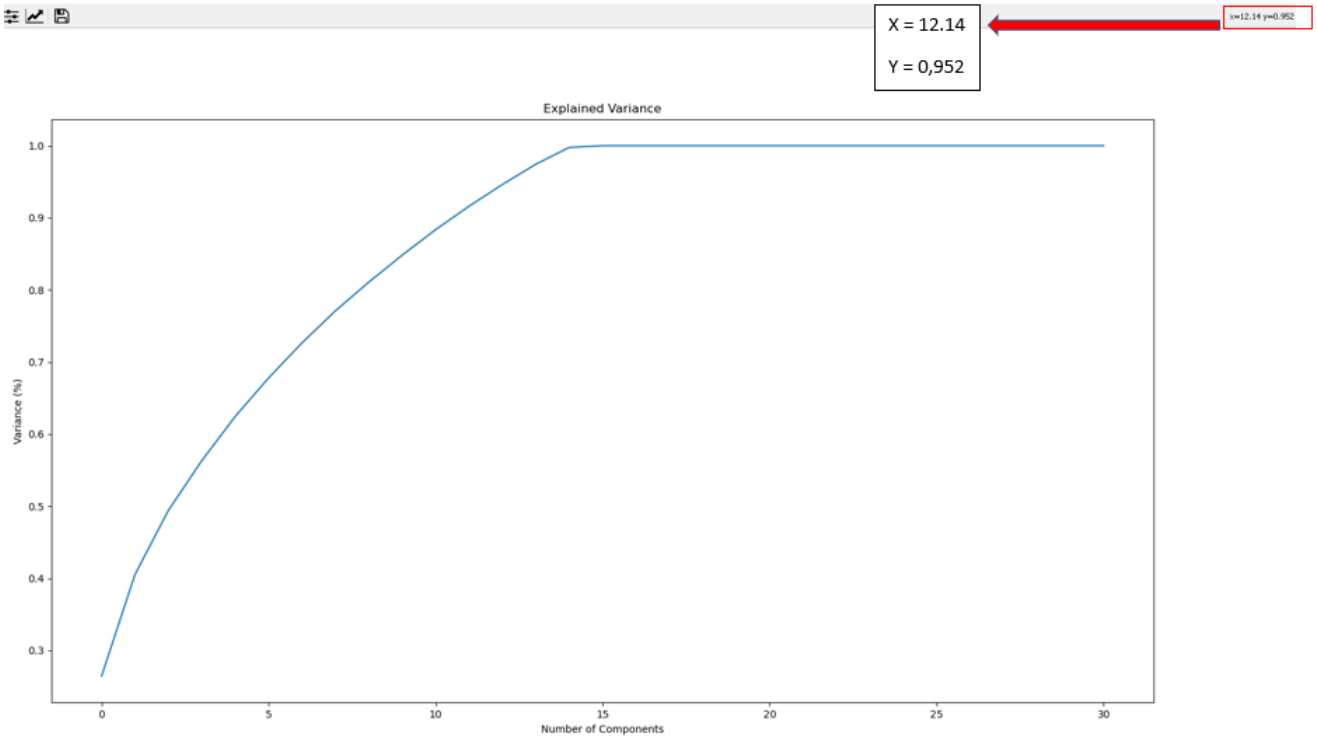
proyecto con nombre plot_variance, que indica cual es la mínima cantidad de variables como se observa en la figura 3-2, para este caso son 12.

Figura 3-1.: Fragmento de nuevo conjunto de datos generado a partir de variables dummy.

sudden weig	Polydipsia_No	Polydipsia_Yes	Polyuria_No	Polyuria_Yes	Gender_Female	Gender_Male	Age	class
0	0	1	1	0	0	1	40	1
0	1	0	1	0	0	1	58	1
0	1	0	0	1	0	1	41	1
1	1	0	1	0	0	1	45	1
1	0	1	0	1	0	1	60	1
0	0	1	0	1	0	1	55	1
0	0	1	0	1	0	1	57	1
1	0	1	0	1	0	1	66	1
0	0	1	0	1	0	1	67	1
1	0	1	1	0	0	1	70	1
0	0	1	0	1	0	1	44	1
0	0	1	0	1	0	1	38	1
0	1	0	0	1	0	1	35	1
1	0	1	0	1	0	1	61	1
0	0	1	0	1	0	1	60	1
0	0	1	0	1	0	1	58	1

Fuente: elaboración propia

Figura 3-2.: Imagen que indica la cantidad de variables que necesita el modelo. Es decir, cuando el valor de Y este más próximo a 1 quiere decir que se toma el valor que genere en X. Entre más alejado sea el valor de 1 para Y, menos preciso puede llegar a ser el modelo.



Fuente: elaboración propia

4. Resultados obtenidos

En el presente trabajo de grado se realizaron pruebas con los modelos redes neuronales, máquinas de vectores de soporte y bosques aleatorios, utilizando validación cruzada con iteraciones de 10 y 5 veces. Usando la búsqueda en Grid (param_grid o lista de diccionarios), los resultados obtenidos se mencionan a continuación.
Param_grid: diccionario con nombres de parámetros (str) como claves y listas de ajustes de parámetros a probar como valores, en cuyo caso se exploran las cuadrículas abarcadas por cada lista (42).

4.1. Máquinas de vectores de soporte

Una vez realizado en preprocesamiento, se realizaron pruebas con el modelo máquinas vectores de soporte usando métricas de rendimiento, en donde se obtuvieron los resultados que se pueden observar en la tabla 4-1.

Tabla 4-1.: Resultados para cada una de las métricas en el modelo SVM

	ROC AUC	F1	Precisión	Recall	Accuracy	Balanced Accuracy	Average precisión
k=10	0.986875	0.976837	0.976522	0.978125	0.971154	0.969063	0.991119
k=5	0.986406	0.976775	0.976157	0.978125	0.971154	0.969063	0.990068

Fuente: elaboración propia

Los parámetros que mejor se adaptan para el modelo usando la búsqueda en grid, se muestran en la tabla 4-2

Tabla 4-2.: Mejores parámetros para el modelo SVM

	C	Gamma	Kernel
Máquinas de vectores de soporte	10	Scale	rbf

Fuente: elaboración propia

C: parámetro de regularización. Debe ser estrictamente positiva.
Gamma: define cuánta influencia tiene un solo ejemplo de formación. Puede elegir entre los

parámetros scale, auto. Cambiado en la versión 0.22 de scikit learn: El valor por defecto de gamma cambió de auto a scale.
Kernel: es una función matemática que transforma un espacio de pocas dimensiones en un espacio de dimensiones mayores. Debe ser uno de los siguientes: lineal, poly, rbf, sigmoide, precomputed. Si no se indica ninguno, se utilizará 'rbf' por defecto (42).

