UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO
FACULTAD DE ECONOMÍA

"COMPARACIÓN ENTRE REGRESIÓN LOGÍSTICA Y PERCEPTRÓN MULTICAPA: CASO APLICADO AL CONJUNTO DE DATOS PIMA INDIAN DIABETES"

TESIS

QUE PARA OBTENER EL TÍTULO DE:

LICENCIADO EN ACTUARÍA

PRESENTA:

BENJAMÍN DE JESÚS QUINTANA CHIMAL

ASESOR:

DR. DANIEL LOZANO KEYMOLEN

REVISORES:

DRA. LIDIA ELENA CARVAJAL GUTIÉRREZ

DR. SERGIO CUAUHTÉMOC GAXIOLA ROBLES LINARES

ÍNDICE

INTRODUCCIÓN

La diabetes mellitus es uno de los trastornos documentados más antiguos de la humanidad y ha desafiado a las civilizaciones durante siglos. Areteo de Capadocia (81-138 d.C.), discípulo de Hipócrates, fue el primero en proponer el término diabetes que significa "correr a través de" o "sifón" para el trastorno de la excesiva ingestión de líquidos y la producción de orina. La diabetes mellitus se define como una enfermedad crónica irreversible, en la que existe presencia irregular de azúcar o de glucosa en la sangre debido a un mal funcionamiento en el páncreas.

El farmacéutico francés Apollinaire Bouchardat fue el primero en proponer una relación entre la diabetes mellitus y el páncreas en 1866, la cual fue confirmada por Oskar Minkowski y Joseph von Mering, de la Universidad de Estrasburgo, Francia. En 1889 mediante la serendipia, durante sus estudios sobre los mecanismos que intervienen en la absorción de las grasas intestinales, extirparon el páncreas de un perro y descubrieron involuntariamente que esto provocaba la poliuria y la glucosuria características de la diabetes mellitus. Posteriormente el patólogo francés Gustave Edouard Laguesse postuló que los islotes de Langerhans producían una nueva hormona que desempeñaba un papel regulador de la digestión, el cual llamó insulina.

Los trabajos de Frederick Grant Banting y Charles Herbert Best confirmaron la existencia de la insulina, lo que condujo a la disponibilidad de un tratamiento eficaz de la diabetes mellitus (inyecciones de insulina). Para 1922 se trató con éxito al primer paciente que la padecía. Otro hito fue establecido en 1960 por la física médica americana Rosalyn Yalow, co-ganadora del Premio Nobel de 1977 y co-desarrolladora del radioinmuno ensayo.

La investigación de la Dra. Yalow condujo al concepto de resistencia a la insulina en pacientes con diabetes mellitus de tipo 2: aunque estos individuos pueden 4producir suficiente insulina, sus tejidos diana (en particular el tejido adiposo, el músculo y el hígado) suelen responder de forma inadecuada a la hormona, por lo que se considera la primera distinción entre los tipos de diabetes, dicha distinción abrió la puerta a

nuevas y más eficaces formas de tratamiento de la diabetes mellitus que progresó de acuerdo con la creciente comprensión en su fisiopatología[1].

Los criterios más recientes para definir la diabetes[2] son los propuestos por la *Asociación Americana de la Diabetes*, que se basan en la concentración de glucosa en la sangre, según la prueba A1C[3]. Entre las principales causas de morbilidad y mortalidad son la obesidad y el aumento de peso y ambos están asociados con un mayor riesgo de diabetes entre las personas. Dentro de las medidas que se pueden adoptar para disminuir el riesgo de padecer diabetes todas implican ejercicio constante y buenos hábitos de alimentación, llegando a controlar así el sobrepeso, los niveles altos de colesterol, la inactividad física y la presión arterial. Cabe mencionar que, así como hay modificaciones en la rutina de las personas con el fin de disminuir el riesgo de padecer diabetes, hay situaciones que no se pueden modificar, tales como el tener antecedentes de diabetes familiar, la raza o etnia a la que se pertenezca y la edad principalmente.

Una vez que se padece diabetes, hay consecuencias en la salud, las cuales se pueden agrupar en dos categorías: complicaciones microvasculares y macrovasculares. Las complicaciones microvasculares son la retinopatía[4], nefropatía[5] y la neuropatía[6], mientras que las complicaciones macrovasculares son la cardiopatía coronaria, derrame cerebral y aterosclerosis. La diabetes también tiene impacto económico, los

---

1 Parte de la biología que estudia el funcionamiento de un organismo o de un tejido durante el curso de una enfermedad.

2 En adelante solo se utilizará el término diabetes dado que la ADA propuso desde el año 2019 que solo se nombre como tal a la enfermedad debido a que se diferencia esta enfermedad por los tipos de diabetes.

3 La prueba denominada A1c (HbA1c) mide la cantidad de azúcar en la sangre (glucosa) adherida a hemoglobina. La hemoglobina es la parte de los glóbulos rojos que transporta el oxígeno de los pulmones al resto del cuerpo.

4 La retinopatía ocurre cuando de forma mantenida en el tiempo existe hiperglucemia, por tanto, las pequeñas arterias de la retina sufren y se lesionan.

5 La nefropatía se define como la patología de personas diabéticas en las que partes del riñón llamadas nefronas lentamente se engruesan y cicatrizan. Esto provoca que se empiece a filtrar patológicamente proteínas como la albúmina y aparezcan en la orina.

6 La neuropatía es una complicación cuando no se controla bien la diabetes, en la cual se daña parte del sistema nervioso.

costos directos de la atención de la salud de las personas con diabetes son generalmente 2-3 veces más que para los que no tienen diabetes y 4-8 veces más si hay complicaciones debido a la diabetes. Los costos indirectos de la diabetes se vuelven cada vez más importantes en los países de bajos ingresos cuando los costos como el costo de los viajes a clínicas y la pérdida de ingresos tienen un mayor impacto en el individuo y la familia.

PLANTEAMIENTO DEL PROBLEMA

La diabetes impone una gran carga económica al sistema mundial de atención de la salud pública y a la economía mundial en general, a nivel mundial, se estima que 422 millones de adultos vivían con diabetes y la población adulta estimada era de 4.97 miles de millones en 2014, en comparación con 108 millones de diabeticos y los 2.3 miles de millones de adultos en 1980 (Organizacion Mundial de la Salud, 2016) . La prevalencia mundial de la diabetes casi se ha duplicado desde 1980, pasando del 4.7% al 8.5% en la población adulta. Esto refleja un aumento de los factores de riesgo asociados como el sobrepeso u obeso. La diabetes causó 1.5 millones de muertes en 2012. Un nivel de glucosa en la sangre superior a la óptima causó un aumento de 2.2 millones de muertes, al aumentar los riesgos de enfermedades cardiovasculares y de otro tipo. El 43% de estos 3.7 millones de muertes ocurren antes de los 70 años (Organizacion Mundial de la Salud, 2016).

En México, las estimaciones existentes son muy variables con cálculos de costos de atención por paciente que padece diabetes, que van desde 700 hasta 3,200 dólares anuales, lo que se traduce en 5 a 14 por ciento del gasto en salud destinado a la atención de esta enfermedad y sus complicaciones, inversión que de acuerdo con la Federación Internacional de Diabetes se relaciona directamente con la tasa de mortalidad por esta causa.

Ya que se tiene el contexto de la diabetes y se dimensiona el problema de salud pública que representa, son notorios los intentos por predecir o pronosticar la enfermedad como una forma de aminorar las complicaciones en salud y económicas

de la misma. A propósito de lo anterior, esta tesis tiene como objetivo comparar dos algoritmos de clasificación, dado que la posible diferencia de precisión entre ambos, se traduce en diagnósticos acertados, es decir, una clasificación superior. Se elige el perceptrón dado que es un algoritmo en apogeo y que aún se investiga al respecto y la regresión logística, dado que es un algoritmo generalmente usado para la clasificación en distintos ámbitos, como por ejemplo identificar ciudadanos que han consumido de drogas en el estado de Kwara, en Nigeria (Balogun et al., 2013); determinar el sector asegurador al que pertenece una aseguradora en Turquía (Ruzgar y Ruzgar, 2007); clasificar enfermedades hepáticas (Abdulqader, 2017) entre otros, es por eso que se eligen estos algoritmos, por lo que se procede a resolver el problema de clasificación empleando la base de datos Pima Indian Diabetes, que pertenece al National Institute of Diabetes and Digestive and Kidney Diseases de los Estados Unidos de América. Las unidades de análisis consisten en 768 mujeres residentes cerca de Phoenix, Arizona, EEUU, pertenecientes a la etnia Pima y con al menos 21 años de edad. La finalidad de resolver el problema de clasificación con dos modelos diferentes es el comparar las propiedades clasificatorias de cada algoritmo y determinarlo siguiente: ¿cuál resuelve el problema con mayor precisión?

HIPÓTESIS DE INVESTIGACIÓN

Previo a la comparación, se plantea que el perceptrón será superior a la regresión logística en las propiedades clasificatorias, así como también será un modelo más robusto (menor varianza en las estimaciones) y estable.