4.2. Bosques aleatorios

Una vez realizado en preprocesamiento, se realizaron pruebas con el modelo bosques aleatorios usando métricas de rendimiento, en donde se obtuvieron los resultados que registra la tabla 4-3.

Tabla 4-3.: Resultados para cada una de las métricas en el modelo bosques aleatorios

	ROC AUC	F1	Precisión	Recall	Accuracy	Balanced Accuracy	Average precisión
k=10	0.990469	0.975374	0.976186	0.97500	0.969231	0.974687	0.993344
k=5	0.986406	0.975334	0.970037	0.98125	0.965385	0.966562	0.992572

Fuente: elaboración propia

Los parámetros que mejor se adaptan para el modelo usando la búsqueda en grid, se visualizan en la tabla 4-4.

Tabla 4-4.: Mejores parámetros para el modelo bosques aleatorios

	Bootstrap	Criterion	Max_features	Min_samples_split
Bosques aleatorios	True	Gini	4	8

Fuente: elaboración propia

Bootstrap: hace que el modelo forme muestras aleatorias con reemplazo. Los parámetros son: falso y verdadero, la opción falso si se usa todo el árbol y verdadero si se usan muestras.
Criterion: es la función con la cual se mide la calidad de una división. Los parámetros son: gini y entropy, gini para la impureza de Gini y entropía para la ganancia de información.
Max_features: indica cuantas características se deben considerar para buscar la mejor división. Los parámetros son: auto, sqrt, log2, en donde: si es auto, entonces max_features=sqrt(n_features), sqrt, entonces max_features=sqrt(n_features) (igual que auto), log2, entonces max_features=log2(n_featu
Min_samples_split: indica el número mínimo de muestras que se necesitan para dividir un nodo interno (42).

4.3. Redes Neuronales

Una vez realizado en preprocesamiento, se realizaron pruebas con el modelo redes neuronales usando métricas de rendimiento, en donde se obtuvieron los resultados que se visualizan en la tabla 4-5.

Tabla 4-5.: Resultados para cada una de las métricas en el modelo redes Neuronales

	ROC AUC	F1	Precisión	Recall	Accuracy	Balanced Accuracy	Average precisión
k=10	0.966875	0.941425	0.956063	0.928125	0.928846	0.929063	0.981607
k=5	0.970313	0.9946807	0.954209	0.940625	0.934615	0.932813	0.983348

Fuente: elaboración propia

Los parámetros que mejor se adaptan para el modelo usando la búsqueda en grid, se observan en la tabla 4-6.

Tabla 4-6.: Mejores parámetros para el modelo de redes neuronales

	Activation	alpha	Hidde_layer_sizes	Learning_rate_init	Max_iter	Random_state	Solver
Redes neuronales	relu	0.0001	75	0.001	5	1	lbfgs

Fuente: elaboración propia

Activation: es la función encargada de activar la capa oculta. Los parámetros son: identity, logistic, tanh, relu, identity, en donde: activación, no-op, útil para implementar el cuello de botella lineal, devuelve f(x) = x; logistic, la función sigmoidea logística, devuelve f(x) = 1 / (1 + exp(-x)); tanh, la función hiperbólica tan, devuelve f(x) = tanh(x); relu, la función lineal unitaria rectificada, devuelve f(x) = max(0, x).
alpha: hace referencia a un parámetro de penalización L2. Por defecto va el parámetro 0,0001.
Hidde_layer_sizes: el elemento representa el número de neuronas en la capa oculta.
Learning_rate_init: hace referencia a la tasa de aprendizaje inicial utilizada y además, se encarga de controlar el tamaño del paso al actualizar los pesos. Por defecto va el parámetro 0,0001.
Max_iter: indica cual va a ser el número máximo de iteraciones.
Random_state: indica la generación de números aleatorios para las circunspecciones y el principio de sesgo.
Solver: Se encarga de la optimización del peso. Los parámetros son: lbfgs, sgd, adam, en donde: lbfgs, es un optimizador de la familia de los métodos quasi-Newton; sgd, se refiere al descenso de gradiente estocástico; adam, se refiere a un optimizador basado en el gradiente

estocástico propuesto por Kingma, Diederik y Jimmy Ba (42).

4.4. Resultados con las métricas de rendimiento

En el presente trabajo de grado se realizaron pruebas con las métricas ROC AUC, F1, *Recall*, *Precision*, *Accuracy*, *Balanced Accuracy*, *Avergae Precision*, los resultados obtenidos con respecto a métricas se mencionan a continuación.

ROC AUC

En términos para la curva ROC AUC los resultados generados se presentan en la tabla 4-7. Como se observa en la figura 4-1, se tiene un mejor resultado con el modelo de bosques aleatorios ya que presenta un porcentaje de 0,990469 %, con validación cruzada de k=10.

Tabla 4-7.: Resultados obtenidos en cada modelo para la métrica curva ROC AUC

Modelo	Iteración	ROC AUC
Redes	k=10	0,966875
	k=5	0,970313
Bosques	k=10	0,990469
	k=5	0,986406
Vectores	k=10	0,986875
	k=5	0,986406

Fuente: elaboración propia

F1

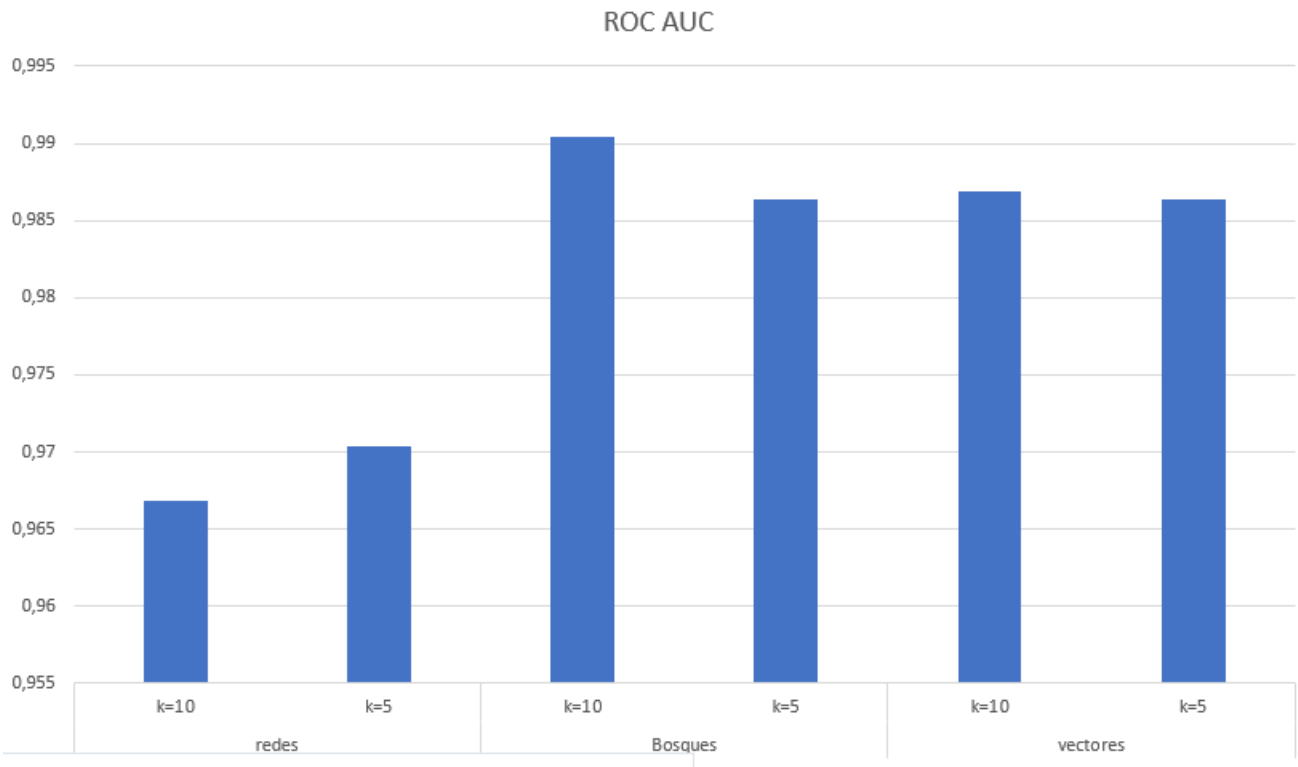
En términos de la métrica F1 se generaron los resultados que se presentan en la tabla 4-8. Como se observa en la figura 4-2, se tiene un mejor resultado con el modelo de redes neuronales ya que presenta un porcentaje de 0,9946807 %, con validación cruzada de k=5.

Tabla 4-8.: Resultados obtenidos en cada modelo para la métrica F1

Modelo	Iteración	F1
Redes	k=10	0,941425
	k=5	0,9946807
Bosques	k=10	0,975374
	k=5	0,975334
Vectores	k=10	0,976837
	k=5	0,976775

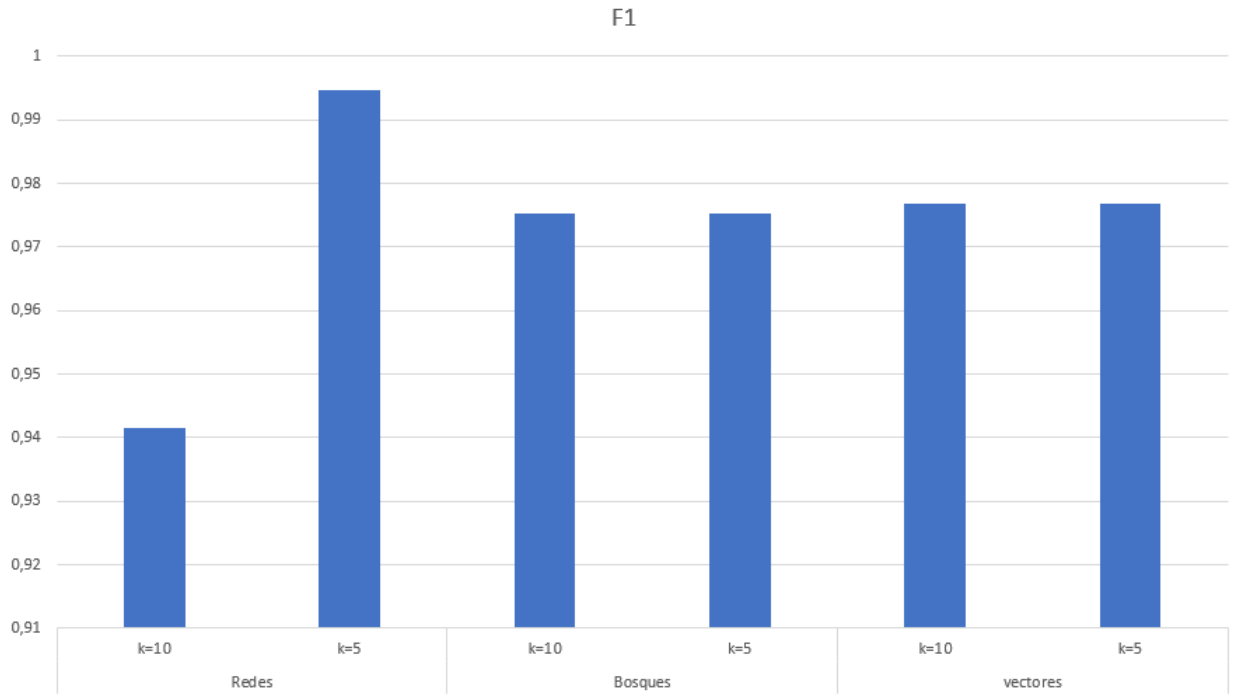
Fuente: elaboración propia

Figura 4-1.: Gráfico de barras de los resultados de la métrica ROC AUC.



Fuente: elaboración propia

Figura 4-2.: Gráfico de barras de los resultados de la métrica F1.



Fuente: elaboración propia

Precision

Los resultados generados para la métrica *Precision* se presentan en la tabla 4-9. Como se observa en la figura 4-3, se tiene un mejor resultado con el modelo de máquinas de vectores de soporte ya que presenta un porcentaje de 0.976522 %, con validación cruzada de k=10.

Tabla 4-9.: Resultados obtenidos en cada modelo para la métrica *Precision*

Modelo	Iteración	Precision
Redes	k=10	0,956063
	k=5	0.954209
Bosques	k=10	0.976186
	k=5	0.970037
Vectores	k=10	0.976522
	k=5	0.976157

Fuente: elaboración propia

Recall

En términos para la métrica *Recall* se generaron los resultados que se presentan en la tabla 4-10. Como se observa en la figura 4-4, se tiene un mejor resultado con el modelo de bosques aleatorios ya que presenta un porcentaje de 0,98125 %, con validación cruzada de k=5.