El primer algoritmo es la regresión logística, donde la variable dependiente o variable de respuesta es el diagnóstico que se tiene en la base de datos el cual dictamina si la persona padece diabetes o no, los demás factores de la base de datos sólo son de interés por su capacidad de ayudar a explicar la variable de respuesta. Los modelos de regresión logística suelen ajustarse por máxima verosimilitud, usando la probabilidad condicional de Y dada X.

El segundo algoritmo es el perceptrón multicapa, que es un conjunto de *k* perceptrones independientes que se combinan para construir una única red neuronal artificial con *k* salidas. Este algoritmo se denomina multicapa dado que precisamente tiene varias capas, la primera es la capa de entrada, la cual está conformada por las variables independientes o explicativas del conjunto de datos. La siguiente capa se denomina capa oculta, la cual interconecta todos los nodos de entrada. Finalmente, la capa de salida es el resultado de la suma ponderada de los pesos de los nodos multiplicado por las variables independientes, la cual en este caso es la predicción del padecimiento de diabetes.

Se seleccionan como variables independientes únicamente las continuas, ya que la investigación se hace bajo el enfoque de la escuela francesa del análisis de datos. Es necesario hacer tratamientos a las variables debido a que tienen diferentes escalas de valores y valores faltantes. El tener distintas escalas representa una desventaja dado que los pesos que asignan a las variables podrían estar sesgados. Los valores faltantes son imputados con la mediana para evitar sesgo en los pesos tanto de la regresión logística como del perceptrón. Una situación que se presenta mucho en los problemas de clasificación es contar con clases desbalanceadas, y el conjunto de datos Pima no es una excepción, por lo que se debe proceder con técnicas estadísticas de balanceo, todo esto con el objetivo de que los algoritmos se entrenen con el mismo número de casos de ambas clases y así los pesos no estén sesgados hacia una clase en particular. Una vez imputados los valores faltantes y balanceadas las clases, se procede a estandarizar las variables con la finalidad de que todas las variables tengan la misma escala.

Ya con los datos preparados, se procede a alimentar los modelos y comparar los resultados. La precisión de los modelos se evalúa con la validación cruzada, un proceso aleatorio que divide el conjunto de datos en *k* partes iguales y las selecciona aleatoriamente de acuerdo a la proporción entrenamiento-prueba, después de *m* iteraciones se calcula la precisión promedio de cada algoritmo, sin embargo, la precisión no es suficiente para determinar qué modelo resuelve mejor el problema,

también se deben contemplar los criterios de cobertura, robustez, estabilidad, interpretabilidad, que permiten determinar si efectivamente las precisiones del modelo son sólidas o si fueron cuestión del azar. También se emplea la curva del Receiver Operating Characteristic (ROC) la cual profundiza más en la clasificación de los casos positivos, los criterios de sensibilidad y especificidad explican precisamente cómo el algoritmo clasifica los casos de personas que en efecto padecen o no diabetes.

Las métricas con las que se evalúan los modelos giran en torno a la precisión, esto debido a que es de suma importancia el poder clasificar de manera adecuada por las implicaciones en salud y económicas que se abordaron previamente. De la precisión depende el diagnóstico oportuno o incluso la prevención de una enfermedad, evitando así las complicaciones fisiológicas, económicas y sociales que la diabetes implica. Sin embargo, el estudio presentado generaliza un poco más, pues se tocan puntos que son importantes desde la perspectiva de la estadística, dado que el presente trabajo pretende comparar los dos algoritmos de clasificación exponiendo ventajas y desventajas de ambos, así como brindar una “introducción” al perceptrón debido a que en el mercado laboral actual se demanda el dominio de herramientas “novedosas” y sofisticadas como el perceptrón.

Se eligió el perceptrón debido a que se presume superior a los algoritmos tradicionales, sin embargo, se exponen los rubros en los cuales la regresión logística puede tener una ventaja respecto del perceptrón, dado que en problemas más generales donde no se depende tanto de la precisión puede ser mejor implementar una regresión logística, por cuestiones de infraestructura, tiempo de cómputo, tiempo de entrenamiento del modelo entre otras. En general el problema de clasificación se presenta en todas las áreas que abarca la actuaría, el aprendizaje automático y la minería de datos cobran cada vez más importancia y es imprescindible comenzar a estudiar y dominar éstas áreas con el fin de ser competitivos y poder ser considerados por las empresas que demandan profesionales capacitados, ya que las áreas de oportunidad son más exigentes día con día.

Considerando el objetivo de esta tesis, el contenido de la misma se expone en 5 capítulos adicionales a esta Introducción. En el capítulo 1 se abordan las causas y consecuencias de la diabetes, la historia de cómo se fue estudiando esta enfermedad a lo largo del tiempo, la clasificación de ésta y las repercusiones en el cuerpo humano de quienes la padecen. En el capítulo 2 se habla del impacto económico y social que tiene la diabetes en el mundo, comparando países de ingresos altos, de ingresos bajos, y el caso particular de México. En el capítulo 3 se exponen la historia y evolución de los algoritmos de clasificación que se comparan, las aplicaciones en distintos ámbitos, el concepto y desarrollo del aprendizaje automático y las técnicas novedosas que de éste se desprenden. En el capítulo 4 se aplica de las metodologías expuestas en el capítulo 3, se discuten los resultados a los que se llega y se implementa una rutina en el lenguaje Python (von Rossum, 1995) con el propósito de que sea reproducible en un futuro. Finalmente, en las conclusiones se analiza y compara el perceptrón multicapa y la regresión logística.

CAPÍTULO 1. DIABETES MELLITUS: CAUSAS Y CONSECUENCIAS

La diabetes mellitus es uno de los trastornos documentados más antiguos de la humanidad y ha desafiado a las civilizaciones durante siglos. Es generalmente aceptado que el Papiro de Ebers, escrito en 1552 A.C. por el médico egipcio Hesy-Ra de la Tercera Dinastía[7] contiene la más antigua descripción de síntomas similares a los de la diabetes, incluyendo la excesiva ingesta de líquidos y la producción de orina. Este documento fue descubierto en 1862 por el egiptólogo alemán Georg Ebers, y también contiene una lista de los remedios contra la enfermedad del exceso de orina[8] (Toelsie et al., 2013).

Areteo de Capadocia (81-138 d.C.), discípulo de Hipócrates, fue el primero en proponer el término diabetes que significa "correr a través de" o "sifón" para el trastorno de la excesiva ingestión de líquidos y la producción de orina. Aproximadamente 250 años después, los antiguos médicos hindúes Charaka, Sushruta y Vagbhata acuñaron el término "orina de miel" en los textos ayurvédicos de la India para caracterizar la orina pegajosa que sabía a miel y atraía a las hormigas y las moscas. Estos eruditos también describieron la aparición temprana y tardía de este trastorno, así como su relación con factores hereditarios, la obesidad, un estilo de vida sedentario y ciertos hábitos alimentarios (Toelsie *et al.,* 2013).

1.1 Definición de diabetes

El siglo XVIII se considera generalmente como la "Edad de Oro" de la medicina. Muchas enfermedades fueron catalogadas y descritas, e importantes avances se hicieron, particularmente en el campo de la medicina interna. Estos desarrollos también llevaron a una mejora notable en la comprensión de la fisiopatología de la diabetes, que se define como una enfermedad crónica irreversible, en la que existe

[7] La tercera dinastía del Antiguo Egipto comienza en 2700 a. C. con el reinado de Sanajt y termina en 2630 a. C., después de Huny. Es la primera de las cuatro dinastías que constituyen el denominado Imperio Antiguo de Egipto.

[8] Se denomina poliuria a la enfermedad que produce orinar en exceso y que tiene varias causas, como la diabetes.

presencia irregular de azúcar o de glucosa en la sangre debido a un mal funcionamiento en el páncreas. En 1774, el médico inglés Matthew Dobson demostró la presencia de "materiales de sacarina" en la orina de los pacientes con diabetes al evaporar dos cuartos de orina de dicho individuo y obtener un residuo granulado que olía y sabía a azúcar. Veinticuatro años después, John Rollo, un médico escocés, introdujo el adjetivo "mellitus" (que significa "dulce" en latín) al término diabetes, para distinguir este trastorno de otras condiciones poliúricas (Toelsie *et al.,* 2013).

Evaluando los datos de los estudios de autopsias, el farmacéutico francés Apollinaire Bouchardat fue el primero en proponer una relación entre la diabetes mellitus y el páncreas en 1866. Oskar Minkowski y Joseph von Mering, de la Universidad de Estrasburgo (Francia), confirmaron esta proposición en 1889 mediante la serendipia durante sus estudios sobre los mecanismos que intervienen en la absorción de las grasas intestinales, extirparon el páncreas de un perro y descubrieron involuntariamente que esto provocaba la poliuria y la glucosuria características de la diabetes mellitus (Toelsie *et al.,* 2013).