Tabla 4-10.: Resultados obtenidos en cada modelo para la métrica *Recall*

Modelo	Iteración	Recall
Redes	k=10	0,928125
	k=5	0.940625
Bosques	k=10	0.97500
	k=5	0.98125
Vectores	k=10	0.978125
	k=5	0.978125

Fuente: elaboración propia

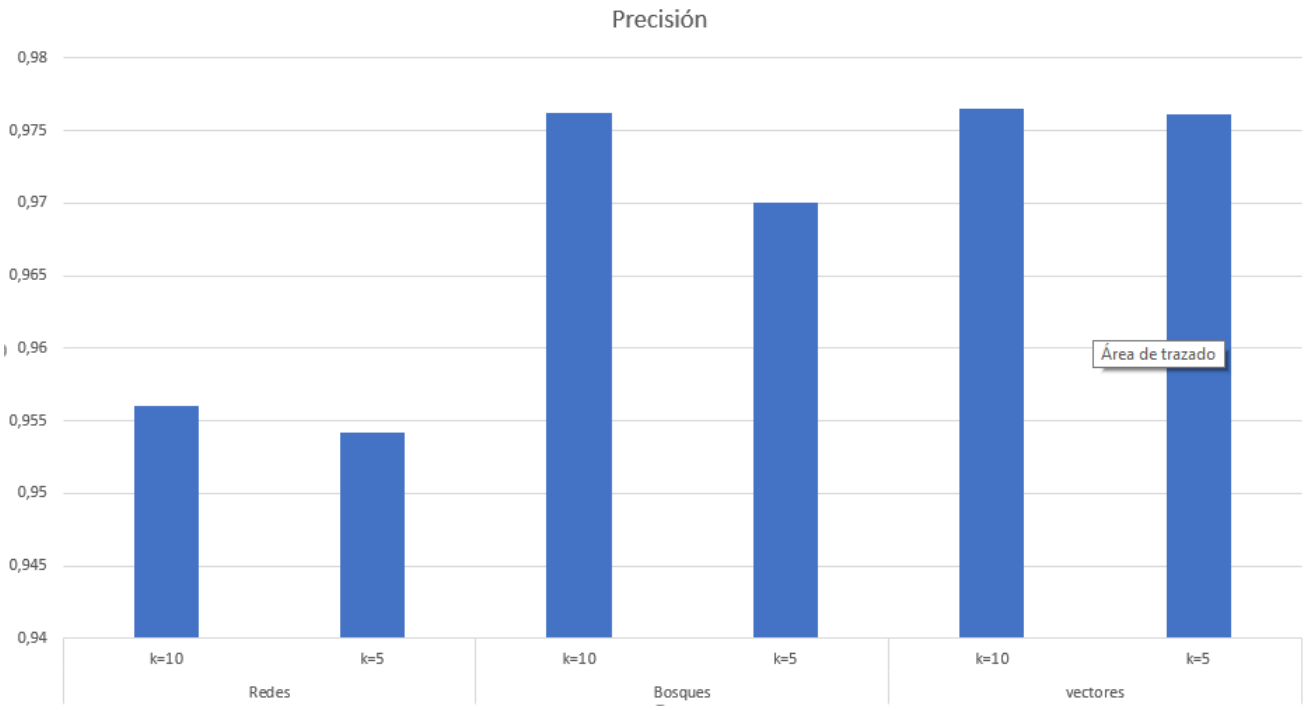
Accuracy

En términos de la métrica *Accuracy* se generaron los resultados que se presentan en la tabla 4-11. Como se observa en la figura 4-5, se tiene un mejor resultado con el modelo de máquinas de vectores de soporte ya que presenta un porcentaje de 0,971154 %, con validación cruzada de k=10.

Balanced Accuracy

En términos de la métrica *Balanced Accuracy* se generaron los resultados que se presentan en la tabla 4-12. Como se observa en la figura 4-6, se tiene un mejor resultado con el modelo de bosques aleatorios ya que presenta un porcentaje de 0,974687 %, con validación cruzada de k=10.

Figura 4-3.: Gráfico de barras de los resultados de la métrica *Precision*.

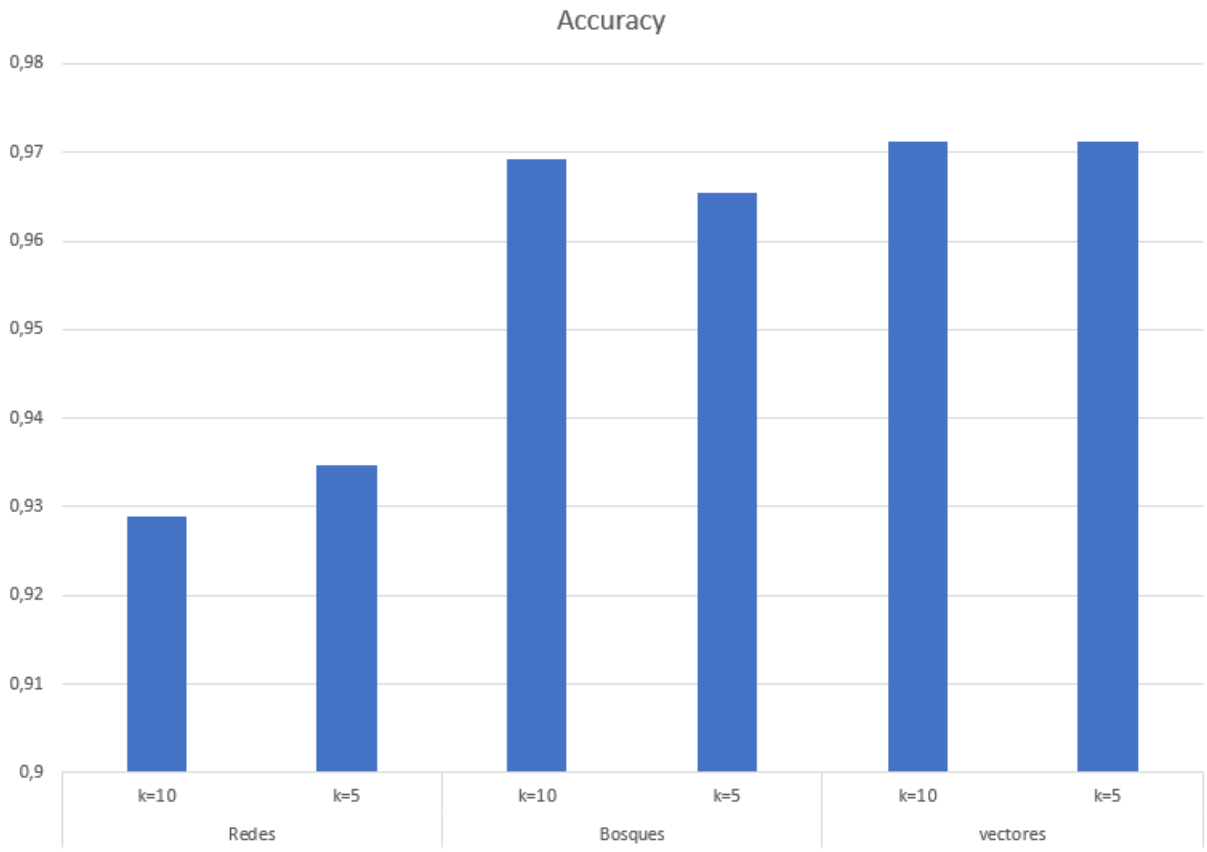


Fuente: elaboración propia

Figura 4-4.: Gráfico de barras de los resultados de la métrica *Recall*.

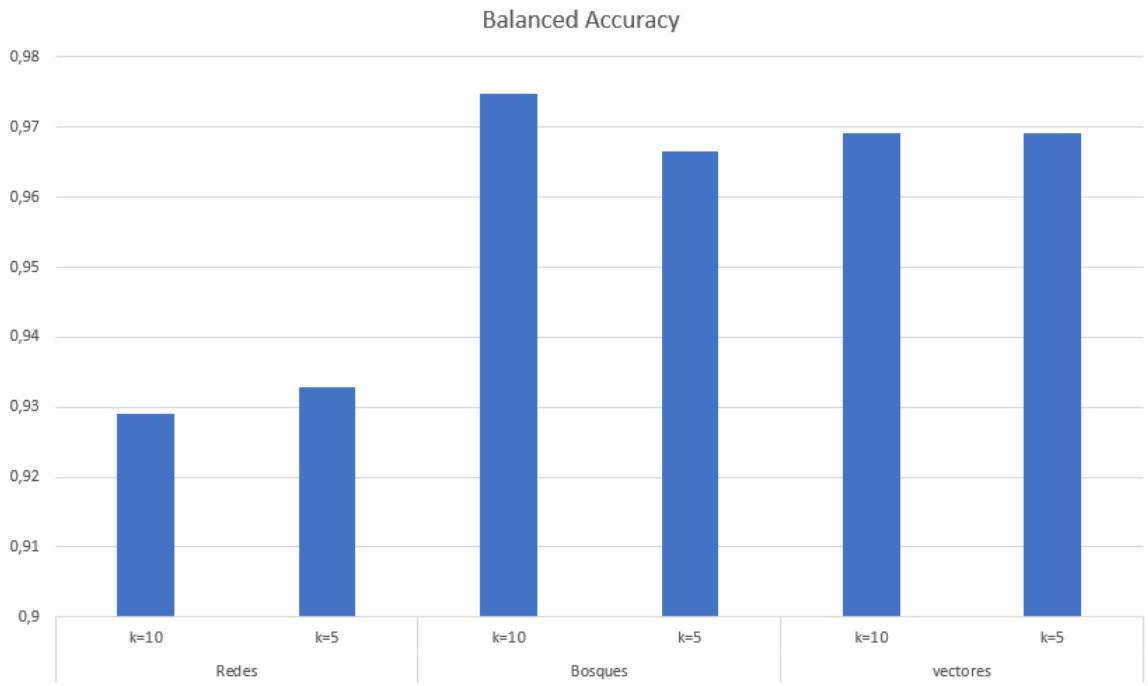
Fuente: elaboración propia

Figura 4-5.: Gráfico de barras de los resultados de la métrica *Accuracy*.



Fuente: elaboración propia

Figura 4-6.: Gráfico de barras de los resultados de la métrica *Balanced Accuracy*.



Fuente: elaboración propia

Tabla 4-11.: Resultados obtenidos en cada modelo para la métrica *Accuracy*

Modelo	Iteración	Accuracy
Redes	k=10	0.928846
	k=5	0.934615
Bosques	k=10	0.969231
	k=5	0.965385
Vectores	k=10	0.971154
	k=5	0.971154

Fuente: elaboración propia

Tabla 4-12.: Resultados obtenidos en cada modelo para la métrica *Balanced Accuracy*

Modelo	Iteración	Balanced Accuracy
Redes	k=10	0.929063
	k=5	0.932813
Bosques	k=10	0.974687
	k=5	0.966562
Vectores	k=10	0.969063
	k=5	0.969063

Fuente: elaboración propia

Avergae Precision

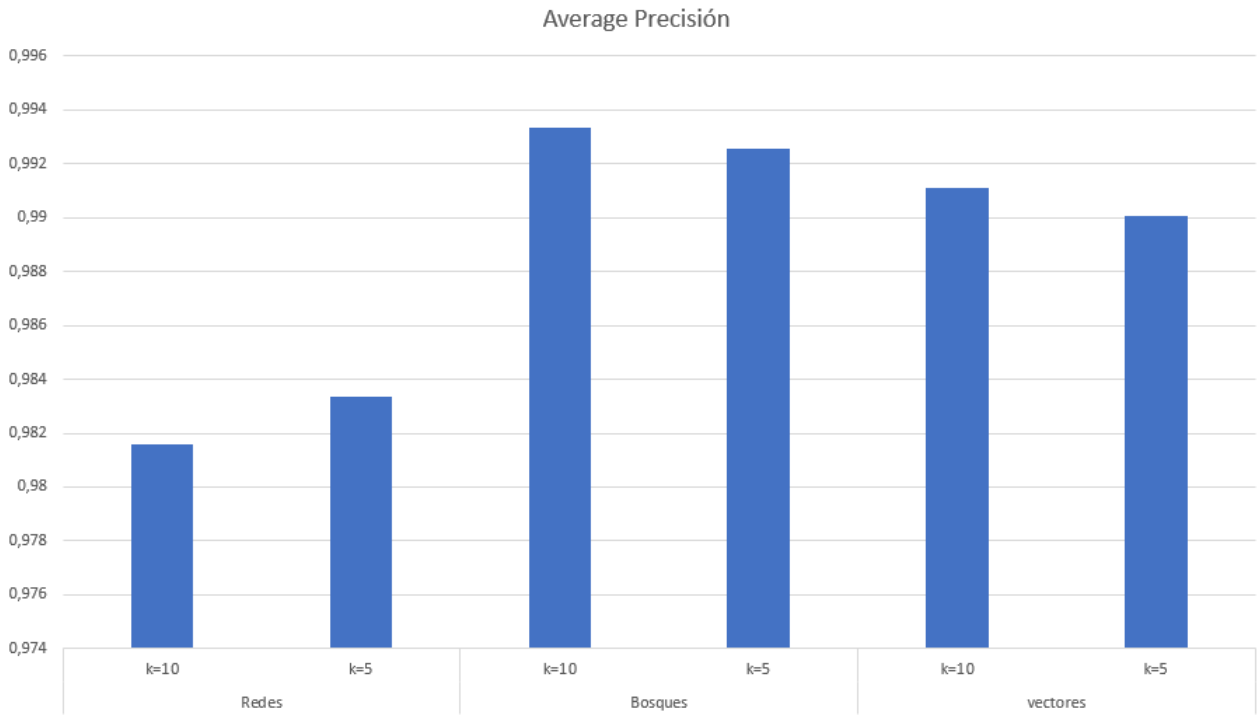
Los resultados generados para la métrica *Avergae Precision* se presentan en la tabla 4-13. Como se observa en la figura 4-7, se tiene un mejor resultado con el modelo de bosques aleatorios ya que presenta un porcentaje de 0,993344 %, con validación cruzada de k=10.

Tabla 4-13.: Resultados obtenidos en cada modelo para la métrica *Avergae Precision*.