Unos años más tarde (en 1893), el patólogo francés Gustave Edouard Laguesse postuló que los islotes de Langerhans producían una nueva hormona que desempeñaba un papel regulador de la digestión. Estas estructuras recibieron su nombre en honor a su descubridor Paul Langerhans (1847-1888) quien, como estudiante de medicina en Alemania, las había descrito en 1869, pero no pudo explicar su función. El término *insulina* derivado de la palabra latina "insula" para "isla" fue introducido en 1909 por el médico Jean de Meyer para referirse a la hipotética nueva hormona producida por las islas de Langerhans (Toelsie *et al.,* 2013).

Repitiendo el trabajo de Von Mering y Minkowski, Frederick Grant Banting y Charles Herbert Best confirmaron la existencia de la insulina demostrando que podían revertir la diabetes en perros a los que se les practicó pancreatectomía, tratándolos con un extracto de los islotes de Langerhans de perros sanos. Banting, Best y sus compañeros de trabajo en la Universidad de Toronto (especialmente el químico James

Collip) continuó purificando la insulina del páncreas de los bovinos. Esto condujo a la disponibilidad de un tratamiento eficaz de la diabetes mellitus (inyecciones de insulina) y en 1922 se trató con éxito al primer paciente, Leonard Thompson, un joven que moría de la enfermedad. Banting y el director del laboratorio John MacLeod recibieron el Premio Nobel en 1923 por su logro, y compartieron el dinero del premio con los otros miembros del equipo que no fueron reconocidos, en particular Best y Collip (Toelsie et al., 2013). Banting es honrado por el *Día Mundial de la Diabetes* que se celebra en el día de su cumpleaños, el 14 de noviembre.

1.1.1 Clasificación de la diabetes

Desde el descubrimiento de la insulina, ha habido muchos avances médicos que prolongaron y facilitaron la vida de las personas que sufren de diabetes mellitus. En 1930, por ejemplo, el profesor de medicina británico Sir Harold Percival Himsworth hizo la distinción entre la diabetes mellitus sensible a la insulina (tipo 1 o de inicio temprano) y la diabetes mellitus insensible a la insulina (tipo 2 o de inicio tardío) (Toelsie *et al.,* 2013).

El descubrimiento de estas dos variantes abrió la puerta a nuevas formas más eficaces de tratamiento de la enfermedad. Otro hito fue establecido en 1960 por la física médica americana Rosalyn Yalow, co-ganador del Premio Nobel de 1977 y co-desarrolladora del radioinmuno ensayo. La investigación de la Dra. Yalow condujo al concepto de resistencia a la insulina en pacientes con diabetes mellitus de tipo 2: aunque estos individuos pueden producir suficiente insulina, sus tejidos diana (en particular el tejido adiposo, el músculo y el hígado) suelen responder de forma inadecuada a la hormona (Toelsie *et al.,* 2013).

El descubrimiento de estas dos variantes abrió la puerta a nuevas y más eficaces formas de tratamiento de la diabetes mellitus que progresó de acuerdo con la creciente comprensión en su fisiopatología (American Diabetes Association, 2011). Hoy en día, la diabetes está clínicamente categorizada en la diabetes tipo 1, tipo 2 diabetes, diabetes mellitus gestacional, así como otros subtipos como los causados por la

genética defectos en la función de las células, defectos genéticos en la insulina, trastornos del páncreas, y ciertas drogas. De acuerdo con estos desarrollos, el diagnóstico se han revisado los parámetros de la diabetes repetidamente. En cuanto a su clasificación, la enfermedad se define como:

a) *Diabetes tipo 1* (debido a la destrucción de las células β[9] autoinmunes, que generalmente conduce a una deficiencia absoluta de insulina).
b) *Diabetes tipo 2* (debido a una pérdida progresiva de la secreción adecuada de insulina de células β con frecuencia en el contexto de resistencia a la insulina).
c) *Diabetes mellitus gestacional* (diabetes diagnosticada en el segundo o tercer trimestre del embarazo que no era claramente una diabetes manifiesta antes de la gestación).
d) *Tipos específicos de diabetes debido a otras causas* (diabetes relacionada con síndromes de diabetes monogénica (como diabetes neonatal y diabetes de madurez en los jóvenes), enfermedades del páncreas exocrino (como fibrosis quística y pancreatitis) y enfermedades inducidas por fármacos o productos químicos. diabetes (como con el uso de glucocorticoides, en el tratamiento del VIH / SIDA o después de un trasplante de órganos) (American Diabetes Association, 2020).

1.1.2 Diagnóstico de la diabetes

Los criterios más recientes para definir la diabetes son los propuestos por la *Asociación Americana de la Diabetes*, que se basan en la concentración de glucosa en la sangre, según lo indicado por la prueba A1C; niveles de glucosa en plasma en ayunas (≥ 126 mg/dL o 7.0 mmol/L); niveles de glucosa en plasma de 2 h (≥ 200 mg/dL o 11.1 mmol/L) durante una prueba de tolerancia a la glucosa oral; o síntomas clásicos de hiperglucemia o crisis hiperglicémica, con un plasma aleatorio glucosa ≥ 200 mg/dL o 11.1 mmol/L (American Diabetes Association, 2020; Jiménez Navarrete, 2000).

[9] Las células beta se encargan de producir insulina, y se encuentran en el páncreas.

1.2 Causas de la diabetes

Las principales causas de morbilidad y mortalidad son la obesidad[10] y el aumento de peso, ambos están asociados con un mayor riesgo de diabetes entre las personas. El deterioro de la función de las células β del páncreas muestra una notable progresión a lo largo del tiempo en la diabetes tipo 2. Factores como el envejecimiento, la obesidad, insuficiente consumo de energía, fumar y beber alcohol, etc. desempeñan un papel importante en la patogénesis de la diabetes de tipo I o tipo II (Papatheodorou et al., 2018).

1.2.1 Factores de riesgo modificables

Existe un número de factores de riesgo modificables para padecer diabetes, haciendo cambios saludables en ciertas áreas, las personas pueden reducir sus riesgos o retrasar el desarrollo de la diabetes y mejorar su calidad de vida en general.

a) Sobrepeso/obesidad: Alrededor del 50 por ciento de los hombres y el 70 por ciento de las mujeres que tienen diabetes son obesos. Tener un peso óptimo puede reducir a la mitad el riesgo de desarrollar prediabetes[11], y el riesgo disminuye aún más a medida que se pierde más peso.
b) Inactividad física: Junto con el sobrepeso/obesidad, la inactividad física se encuentra entre los principales factores de riesgo modificables para la prediabetes y el tipo 2 diabetes.
c) Presión arterial alta (hipertensión): Además de causar daños al sistema cardiovascular, la hipertensión no tratada se ha relacionado con el desarrollo de la diabetes.
d) Niveles anormales de colesterol (lípidos): Un bajo nivel de colesterol HDL "malo" y/o altos triglicéridos pueden aumentar el riesgo de diabetes tipo 2. Ambas anomalías pueden aumentar el riesgo de enfermedades cardiovasculares (Papatheodorou et al., 2018).

[10] Según la OMS, la obesidad se define como una acumulación anormal o excesiva de grasa que puede ser perjudicial para la salud

[11] La prediabetes se define como la afección en la que el nivel de azúcar en sangre es elevado, pero no lo suficiente para ser diabetes de tipo 2.

1.2.2 Factores de riesgo no modificables

Hay una serie de factores de riesgo que aumentan el riesgo de que una persona desarrolle prediabetes y, en última instancia, diabetes tipo II, y que están más allá del control de una persona, como:

a) Historial familiar: Si se tiene un familiar que padeció diabetes, el riesgo de desarrollarla aumenta significativamente.
b) Raza u origen étnico: Si se tiene ascendencia afroamericana, asiática-americana, latina/hispanoamericana, nativa americana o de las islas del Pacífico, se tiene una mayor probabilidad de desarrollar diabetes.
c) Edad: Cuanta más edad se tenga, mayor será el riesgo. Generalmente, la diabetes de tipo 2 se presenta en adultos de mediana edad, con mayor frecuencia después de los 45 años.
d) Antecedentes de diabetes gestacional: Si se desarrolla diabetes durante el embarazo o se dio a luz a un bebé de más de 4 kilogramos, hay un mayor riesgo (Papatheodorou et al., 2018).

1.3 Consecuencias de la diabetes

Las consecuencias de la diabetes se pueden agrupar en 2.

1.3.1 Complicaciones microvasculares

Están complicaciones afectan principalmente a la calidad de vida de la persona que la padece, dichas complicaciones son la retinopatía diabética, la nefropatía diabética y la neuropatía diabética (Papatheodorou et al., 2018).

Retinopatía diabética. La retina es la capa más interna y delicada del globo ocular. Si el ojo fuera una cámara de fotos, la retina sería como el sensor que transforma los estímulos de luz en impulsos eléctricos para transmitir al cerebro la imagen. Cuando de forma mantenida en el tiempo existe hiperglucemia[12], las pequeñas arterias de la

[12] La hiperglucemia quiere decir azúcar o glucosa alta en la sangre. Esta glucosa proviene de los alimentos ingeridos.

retina sufren y se lesionan, puede aparecer proliferación de pequeños vasos (Papatheodorou et al., 2018).