Modelo	Iteración	Average Precisión
Redes	k=10	0.981607
	k=5	0.983348
Bosques	k=10	0.993344
	k=5	0.992572
Vectores	k=10	0.991119
	k=5	0.990068

Fuente: elaboración propia

Figura 4-7.: Gráfico de barras de los resultados de la métrica *Avergae Precision*.



Fuente: elaboración propia

Este trabajo de grado tiene como referencia la investigación realizada previamente para la predicción de diabetes en una etapa temprana usando técnicas de minería de datos. La investigación hace uso de cuatro modelos de clasificación para el análisis del conjunto de datos público, estos son bosques aleatorios, regresión logística, árbol de decisión y bayes. Este estudio utiliza métricas de rendimiento como *TP rate*, *FP rate*, *Precision*, *Recall* y *F-measure* para elegir el modelo con el mejor resultado. Se concluye que el modelo de bosques aleatorios es el mejor en la prueba de evaluación dividida porcentual.

Al comparar la investigación realizada anteriormente y el presente trabajo de grado, tienen similitud en que se usaron las métricas *Precision*, *Recall* y *F-measure* en ambas investigaciones; sin embargo, presentan diferentes resultados: para métrica *Precision* en la investigación el mejor modelo fue RF con 0,974 % y para el trabajo de grado fue SVM con 0,976 %; para métrica *Recall* en la investigación el mejor modelo fue RF con 0,974 % y para el trabajo de grado también para RF con 0,981 %; para métrica F1 en la investigación el mejor modelo fue RF con 0,974 % y para el trabajo de grado fue ANN con 0,994 %, tal como se puede observar en la tabla 4-14. En este trabajo de grado se tuvo en cuenta la curva de ROC AUC, en donde se obtuvo la mejor métrica para RF.

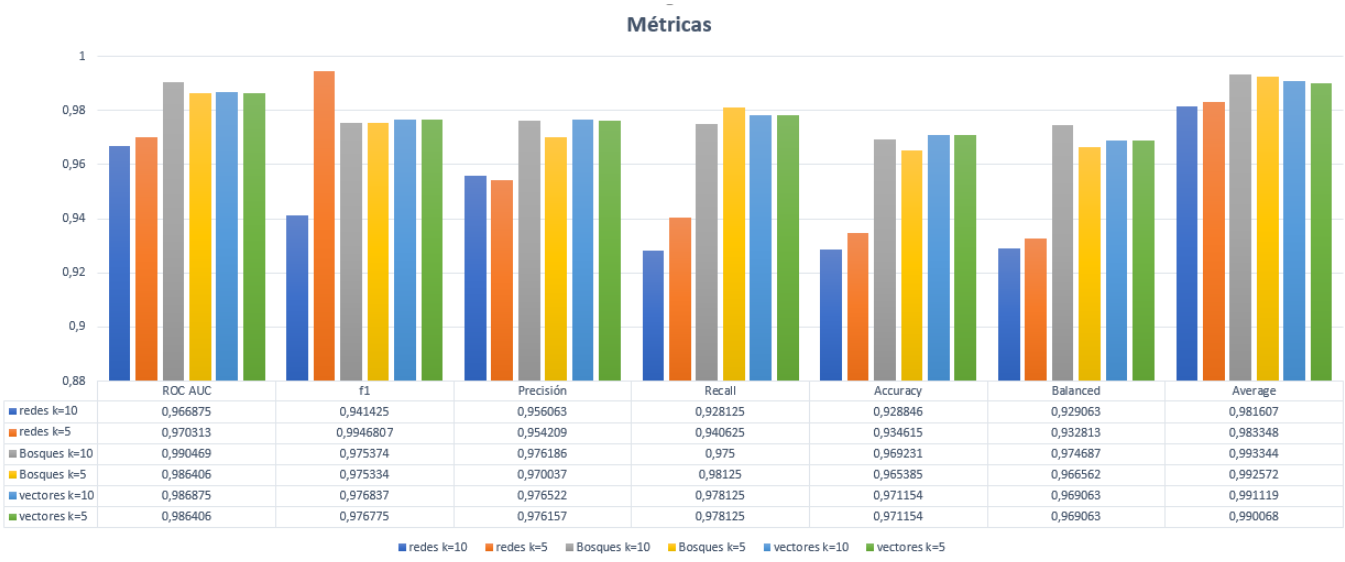
Tabla 4-14.: Comparación de los resultados de la investigación realizada anteriormente con los resultados del presente trabajo de grado.

Métrica	Investigación	Trabajo de grado
Precision	RF 0,974 %	SVM 0,976 %
Recall	RF 0,974 %	RF 0,981 %
F-measure	RF 0,974 %	ANN 0,994 %
ROC AUC	No registra	RF 0,990 %

Fuente: elaboración propia

De forma general, se puede concluir que con las técnicas de aprendizaje y métodos de preprocesamiento utilizados, el modelo con el cual se obtiene el mejor resultado es RF y SVM como se observa en la figura 4-8; en donde: para ROC AUC el mejor modelo es RF con precisión de 0,990 % y SVM con 0,986 %; para la F1 el mejor modelo es ANN con precisión de 0,994 % y SVM con 0,976 %; para la *Precision* el mejor modelo es SVM con precisión de 0,9765 % y RF con 0,9761 %; para la *Recall* el mejor modelo es RF con precisión de 0,9815 % y SVM con 0,978 %; para la *Accuracy* el mejor modelo es SVM con precisión de 0,971 % y RF con 0,969 %; para la *Balanced Accuracy* el mejor modelo es RF con precisión de 0,974 % y SVM con 0,969 %; finalmente para la métrica *Average Precision* el mejor modelo es RF con precisión de 0,993 % y SVM con 0,991 %.

Figura 4-8.: Gráfico de barras de los resultados de las métricas.



	ROC AUC	f1	Precisión	Recall	Accuracy	Balanced	Average
redes k=10	0,966875	0,941425	0,956063	0,928125	0,928846	0,929063	0,981607
redes k=5	0,970313	0,9946807	0,954209	0,940625	0,934615	0,932813	0,983348
Bosques k=10	0,990469	0,975374	0,976186	0,975	0,969231	0,974687	0,993344
Bosques k=5	0,986406	0,975334	0,970037	0,98125	0,965385	0,966562	0,992572
vectores k=10	0,986875	0,976837	0,976522	0,978125	0,971154	0,969063	0,991119
vectores k=5	0,986406	0,976775	0,976157	0,978125	0,971154	0,969063	0,990068

Fuente: elaboración propia

5. Conclusiones y recomendaciones

5.1. Conclusiones

La diabetes es una de las enfermedades con alto grado de mortalidad en cualquier etapa de la vida, no solo en Colombia sino en el mundo. De igual forma, se tienen tratamientos que pueden ser indispensables para manejar adecuadamente y evitar la enfermedad como: control de glucosa, dieta saludable, actividad física, tomar agua, no consumo de tabaco y alcohol, que si ya diagnosticado el paciente puede ayudarle a llevar un mejor control. Sin embargo, es pieza clave la detección temprana de la enfermedad para el tratamiento y poder evitar complicaciones graves a las que conlleva la diabetes. Para ello, en el presente trabajo de grado, se implementaron modelos de aprendizaje como: máquinas de vectores de soporte, bosques aleatorios y redes neuronales, para predecir el diagnóstico de la diabetes. Se utilizó el conjunto de datos público, recopilado por el hospital de Sylhet, Bangladesh y usando varias métricas de desempeño incluidas CURVA ROC y F1. Dando cumplimiento al objetivo del trabajo de grado, se identifica al modelo bosques aleatorios como el modelo que muestra el mejor resultado para predecir el diagnóstico de la enfermedad.