Nefropatía diabética. Cada riñón está compuesto de cientos de miles de unidades pequeñas llamadas nefronas. Estas estructuras filtran la sangre, ayudan a eliminar los residuos del cuerpo y controlan el equilibrio de líquidos. En personas con diabetes, las nefronas lentamente se engruesan y cicatrizan. Esto provoca que se empiece a filtrar patológicamente proteínas como la albúmina y aparezcan en la orina. En la progresión de la enfermedad influyen diversos factores: control glucémico, tensión arterial, colesterol, obesidad, tabaquismo, la enfermedad cardiovascular y el consumo de antiinflamatorios (Papatheodorou et al., 2018).

Neuropatia diábetica. Cuando la diabetes está mal controlada también puede afectar al sistema nervioso. Cuando los nervios que controlan la digestión resultan afectados, pueden aparecer problemas para digerir los alimentos y esto además puede dificultar el control de la diabetes (Papatheodorou et al., 2018).

1.3.2 Complicaciones macrovasculares

Arterosclerosis. El mecanismo patológico central de la enfermedad macrovascular es el proceso de arterosclerosis, que conduce al estrechamiento de las paredes arteriales en todo el cuerpo. Se cree que la arterosclerosis es el resultado de una inflamación crónica y una lesión de la pared arterial en el sistema vascular periférico o coronario. El resultado neto del proceso es la formación de una lesión aterosclerótica rica en lípidos con un capuchón fibroso. La ruptura de esta lesión provoca un infarto vascular agudo (Papatheodorou et al., 2018).

Cardiopatía coronaria. Entre las complicaciones de la diabetes macrovascular, la cardiopatía coronaria se ha asociado con la diabetes. Estudios más recientes han demostrado que el riesgo de infarto de miocardio (IM) en personas con diabetes es equivalente al riesgo en pacientes no diabéticos con antecedentes de IM (Papatheodorou et al., 2018).

La diabetes de tipo 2 se produce típicamente en el contexto del síndrome metabólico[13], que también incluye obesidad abdominal, hipertensión, hiperlipidemia y aumento de la coagulabilidad. Estos otros factores también pueden actuar para promover las enfermedades cardiovasculares. Incluso en este entorno de múltiples factores de riesgo, la diabetes de tipo 2 actúa como un factor de riesgo independiente para el desarrollo de la enfermedad isquémica, el accidente cerebrovascular y la muerte. La presencia de enfermedades microvasculares es también un factor de predicción de los acontecimientos cardíacos coronarios (Papatheodorou et al., 2018).

Derrame cerebral. La diabetes es también un fuerte predictor independiente del riesgo de derrame cerebral y enfermedad cerebrovascular, como en la enfermedad de las arterias coronarias. Los pacientes con diabetes de tipo 2 tienen un riesgo mucho mayor de derrame cerebral, con un aumento del riesgo de 1.5 al 4 veces. El riesgo de demencia y recurrencia relacionadas con los accidentes cerebrovasculares, así como la mortalidad relacionada con ellos, es elevado en los pacientes con diabetes (Papatheodorou et al., 2018).

1.3.2 Consecuencias sociales y económicas

Los costos directos de la atención de la salud de las personas con diabetes son generalmente 2-3 veces más que para los que no tienen diabetes u otros padecimientos relacionados y 4-8 veces más si hay complicaciones de la diabetes. Los costos indirectos de la diabetes se vuelven cada vez más importantes. En los países de bajos ingresos cuando los costos como el de los viajes a clínicas y la pérdida de ingresos tienen un mayor impacto en toda la familia. Las investigaciones económicas sobre la salud han identificado este efecto dominó en las familias de bajos ingresos que soportan una mayor proporción de los gastos de salud como gastos de bolsillo (Silink, 2009).

[13] El síndrome metabólico se define como el conjunto de trastornos que aumentan el riesgo de padecer enfermedades cardíacas, un derrame cerebral y diabetes.

La diabetes impone una gran carga económica al sistema mundial de atención de la salud y a la economía mundial en general. Esta carga puede ser medido a través de los costos médicos directos, los costos indirectos asociados con la pérdida de productividad, la mortalidad prematura y el impacto negativo de la diabetes en el producto interno bruto (PIB) de las naciones (O'Connell y Manson, 2019).

Los costos médicos directos asociados con la diabetes incluyen gastos para prevenir y tratar la diabetes y sus complicaciones. Estos incluyen la atención ambulatoria y de emergencia; hospitalización, cuidado, medicamentos y atención médica; suministros como los dispositivos de inyección y consumibles de autocontrol; y cuidados a largo plazo. Sobre la base de las estimaciones de costos de un examen sistemático reciente, ha se ha estimado que el costo anual directo de la diabetes para el mundo es de más de 827,000 millones de dólares. La Federación Internacional de Diabetes (FID) estima que el gasto total mundial en atención de la salud en materia de diabetes se ha triplicado con creces en el período comprendido entre 2003 y 2013, como resultado del aumento del número de personas con diabetes y el incremento de la diabetes per cápita de gasto, en 2003 el gasto anual se estimó en 4060 dólares, mientras que para 2013 el gasto fue estimado en 11360 dólares (O'Connell y Manson, 2019).

Además de la carga económica sobre el sistema de salud se cada país, también se ven afectadas las familias en las que algún integrante padece diabetes por la pérdida de ingresos familiares asociada con discapacidad y pérdida prematura de la vida (Forsham, 1982).

La relación entre la diabetes y el riesgo de que se produzca un escenario de gastos médicos mayores de los individuos y las familias ha sido explorada en 35 países en desarrollo, en los cuales las personas con diabetes tienen una significativamente mayor posibilidad de incurrir en un gasto médico grande comparado a individuos en estatus sociales similares sin diabetes. Los efectos asociados a la diabetes son más marcados en los países de ingresos más bajos. La diabetes es reconocida como una

importante causa de muerte prematura y discapacidad. Es una de las cuatro enfermedades no transmisibles prioritarias en las políticas globales de salud (Forsham, 1982).

CAPÍTULO 2. EL CONTEXTO DE LA DIABETES MELLITUS

2.1 Diabetes en el mundo

A nivel mundial, se estima que 422 millones de adultos vivían con diabetes en 2014, en comparación con 108 millones en 1980 (Organización mundial de la salud, 2016). La prevalencia mundial (normalizada por edad) de la diabetes casi se ha duplicado desde 1980, pasando del 4.7% al 8.5% en la población adulta.

En el último decenio, la prevalencia de la diabetes ha aumentado más rápidamente en los países de ingresos bajos y medianos que en los países de ingresos medianos-altos y que en los países de altos ingresos. Alemania en el 2010 tenía un índice de diabetes de 8.9 por ciento, mientras que en 2019 aumentó a 10.4 por ciento, mientras que en India paso de un índice de 7.8 por ciento a 10.4 por ciento en 2010 y 2019 respectivamente (Abolghasem, Sepideh; Amézquita, 2019).

La diabetes causó 1.5 millones de muertes en 2012 (Organización Mundial de la Salud, 2016). Un nivel de glucosa en la sangre superior a la óptima causó un aumento de 2.2 millones de muertes, al aumentar los riesgos de enfermedades cardiovasculares y de otro tipo (Organización Mundial de la Salud, 2016). El 43 por ciento de estos 3.7 millones de muertes ocurren antes de los 70 años. El porcentaje de muertes atribuible a los elevados niveles de glucosa en la sangre, que se produce antes de los 70 años, y que es más alta en los países de ingresos bajos y medios que en los de ingresos altos, debido a que normalmente se requieren sofisticadas pruebas de laboratorio para distinguir entre la diabetes de tipo 1 (que requiere inyecciones de insulina para sobrevivir) y la diabetes de tipo 2 (en la que el cuerpo no puede utilizar adecuadamente la insulina que produce), no se han hecho estimaciones mundiales separadas de la prevalencia de la diabetes para el tipo 1 y el tipo 2. La mayoría de las personas con diabetes se ven afectadas por la diabetes de tipo 2. Esto solía ocurrir casi exclusivamente entre los adultos, pero ahora también ocurre en los niños.

2.2 Diabetes en algunas potencias mundiales

Durante las crisis económicas, la salud de la población de un país empeora debido a la reducción de los ingresos familiares y al menor acceso a la atención de la salud. Los pobres de los países de bajos ingresos son los más afectados porque pagan una gran parte de sus gastos de salud de su bolsillo sin el beneficio de las redes de seguridad social. Las repercusiones económicas de las enfermedades cardiovasculares, los accidentes cerebrovasculares y la diabetes en los países en desarrollo son severas.

La OMS estima que la mortalidad por diabetes, enfermedades cardíacas y accidentes cerebrovasculares cuesta unos 250,000 millones de dólares internacionales en China, 225,000 millones en la Federación de Rusia y 210,000 millones en la India. Gran parte de las enfermedades cardíacas y los accidentes cerebrovasculares que figuran en esas estimaciones están vinculadas a la diabetes. La OMS estima que la diabetes, las enfermedades cardíacas y los accidentes cerebrovasculares costarán en conjunto unos 555,000 millones de dólares de los EE.UU. en pérdidas de ingresos en China durante el período 2005-2010, 303,000 millones de dólares de los EE.UU., 2,500 millones de dólares incluso en un país muy pobre como Tanzania. Estas estimaciones se basan en la productividad perdida, que resulta principalmente de una muerte prematura (Organización Mundial de la Salud, 2016).

La prevención de la diabetes propiamente dicha (prevención primaria) y sus complicaciones (prevención secundaria) son en gran medida factibles y claramente deseables. A nivel mundial, los beneficios de prevenir que 7 millones más de personas tengan diabetes anualmente serían considerables. La mayoría de las economías desarrolladas ya están gastando entre el 10% y el 12% de sus presupuestos de atención de la salud en la diabetes (Silink, 2009).

2.2.1 Diabetes en países de ingresos altos

Un panorama general de la distribución y las condiciones de la diabetes puede apreciarse entre los países de ingresos altos como los Estados Unidos de América,

Suecia, los Países Bajos y China, en los países de ingresos medios y bajos como India y México. A pesar de esto, debe señalarse que incluso entre estas sociedades existen diferencias en la distribución y factores asociados a la diabetes. La definición utilizada para países de ingresos altos, medios y bajos se retoma de la clasificación del Banco Mundial (Fantom y Serajuddin, 2016).

Diabetes en los Estados Unidos de América. Más de 30 millones de personas sufrieron de diabetes en 2015, lo que representa el 9.4% de toda la población de los Estados Unidos y el 12.2% de todos los adultos. Una de cada cuatro de estas personas no sabía que tenía diabetes. Además, en el mismo año, se creía que aproximadamente 84.1 millones de adultos estadounidenses, más de un tercio de los adultos estadounidenses (33.9%), tenían prediabetes, el 90% de los cuales tampoco sabían que padecían la enfermedad.

La alta prevalencia de la diabetes mellitus gestacional (DMG) (6.0%) entre las mujeres que dieron a luz en 2016 subraya aún más la gravedad de esta epidemia. La carga de la diabetes no se comparte de manera uniforme, ya que varía en función de la edad, la educación, los ingresos, la ubicación, la raza/etnia y otros determinantes sociales de la salud (O'Connell y Manson, 2019).

Es evidente que la carga es mayor entre los adultos con menor nivel de educación e ingresos familiares que entre los adultos de mayor nivel socioeconómico, disparidades que se han ampliado con el tiempo. En 2015, en comparación con los blancos no hispanos (7.4%), la prevalencia ajustada por edad de la diabetes diagnosticada fue mayor entre los adultos indígenas americanos y nativos de Alaska (15.1%), los adultos negros no hispanos (12.7%), los adultos de etnia hispana (12.1%) y los adultos asiáticos (8.0%). La prevalencia de la diabetes no diagnosticada, la prediabetes y la diabetes mellitus gestacional también varió (O'Connell y Manson, 2019).

Los gobiernos en sus niveles federal, estatal y local soportan la mayor parte de los costos relacionados con la diabetes. Por ejemplo, la carga relacionada con la diabetes

de Medicare[14], aumentó en los últimos años al incrementar la prevalencia de diabetes. Es fundamental que los gobiernos, los empleadores, otros contribuyentes y los proveedores de servicios de salud dispongan de información actualizada y exhaustiva sobre la carga económica de las afecciones relacionadas con la diabetes, a fin de evaluar las oportunidades de mejorar la prestación de servicios y, en última instancia, los resultados en materia de salud (O'Connell y Manson, 2019).

La American Diabetes Association cifró el costo de la diabetes diagnosticada en los Estados Unidos en 2017 en 327,200 millones de dólares. La diabetes no diagnosticada (7.9%, 31,700 millones de dólares), la prediabetes (10.7%, 43,400 millones de dólares) y la diabetes mellitus gestacional (0.4%, 1.600 millones de dólares) se combinan con la estimación previa de la diabetes diagnosticada para totalizar 403,900 millones de dólares anuales (O'Connell y Manson, 2019).

La diabetes diagnosticada representaba el 81.0% de este total. Se proyectó que el costo económico promedio por persona era de 13,240 dólares para la diabetes diagnosticada, 4,250 dólares para la diabetes no diagnosticada, 500 dólares para la prediabetes y 5,800 dólares para la diabetes gestacional. Esas estimaciones incluían los gastos médicos que superaban los niveles que se producían en ausencia de diabetes o prediabetes, así como los costos indirectos debidos a las pérdidas de productividad asociadas a la morbilidad y la mortalidad conexas. No es de extrañar, pues, que el gasto médico para la diabetes diagnosticada se encuentre entre los más elevados para todas las afecciones. Para situar los costos médicos e indirectos totales de 2017 dentro de la economía de los Estados Unidos, se comparan esos costos con el producto interno bruto de los Estados Unidos; los 403,900 millones de dólares de costos económicos de la diabetes y la prediabetes son aproximadamente el 2.1% del PIB de los Estados Unidos en 2017 (O'Connell y Manson, 2019).

[14] Medicare es un programa de cobertura de seguridad social administrado por el gobierno de Estados Unidos, el cual provee atención médica a población vulnerable, generalmente adultos mayores y discapacitados.

A la luz de estas circunstancias, aumentar el acceso a programas para prevenir la diabetes, la prediabetes y los factores de riesgo asociados a estas condiciones (por ejemplo, la obesidad, la falta de ejercicio) se vuelve aún más crucial. El Programa Nacional de Prevención de la Diabetes (PPD), específicamente la intervención individual en el estilo de vida y el uso de metformina, fueron efectivos en la prevención de la diabetes y rentables. En estudios posteriores se determinó que los programas de grupo de estilo de vida del PPD eran eficaces en función de los costos. De hecho, un estudio de Medicare de tales programas reveló que estaban asociados con reducciones significativas en los gastos de Medicare, admisiones de pacientes hospitalizados y visitas a la sala de emergencias en el grupo de intervención en relación con el grupo de comparación. Los programas PPD pueden calificar para el reembolso por parte de Medicare, lo que lleva a muchos planes de salud privados y a algunas agencias estatales de Medicaid[15] a apoyar su expansión. Dicha expansión, en particular entre las poblaciones de alto riesgo, puede tener un impacto considerable en la prevalencia de la diabetes y los costos asociados (O'Connell y Manson, 2019).

En 2018, 34.2 millones de estadounidenses, o el 10.5% de la población, tenían diabetes. Casi 1.6 millones de estadounidenses tienen diabetes tipo 1, incluyendo unos 187,000 niños y adolescentes. De los 34.2 millones de adultos con diabetes, 26.8 millones fueron diagnosticados, y 7.3 millones no fueron diagnosticados. El porcentaje de estadounidenses mayores de 65 años sigue siendo alto, con un 26.8%, o 14.3 millones de ancianos (diagnosticados y no diagnosticados). Aproximadamente 1.5 millones de estadounidenses son diagnosticados con diabetes cada año. En 2015, 88 millones de estadounidenses mayores de 18 años tenían prediabetes (O'Connell y Manson, 2019).

Diabetes en Suecia. Los costos anuales medios en Suecia asociados a los ingresos hospitalarios de un varón de 60 años de edad en el año en que se produjo el primer evento son de aproximadamente 6,488 euros por coma diabético; 6,850 euros por

[15] Medicaid es un programa de cobertura de seguridad social administrado por el gobierno de Estados Unidos, el cual provee atención médica, generalmente a niños y personas de bajos recursos

insuficiencia cardiaca; 7,853 euros por accidente cerebrovascular no mortal; 8,121 euros por complicaciones circulatorias periféricas; 8,736 euros por IM no mortal; 10,360 euros por cardiopatía isquémica; 11,411 euros por insuficiencia renal; y 14,949 euros por amputación. En promedio, los costos son más altos cuando se tiene en cuenta la comorbilidad[16] (Gerdtham et al., 2009).

Entre 2007 y 2013 la prevalencia de la diabetes aumentó del 5.8 al 6.8% en Suecia, pero la incidencia se mantuvo constante en el 4.4 por 1000. Con una incidencia constante y una mejora continua de la supervivencia relativa, la prevalencia aumentará hasta el 10.4% para el año 2050 y el número de personas afectadas aumentará hasta 940,000. De este aumento, el 30% se debe a los cambios en la estructura de edad de la población y el 14% a la mejora de la supervivencia relativa de las personas con diabetes. La hipótesis de un aumento anual del 1% en la incidencia se traducirá en una prevalencia del 12.6% y 1,136,000 casos. Incluso con una incidencia[17] decreciente del 1% anual, la prevalencia de la diabetes seguirá aumentando. Se espera que la prevalencia de la diabetes aumente sustancialmente en Suecia en los próximos 35 años como resultado de los cambios demográficos y la mejora de la supervivencia de las personas con diabetes (Andersson et al., 2015).