Hoy en día el uso de la tecnología es de gran influencia por sus grandes contribuciones en cada uno de los sectores del mundo, el poder ayudar a la humanidad a prever enfermedades como la diabetes y dar un diagnóstico temprano de la enfermedad es de gran satisfacción personal, y con esto, posicionar la carrera de ingeniería de sistemas, con la ayuda de la inteligencia artificial, como una carrera que contribuye a la sociedad, no solo en adelantos tecnológicos, sino también, con encontrar algoritmos que usando un conjunto de datos puedan ayudar en la ciencias médicas prediciendo el riesgo de padecer enfermedades. Entonces, utilizando aprendizaje automático, en el presente trabajo de grado se emplearon los modelos mencionados anteriormente, para predecir el diagnóstico de diabetes a partir de perfiles clínicos de pacientes; se descubrió que: para las métricas curva ROC AUC, *Recall, Balanced Accuracy y Average Precision* el modelo bosques aleatorios presenta la mejor precisión, para F1 la mejor precisión es de redes neuronales, para *precision y Accuracy* es máquinas de vectores de soporte. De forma general, el mejor resultado es para Bosques aleatorios con 0.99 % de precisión.

Fundamentalmente, la ventaja que ayudó para obtener un mejor resultado que en la investigación realizada anteriormente, fue realizar un preprocesamiento diferente, el cambio de variables y posteriormente la reducción de dimensionalidad. No obstante, como se ha venido mencionando, queda abierta la posibilidad de emplear otras técnicas de aprendizaje u otros

métodos de preprocesamiento o aplicando otras tecnologías como minería de datos, se puede llegar a mejorar los resultados obtenidos en el este trabajo de grado. Adicional, se tenía la limitante de un dataset público, con un conjunto de datos recopilado en un país específico, se puede llegar a un análisis más exhaustivo con referente a la presencia de la enfermedad en dicho país.
Finalmente, con este trabajo de grado, se ha propuesto una herramienta adicional para la predicción del diagnóstico de la diabetes utilizando aprendizaje automático. Es de aclarar que el conocimiento de un médico es fundamental y el presente trabajo de grado puede ser de gran ayuda para ellos, por los grandes avances tecnológicos que han existido con el paso del tiempo.

5.2. Recomendaciones

El lenguaje utilizado para realizar el trabajo de grado fue python porque tienen bibliotecas que hacen más sencillo de los algoritmos y el procesamiento de los datos; lo aconsejable es documentarse, indagar, aprender sobre el lenguaje para que a la hora de desarrollar se tenga mayor fluidez, investigar los parámetros que se usan para cada modelo, algo indispensable es conocer los datos o el conjunto de datos para poder elegir una técnica y/o preprocesamiento acorde al resultado que se espera; finalmente, no tener miedo a errar, todos podemos equivocarnos y con ello se acumula conocimiento y experiencia.
Es importante seguir trabajando en el tema, en proyectos que ayuden a las personas a mejorar su salud, su vida, su entorno, ayudar a la comunidad en general. Con esto inculcar a los estudiantes la investigación con el fin de ayudar y además lograr posicionar a la Universidad como una de las mejores o la mejor en temas de investigación del país.

5.3. Trabajo futuro

Como se menciona en el contenido del presente trabajo de grado, hay posibilidad de emplear otras técnicas de aprendizaje u otros métodos de preprocesamiento, aplicar otras tecnologías como minería de datos, usando otro conjunto de datos que contenga más o mejores variables, o por ejemplo, un conjunto de datos que fuese recopilado en el país, para obtener diferentes y porque no, mejores resultados que los obtenidos en el presente trabajo de grado o en investigaciones previas; lo importante es no cambiar la meta, ayudar a las personas por medio de la tecnología.

A. Apéndice

A.1. Manual de usuario

.

Bibliografía

[1] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," Adv. Intell. Syst. Comput., vol. 992, pp. 113-125, 2020, doi: 10.1007/978-981-13-8798-2-12.

[2] B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on Machine Learning," IEEE J. Biomed. Heal. Informatics, vol. 20, no. 1, pp. 39–46, Jan. 2016, doi: 10.1109/JBHI.2015.2396520.

[3] B. J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on Machine Learning," IEEE J. Biomed. Heal. Informatics, vol. 20, no. 1, pp. 39–46, Jan. 2016, doi: 10.1109/JBHI.2015.2396520

[4] SÁNCHEZ RIVERO, Germán. Historia de la diabetes. Gac Med Bol [online]. 2007, vol.30, n.2 [citado 2021-02-15], pp. 74-78 . Disponible en: http://www.scielo.org.bo/scielo.php?script=sci- arttext y pid=S1012-29662007000200016&lng=es&nrm=iso¿. ISSN 1012-2966.

[5] Villalobos A, Rojas-Martínez R, Aguilar-Salinas CA, et al. Atención médica y acciones de autocuidado en personas que viven con diabetes, según nivel socioeconómico. salud pública mex. 2019;61(6):876-887.

[6] Gómez-Encino, Guadalupe del Carmen, Cruz-León, Aralucy, Zapata-Vázquez, Rosario, Morales- Ramón, Fabiola Nivel de conocimiento que tienen los pacientes con Diabetes Mellitus tipo 2 en relación a su enfermedad. Salud en Tabasco [en linea]. 2015, 21(1), 17-25[fecha de Consulta 15 de Febrero de 2021]. ISSN: 1405-2091. Disponible en: https://www.redalyc.org/articulo.oa?id=48742127004

[7] D. I. Conget, “Diagnosis, classification and pathogenesis of diabetes mellitus,” Rev. Esp. Cardiol., vol. 55, no. 5, pp. 528–535, Jan. 2002, doi: 10.1016/S0300-8932(02)76646-3.

[8] AMERICAN DIABETES ASSOCIATION, “Diagnosis and Classification of Diabetes Mellitus,” 2005.

[9] World Health Organization, “OMS — Diabetes,” 2020. https://www.who.int/diabetes/action-online/basics/es/index3.html (accessed Sep. 06, 2020).

[10] A. D. Association, “Classification and diagnosis of diabetes,” Diabetes Care, vol. 40, no. Supplement 1, pp. S11–S24, Jan. 2017, doi: 10.2337/dc17-S005.

[11] D. Gan et al., “Diabetes Atlas Second Edition,” 2003. Accessed: Sep. 06, 2020. [Online]. Available: www.idf.org.

[12] María Estela Raffino, “Dato - Qué es, concepto, ejemplos y tipos de datos,” Aug. 20, 2020. https://concepto.de/dato/ (accessed Sep. 13, 2020).

[13] Ethem Alpaydin, Introduction to Machine Learning - Ethem Alpaydin - Google Libros, Fourth edi. 2020.

[14] ”Diferencia entre clasificación y regresión en machine learning – Ingenierobeta.com.” https://ingenierobeta.com/clasificacion-vs-regresion-machine-learning/ (accessed Sep. 07, 2020).

[15] “Transparencia Habeas Data.” https://www.ins.gov.co/Transparencia/habeasdata (accessed Sep. 07, 2020).

[16] Informe de actor interesado, “El derecho a la intimidad en Colombia,” Exam. Periódico Univers. 30O período Ses. - Colomb., Oct. 2017, doi: 10.2139/ssrn.1951416.

[17] L. Fernando Espinel, “Fundación Tecnológica Liderazgo Canadiense Internacional LCI Reglamento de propiedad intelectual. Octubre 27 de 2.011.”

[18] Eliana Aldana, “Predicción de riesgo de diabetes tipo 2 en adultos jóvenes en presencia de componente heredo-familiar - Revista Electrónica de PortalesMedicos.com,” May 30, 2011.

[19] F. Soriguer et al., “Validación del FINDRISC (FINnish Diabetes Risk SCore) para la predicción del riesgo de diabetes tipo 2 en una población del sur de España. Estudio Pizarra,” Med. Clin. (Barc)., vol. 138, no. 9, pp. 371–376, Apr. 2012, doi: 10.1016/j.medcli.2011.05.025.

[20] C. Pérez Gandía, “Propuesta de algoritmos de predicción de glucosa en pacientes diabéticos,” oct. 2014.

[21] O. D. Castrillón, W. Sarache, and E. Castaño, “Sistema bayesiano para la predicción de la diabetes,” Inf. Tecnol., vol. 28, no. 6, pp. 161–168, 2017, doi: 10.4067/S0718-07642017000600017.

[22] “Página de la organización panamericana de la salud” https://www.paho.org/es/temas/diabetes y https://www.paho.org/hq/dmdocuments/2009/Perfil-ESP.pdf

[23] “Página del ministerio de salud Colombia” https://www.minsalud.gov.co/Paginas/Tres-de-cada-100-colombianos-tienen-diabetes.aspx

[24] Diagnosis and management of type 2 diabetes (HEARTS-D). Geneva]: World Health Organization; 2020 (WHO/UCN/NCD/20.1). Licence: CC BY-NC-SA 3.0 IGO

[25] ARNOLD RODRÍGUEZ, Mónica et al. Pesquisaje y prevención de la diabetes mellitus tipo 2 en población de riesgo. Rev Cubana Hig Epidemiol [online]. 2012, vol.50, n.3 [citado 2021-02-21], pp.380-391. Disponible en: ¡http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1561-30032012000300012&lng=es&nrm=iso¿. ISSN 1561-3003.

[26] ”Neuronas”. Autor: Julia Máxima Uriarte. Para: Caracteristicas.co. Última edición: 17 de octubre de 2019. Disponible en: https://www.caracteristicas.co/neuronas/.

[27] Ïntroducción a las redes neuronales con Scikit-Learn". Autor: Scott Robinson. Disponible en: https://stackabuse.com/introduction-to-neural-networks-with-scikit-learn/.

[28] "Machine learning". Autor: Daniel Burrueco. Disponible en: https://www.interactivechaos.com/es/manual/tutorial-de-machine-learning.

[29] Piloto-Rodríguez, Ramón. (2017). Redes Neuronales Artificiales. Conceptos básicos y algunas aplicaciones en Energía. 10.13140/RG.2.2.23326.54083.

[30] Lasse Petteri Rouhiainen. (2018).Inteligencia artificial. 101 cosas que debes saber hoy sobre nuestro futuro, pp. 19-21, 2018, ISBN: 978-84-17568-08-5.

[31] "bosques aleatorios". Autor: Hebert Yuri Puma. Disponible en: https://medium.com/@hpumah/bosques-aleatorios-482163ace92e

[32] "Documentación de python". Autor: Página de python. Disponible en: https://docs.python.org/es/3/tutorial/index.html

[33] Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B y Varoquaux G (2014) Machine learning for neuroimging with scikit-learn. Parte delantera. Neuroinform . 8 : 14. doi: 10.3389 / fninf.2014.00014

[34] Hao, Jiangang, and Tin Kam Ho. Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. Vol. 44. Los Angeles, CA: SAGE Publications, 2019. Web.

[35] Borja, Ricardo & MonleonGetino, Antonio & Benedé, Jose. (2020). Estandarización de Métricas de Rendimiento para Clasificadores Machine y Deep Learning.

[36] "Solo para entendidos". Autor: AVB. Disponible en: http://www.soloentendidos.com/regresion-lineal-machine-learning-python-2207

[37] "Variables dummy". Autor: Ravindu. Disponible en: https://datasciencelk.com/dummy-variables-in-regression/

[38] Nagesh, Chauhan Sep 02, 2020 https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

[39] "Selección métricas de clasificación". Autor: sitio big data. Disponible en: https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/

[40] .ªnálisis de componentes principales". Autor: Rukshan Pramoditha. Disponible en: https://towardsdatascience.com/principal-component-analysis-pca-with-scikit-learn-1e84a0c731b0

[41] Di Masso, Mauro. Granitto, Pablo. Selección estable de variables independientes con RFE. 2000, ISSN: 1850-2784 pp. 27-28 . Disponible en: https://43jaiio.sadio.org.ar/proceedings/ASAI/4.pdf.

[42] "Documentación de Scikit-Learn". Autor: Página de scikit-learn. Disponible en: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[43] .ªrboles de decisión". Autor: Rahul Saxena. Disponible en: https://dataaspirant.com/decision-tree-algorithm-python-with-scikit-learn/

[44] ïnteligencia computacional". Autor: Maricela Bravo. Disponible en: http://aisii.azc.uam.mx/mcbc/Cursos/IntCompt/Sesion %204. %20Redes %20Neuronales.pdf

[45] Çodificaciones ordinales y One-Hot para datos categóricos". Autor: Jason Brownlee. Disponible en: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/ https://www.paho.org/hq/dmdocuments/2009/Perfil-ESP.pdf