Diabetes en los Países Bajos. En los Países Bajos la prevalencia total de la diabetes conocida por encima de los 30 años de edad en 1993 fue estimada en un 2.7% sobre la base de los casos notificados por los médicos generales y en un 3.2% sobre la base de los casos notificados por los propios pacientes entrevistados en las encuestas. La prevalencia según la edad aumentó en un 7.1% por año de vida para los hombres y en un 7.7% para las mujeres (Baan et al., 1998).

Estas asociaciones fueron esencialmente similares en todos los estudios, como el examen sistemático, realizado con una glucosa oral prueba de tolerancia, el cual

[16] La comorbilidad se define como la presencia de uno o más trastornos además de la enfermedad o trastorno primario.

[17] La incidencia es la influencia de determinada cosa en un asunto o efecto que causa en él.

reveló una prevalencia que era de 1.5 a 2 veces mayor. Según el método utilizado, el número de los sujetos con diabetes conocida en los Países Bajos en 1993 variaron entre 235,000 y 285,000 (Baan et al., 1998).

En hogares de ancianos y casas de reposo de los Países Bajos es donde existe mayor prevalencia de diabetes, 13% para los hombres de 75-79 años y el 21% para mujeres de esta edad. El 4.3% de todos los hombres y el 7.9% de todas las mujeres de 75 a 79 años están registrados como diabéticos (Baan et al., 1998). Los Países Bajos parecen tener un nivel bajo en la jerarquía en comparación a otros países europeos. En otros países europeos, la prevalencia de la diabetes conocida entre 30 y 64 años años de edad varía desde el 1.8% en Rusia hasta el 10.7% en Italia (Baan et al., 1998).

La prevalencia de la diabetes diagnosticada aumenta adicionalmente para los hombres en un 7.1% por cada año de edad y en un 7.7% para mujeres. La edad parece ser un predictor fiable para predecir el número de pacientes con diabetes en una sociedad que envejece (Baan, 1998).

Diabetes en China. Al ser el país más poblado, el rápido aumento de la morbilidad y la mortalidad por enfermedades no transmisibles[18] en China contribuyó en gran cantidad a esta pandemia de diabetes. Las enfermedades no transmisibles representaron aproximadamente el 80% de las muertes y el 70% de la carga total de morbilidad en China en 2005. La prevalencia de la diabetes era inferior al 1% en la población china en 1980. En encuestas nacionales posteriores realizadas en 1994 y 2000-2001, la prevalencia de la diabetes era del 2.5% y el 5.5%, respectivamente (Yu et al., 2012).

Los datos documentan un rápido aumento de la diabetes en la población china para 2010, sin embargo la prevalencia de la diabetes podría haber sido subestimada en las

[18] Las enfermedades no transmisibles también llamadas crónicas son afecciones de larga duración con una progresión generalmente lenta.

encuestas nacionales basadas en los criterios de la ADA (American Diabetes Association, 2011), la prevalencia de la diabetes era de aproximadamente 11.6% en los adultos chinos, del 12.1% en los hombres y del 11.0% en las mujeres, con una prevalencia estimada del 8.1% para la diabetes recientemente detectada: 8.5% en los hombres y 7,7% en las mujeres y fue de 3,5% para aquellos con diabetes previamente diagnosticados: 3,6% en los hombres y 3,4% en las mujeres (Yu et al., 2012). Entre los tres parámetros glucémicos, una concentración de glucosa en plasma de dos horas de duración de 200 mg/dL o más fue menos frecuente (3,5%) que una concentración de glucosa en plasma en ayunas de 126 mg/dL o más (4. 5%) o una concentración de HbA1 de 6.5% o más (4.6%) entre individuos sin antecedentes de diabetes (Yu et al., 2012).

La prevalencia de la diabetes es mayor en los residentes urbanos que en los rurales, tanto en hombres como en mujeres. Además, la prevalencia de la diabetes aumentó con la edad tanto en hombres como en mujeres, y los hombres menores de 50 años tienen una mayor prevalencia, mientras que las mujeres mayores de 60 años tienen una mayor prevalencia. Además, la prevalencia de la diabetes aumentó con el desarrollo económico, al igual que en el caso del sobrepeso y la obesidad. La prevalencia estimada de la prediabetes fue del 50.1% en los adultos chinos: 52.1% en los hombres y 48.1% en las mujeres (Yu et al., 2012). La prevalencia estimada por la glucosa plasmática en dos horas solamente fue mucho más baja que la de la glucosa plasmática en ayunas o la de la HbA1. Los residentes de rurales tienen una prevalencia ligeramente mayor de prediabetes que los residentes urbanos, especialmente en los hombres. La prevalencia de la prediabetes aumenta con la edad, y es mayor en los hombres menores de 50 años. Además, la prediabetes es más frecuente en las regiones económicamente subdesarrolladas, así como en las personas con sobrepeso y obesas (Yu et al., 2012).

Para 2019, China era el país con mayor número de personas con diabetes (114 millones), principalmente de tipo 2, dentro de una sola nación. El control médico de la

diabetes, solo, antes de complicaciones, se estima que representó el 8.5% del gasto nacional en salud en China en 2019 (Luo et al., 2020).

2.2.2 Diabetes en países de ingresos bajos

Diabetes en la India. En 2005 había en la India aproximadamente 33 millones de personas diabéticas, a lo que contribuye brevemente la población urbana. El panorama cambió rápidamente debido a la transición socioeconómica que se está produciendo también en las zonas rurales. La disponibilidad de medios de transporte mejorados, y menos intensos como en las cercanías, dieron lugar a una disminución de las actividades físicas (Ramachandran, 2005).

Las mejores condiciones económicas produjeron cambios en los hábitos alimentarios. Las condiciones son más favorables para la expresión de la diabetes en la población, que ya tiene una susceptibilidad genética a la enfermedad. Las condiciones prediabéticas, como la disminución de la tolerancia a la glucosa y la disminución de la glucosa en ayunas, también van en aumento, lo que indica la posibilidad de que siga aumentando la prevalencia de la diabetes. El síndrome metabólico, que es una constelación de factores de riesgo cardiovascular, de los que la hiperglucemia y la resistencia a la insulina son componentes, también está ampliamente difundido. El padecer diabetes se favoreció por los bajos umbrales de los factores de riesgo, como la edad, el índice de masa corporal y la adiposidad de la parte superior del cuerpo (Ramachandran, 2005).

Las personas indígenas tienen un fenotipo genético caracterizado por un bajo índice de masa corporal[19], pero con una elevada adiposidad de la parte superior del cuerpo, un alto porcentaje de grasa corporal y un alto nivel de resistencia a la insulina. Con una alta predisposición genética y la gran susceptibilidad a los insultos del medio

[19] El índice de masa corporal (IMC) es un indicador simple de la relación entre el peso y la talla que se utiliza frecuentemente para identificar el sobrepeso y la obesidad en los adultos. Se calcula dividiendo el peso de una persona en kilos por el cuadrado de su talla en metros (kg/m2)

ambiente, la población india se enfrenta a un alto riesgo de diabetes y sus complicaciones asociadas (Ramachandran, 2005).

En 2015 se estimó que 66.8 millones de personas padecían diabetes en India. Ese aumento de la carga de la diabetes afectó en gran medida al sector de la atención de la salud y la economía de la India. En india casi la mitad de las personas con diabetes no son detectadas, lo que genera complicaciones en el momento del diagnóstico. A pesar del gran número de personas con diabetes en la India, la concientización es escasa, generalmente no se aborda (Joshi, 2015).

Estudiando la prevalencia de diabetes y la tolerancia anormal a la glucosa en una zona rural del distrito de Nagpur, se observó que 3.6% de los habitantes padecían diabetes, 5.9% tenían una tolerancia a la glucosa disminuida y 3.5% tenían glicemia en ayunas. Se observó que 13.2% tenía una tolerancia anormal a la glucosa. La prevalencia fue alta en aquellos que pertenecen a las clases socioeconómicas altas (23.6%) en comparación con las personas que pertenecen a los sectores socioeconómicos más bajos (8.9%), en los que consumen alcohol (22.2%) en comparación con los no alcohólicos (11.4%), en los que tienen IMC ≥25 kg/m2 (27.4%) en comparación con los que tienen un IMC <25 kg/m2 (9.7%), en las personas que tienen familia historia de la diabetes (46.9%) en comparación con aquellos con no hay tal historia (11.3%) (Singh, 2016).

En la ciudad de Tamilnadu, la diabetes aumentó del 13.9 al 18.6% en seis años y el deterioro de la tolerancia a la glucosa disminuyó significativamente. La ciudad de Kanchipuram y la ciudad de Chennai tienen una prevalencia similar, las aldeas tienen una menor prevalencia de la diabetes, pero la prevalencia había aumentado en comparación con 2006 (Singh, 2016).

Por otro lado, se detectó diabetes mellitus gestacional en 739 mujeres (17.8%) en zonas urbanas, 548 en las zonas semiurbanas (13.8%) y 392 en las rurales (9.9%). De 1,679 mujeres con diabetes mellitus gestacional, 1,204 se detectaron en la primera

visita (72%) y el resto 28% en las visitas posteriores. Un aumento significativo en la prevalencia de la diabetes mellitus gestacional se observó con antecedentes familiares de diabetes, aumento de la edad materna y del índice de masa corporal (Singh, 2016).

2.3 Diabetes en México

El desafío para la sociedad y los sistemas de salud de México es enorme, debido al costo económico y la pérdida de calidad de vida para quienes padecen diabetes y sus familias, así como por los importantes recursos que requieren en el sistema público de salud para su atención. En México, las estimaciones existentes al 2013 son muy variables con cálculos de costos de atención por paciente que van desde 700 hasta 3,200 dólares anuales, lo que se traduce en 5 a 14% del gasto en salud destinado a la atención de esta enfermedad y sus complicaciones, inversión que de acuerdo con la Federación Internacional de Diabetes se relaciona directamente con la tasa de mortalidad por esta causa (Hernández-Ávila et al., 2013).

Los estilos de vida poco saludables son altamente prevalentes en la población mexicana, propiciando un aumento importante de la obesidad y sobrepeso, principal factor de riesgo modificable de la diabetes (Hernández-Ávila et al., 2013). Así, la prevalencia de la diabetes en esta población ha incrementado sustancialmente en las últimas décadas: en 1993 la prevalencia de los diabéticos con diagnóstico conocido en población mayor de 20 años fue de 4%, mientras que en 2000 y 2007 se describió una prevalencia del 5.8 y 7%, respectivamente (Hernández-Ávila et al., 2013). Por otro lado, de acuerdo con las encuestas nacionales de esos mismos años, se ha demostrado la alta prevalencia de condiciones comórbidas en la población diabética y problemas en la calidad de la atención, lo cual contribuye de manera importante a la mayor incidencia de complicaciones macro y microvasculares (Hernández-Ávila et al., 2013).

Las estrategias de prevención implementadas a nivel nacional en países con elevado riesgo que logren modificar estilos de vida, en particular en la dieta, actividad física y

tabaquismo, pueden ser altamente costo efectivas al reducir la aparición de la diabetes y retrasar la progresión de la misma. México tiene condiciones de alto riesgo, por lo que recientemente se han impulsado políticas intersectoriales relacionadas con la salud alimentaria y con ello combatir uno de los más importantes factores de riesgo para diabetes, la obesidad (Hernández-Ávila et al., 2013). Al mismo tiempo se han diseñado, ya desde hace más de una década, estrategias (PREVENIMSS, PREVENISSSTE), grupos de autoayuda, Unidades de Especialidades Médicas para Enfermedades Crónicas, entre otras al interior de las principales instituciones de salud con el propósito de mejorar la atención que se otorga a los pacientes que ya padecen la enfermedad (Hernández-Ávila et al., 2013).

Sin embargo, el estado actual de los diabéticos mexicanos se conoce sólo parcialmente, información que es necesaria para cimentar y fortalecer los esfuerzos que se requieren en prevención a todos los niveles a fin de contener una de las más grandes y emergentes amenazas de la viabilidad de los sistemas de salud: la diabetes (Hernández-Ávila et al., 2013).

La diabetes mellitus ha mostrado un comportamiento epidémico en México desde la segunda mitad del siglo pasado. En la actualidad, México es uno de los países con mayor ocurrencia de diabetes mellitus en el mundo. En 1995 ocupaba el noveno lugar con mayor número de casos de diabetes y se espera que para el año 2030 ocupe el séptimo con casi 12 millones de pacientes con diabetes tipo 2. La diabetes es actualmente la primera causa de mortalidad en México y su tendencia muestra un incremento progresivo en los últimos años (Escobedo-De La Peña et al., 2011).

En 2008 hubo más de 75,500 defunciones por diabetes en el país, para una tasa de mortalidad de 70.9%. En la población con derechohabiencia al Instituto Mexicano del Seguro Social, la diabetes es la primera causa de mortalidad, de años perdidos por muerte prematura, de años vividos con discapacidad y de años de vida saludable perdidos. En el 2000, la diabetes contribuyó con 13.3% de los años de vida saludables perdidos en el IMSS, que se estima en 970,000 (Escobedo-De La Peña et al., 2011).

La diabetes es un claro ejemplo de la transición epidemiológica[20] que vive el país, así como de la transición de la atención a la salud. Se ha estimado que los costos de la atención a la diabetes en México superan los 300 millones de dólares al año (2010) y el comportamiento muestra un patrón ascendente en los próximos años (Escobedo-De La Peña et al., 2011).

2.3.1 Estadísticas de la diabetes por estados

En el año 2008, la Ciudad de México concentró 12% de las defunciones por diabetes en hombres en el país y su tasa de mortalidad ajustada por edad, la segunda más alta en el país (123 por 100,000 hombres). En las mujeres, 11% de las defunciones por diabetes ocurre en la Ciudad de México, para una tasa ajustada por edad de 94.0 por 100 000 mujeres. En la Encuesta Nacional de Salud (ENSA) realizada en el año 2000, la prevalencia de diabetes en la Ciudad de México fue de 8.5%, que representó el séptimo lugar de mayor ocurrencia entre todos los estados del país (Escobedo-De La Peña et al., 2011).

Del total de la población de adultos en México, 9.17% reportó tener un diagnóstico previo de diabetes por un médico, lo que equivale a 6.4 millones de personas. Por sexo, este porcentaje fue de 8.60% entre los hombres y 9.67% entre las mujeres, lo que equivale a 2.84 millones de hombres y 3.56 millones de mujeres. Por sexo, en el caso de los hombres las entidades con mayor proporción de individuos con diagnóstico de diabetes son la Ciudad de México (12.7%), Estado de México (11.5%), y Veracruz (10.7%), en tanto que, para las mujeres, las entidades con mayor proporción de personas con diagnóstico de diabetes son Nuevo León (15.5%), Tamaulipas (12.8%), y la Ciudad de México (11.9%) (Hernández-Ávila et al., 2013).

[20] El cambio general de enfermedades infecciosas agudas y deficiencias características de las enfermedades crónicas no transmisibles características de la modernización y los niveles avanzados de desarrollo se conoce generalmente como la "transición epidemiológica" (Wahdan, 1996) .

Del total de personas con diagnóstico de diabetes, únicamente 85.7% atiende esta condición de salud. De ellos, la mayoría acude al IMSS (39%), en segundo lugar a instituciones financiadas por el entonces Sistema de Protección Social en Salud (SPSS) (28.2%), seguido del sector privado (21.3%) y otras instituciones de seguridad social (11.4%). Los que no se atienden presentan una importante variación por condición se aseguramiento: en tanto que únicamente 4% de los que reportaron contar con aseguramiento privado no se atiende, 27.5% de los diabéticos que no cuentan con protección en salud (cerca de 280,000 individuos) no ha acudido para atenderse de este padecimiento durante al menos un año. Entre los afilados al SPSS, el porcentaje de los que no se atienden es de 13% (256,000 personas) y de 11% (378,000 personas) para los que cuentan con derechohabiencia a la seguridad social (Hernández-Ávila et al., 2013).

En términos de las diferencias por nivel socioeconómico (NSE), para los cinco quintiles se observa el incremento con la edad en la proporción de personas con diagnóstico previo, tanto para hombres como mujeres. De forma general, se encontraron proporciones menores entre las personas del primer quintil (menor nivel) que en todos los casos presentan proporciones menores al promedio para el grupo de edad y sexo, en tanto que en todos los casos, las personas en el quinto quintil (mayor nivel) presentan proporciones de diagnóstico de diabetes mayores al promedio del grupo de edad y sexo (Hernández-Ávila et al., 2013).

En lo que se refiere a la población de adolescentes, el diagnóstico previo de diabetes se reportó para 0.6% de los adolescentes, siendo de 0.5% entre los hombres, y 0.7% entre las mujeres. Esto representa alrededor de 155 000 individuos en este rango de edad que ya han sido diagnosticados con diabetes (Hernández-Ávila et al., 2013).

De acuerdo con la Encuesta Nacional de Salud y Nutrición (ENSANUT) en 2016 la prevalencia total de diabetes fue de 13.7% (9.5% diagnosticada, 4.1% no diagnosticada); 68.2% de los diagnosticados presentó descontrol glucémico. Mayor tiempo de diagnóstico, vivir en el centro/sur del país y ser atendido en farmacias se

asoció con descontrol glucémico, mientras que ser atendido en los servicios de seguridad social se asoció con mejor control glucémico (Basto-Abreu et al., 2020).

La prevalencia de diabetes por diagnóstico médico en 2016 fue de 9.4%. El incremento de 2.2% respecto a 2012 no fue significativo y se observó únicamente en los mayores de 60 años. Aunque las acciones preventivas han aumentado, el acceso al tratamiento médico y los estilos de vida no han mejorado. Se observó un aumento en insulina y una disminución en hipoglucemiantes (Rojas-Martínez et al., 2018).

Si bien México es un país afectado gravemente por la diabetes, es importante que se logre caracterizar la distribución de la enfermedad en el nivel subnacional (estatal, municipal, grupal) y determinar las mejores acciones para enfrentar las consecuencias sociales, económicas y de la salud. Este es el caso de la situación de la población PIMA de México que ha mostrado una de las prevalencias de diabetes más elevadas en el mundo y que constituye la población de análisis de esta tesis.

2.3.2 Poblaciones específicas de México con diabetes: los PIMA de México

Los indígenas Pima de México se localizan en una zona de la sierra denominada zona baja, que comprende una porción oriental del estado de Sonora y parte del occidente de Chihuahua, desde las elevaciones de la sierra Madre Oriental hasta los cauces bajos de los ríos Sonora, Matape y Yaqui; y la Pimería alta, que se ubica entre los 30 y 34 grados de latitud, en las tierras desérticas del noroeste de Sonora y el suroeste de Arizona.

Hoy, los pimas bajos, quienes en mayor número representan al pueblo pima en territorio mexicano, se concentran en las áreas circundantes a Maycoba, en el municipio de Yécora, Sonora, y a Yepáchic y Mesa Blanca, en los municipios chihuahuenses de Temosáchic y Madera, respectivamente. La población de Maycoba está constituida por dos grupos de pobladores, los indígenas Pimas y los blancos (término local para las personas no relacionadas consanguíneamente con el grupo Pima ni con ningún otro grupo indígena). Los blancos habitan principalmente en el

poblado rural de Maycoba, mientras que alrededor del 50% de los indígenas Pima habitan en Maycoba y el resto en las áreas periféricas del poblado y pequeños ranchos familiares aledaños (Urquidez-Romero et al., 2015).

Los indígenas Pima mexicanos tienen un consumo de grasa y fibra dietaria, y mayor nivel de actividad física comparados con los Pima de Estados Unidos. El estudio de los Pima indica que aun en poblaciones genéticamente predispuestas a estas condiciones, su desarrollo puede estar determinado principalmente por circunstancias relacionadas con el estilo de vida (Urquidez-Romero et al., 2015).

Las actividades de los pima mexicanos se basan en el cultivo de unos cuantos productos. El cultivo del maíz, parte fundamental de la economía pima desde tiempos prehispánicos, al igual que el del trigo y la papa, se rota año con año para hacer más productivos los campos. Los pimas cultivan con azadón y palo sembrador o coa y generalmente compran o rentan animales para arar. Complementan su producción con la cría de animales domésticos. También practican la caza y la recolección. La distribución de la tierra arable determina la ubicación de sus rancherías. Desde hace muchos años, los pimas bajan de la sierra para contratarse como jornaleros en los campos agrícolas de algodón, tomate, maíz y uva en Sonora, algunos se van hasta California o Arizona. Migran en pequeños grupos de parientes, hombres jóvenes, a veces mujeres y familias completas que generalmente van a lugares donde ya están establecidos otros pimas (Urquidez-Romero et al., 2015).

La menor prevalencia de diabetes mellitus 2 en Pimas mexicanos, en comparación con la presentada por los Pimas de EUA, parece estar explicada mayormente por las diferencias contrastantes en estilos de vida entre ambos grupos, esto indica que, en poblaciones genéticamente predispuestas a la diabetes, su desarrollo está influido principalmente por circunstancias ambientales, condiciones que son susceptibles a modificarse como método de prevención (Urquidez-Romero et al., 2015).

CAPÍTULO 3. METODOLOGÍA DE INVESTIGACIÓN

El objetivo de este capítulo es exponer la metodología de investigación de esta tesis. Dado que el objetivo de la presente tesis es comparar dos modelos de clasificación, para presentar la metodología se muestra la fuente de datos de la cual se obtiene la información para el análisis. Luego, se caracteriza la muestra de análisis. Posteriormente, se resumen las técnicas de análisis de los datos para lograr el objetivo del proyecto.

3.1 Fuente de datos

Para el proyecto se utilizó la base de datos denominada *Pima Indians Diabetes Database* (PIDD) (1990), cuya propiedad original pertenece al National Institute of Diabetes and Digestive and Kidney Diseases de los Estados Unidos de América (EUA), la población para este estudio era la población india Pima cerca de Phoenix, Arizona. El conjunto de datos fue levantado en 1990. Estos datos fueron obtenidos de la University of California in Irving (UCI) Machine Learning Repository – Pima Indians Diabetes Data Set (datos disponibles en: https://www.kaggle.com/uciml/pima-indians-diabetes-database). Las unidades de análisis consistieron en 768 mujeres pertenecientes a la etnia Pima y con al menos 21 años de edad. En ellas fueron registradas 9 variables: 8 numéricas y una clasificatoria, detalladas a continuación:

1. Concentración de glucosa plasmática a las 2 horas de una prueba de tolerancia oral a la glucosa (G120 mg/dl).
2. Concentración de insulina sérica a las 2 horas de una prueba de tolerancia oral a la glucosa (I120 mU/ml).
3. Presión arterial diastólica (PAD mmHg).
4. Grosor del pliegue de la piel del tríceps (GPPT mm).
5. Índice de masa corporal (IMC= peso /altura al cuadrado= kg/m^2).
6. Antecedentes Familiares o función de pedigrí de diabetes (FPD).
7. Número de embarazos.
8. Edad (En años).

9. Variable clasificatoria (D: 0 – 1, donde 1 es interpretado como “test positivo para diabetes”). El diagnóstico estuvo basado en el criterio de la OMS (i.e.: G120 ≥ 200 mg/dl en cualquier examen o evaluación de rutina médica).

Se elige esta base de datos debido a que la etnia posee una alta prevalencia de diabetes tanto en las poblaciones Pima residentes en México o en los EUA. Se obtuvieron datos completos de tolerancia a la glucosa para el 77.2% de los Pima mexicanos elegibles (Schulz et al., 2006). Entre los indígenas Pima mexicanos, el 5.6% de los hombres y el 8.5% de las mujeres tenían diabetes, prevalencias significativamente menores que las de los Pima estadounidenses, de los cuales el 34.2% de los hombres y el 40.8% de las mujeres tenían la enfermedad (Schulz et al., 2006). De los Pima mexicanos, 141 (58%) eran de herencia Pima completa, de los cuales el 5.6% de los hombres y el 7.1% de las mujeres tenían diabetes (Schulz et al., 2006). Entre los Pima de los Estados Unidos de América, 622 (70%) eran de plena herencia Pima, de los cuales el 34.3% de los hombres y el 46.8% de las mujeres tenían diabetes. La prevalencia era aún menor en los mexicanos no Pima, de los cuales ninguno de los hombres y el 5% de las mujeres se veían afectados, pero las diferencias entre los mexicanos Pima y los no Pima no eran estadísticamente significativas (Schulz et al., 2006).

La prevalencia ajustada por edad y sexo en los Pima de los EUA fue 5.5 veces mayor que la de los Pima mexicanos y 16 veces mayor que la de los mexicanos no Pima. Entre los Pima de México, aunque la prevalencia fue 2.8 veces mayor que la de los mexicanos no Pima, no fue estadísticamente significativa. La deficiencia en la tolerancia a la glucosa (DTG) estaba presente en el 6.5% de los hombres Pima mexicanos y el 6.0% de las mujeres, en el 4.4% de los hombres no Pima y el 12.9% de las mujeres, y en el 9.9% de los hombres Pima de EE.UU. y el 12.4% de las mujeres. No hubo diferencia en la prevalencia de DTG entre los no Pima y Pima en México, pero la IGT fue más prevalente en los Pima de EE.UU., que en los dos grupos mexicanos. La prevalencia combinada de DGT y la diabetes no difirió significativamente entre los no Pima (11.4%) y Pima mexicanos (13.4%), pero fue

mayor entre los Pima de EE.UU. (49.4%) (Schulz *et al.,* 2006). Por lo anteriormente descrito se puede concluir que hay una prevalencia importante de diabetes en los Pima, y que a su vez es un motivo que despierta el interés de hacer investigación al respecto.

Se ha encontrado que la etnia Pima tiene la mayor prevalencia de diabetes tipo 2 de cualquier población en el mundo. En un estudio etnográfico, Smith-Morris propone que, entre los Pima, la diabetes es más que un fenómeno biomédico, propone que la solución de la diabetes entre los Pima, que tiene profundas raíces biológicas, históricas y culturales, requiere nuevas soluciones de base comunitaria que prioricen, pero no exploten las perspectivas de las mujeres (Welch, 2009).

Por último, otro motivo importante del por qué se elige esta base de datos es la disponibilidad de la información, siendo ésta una limitante para quienes se preparan para obtener el título de licenciatura, debido a que para acceder a bases de datos más recientes, con mayor calidad de recopilación de datos es necesario tener un grado académico para acceder a ellas, es decir, pertenecer a alguna institución que se dedique a la investigación y firmar una solicitud, así como demostrar que en efecto se posee el grado académico. La base de datos *Pima Indian Diabetes* se encuentra disponible en varios repositorios de Internet y puede ser descargada por cualquier persona, lo que representa una ventaja, el fácil acceso a esta.

3.2 Muestra de análisis

Se analizan los 768 casos disponibles en la base de datos (1990), sin embargo, solo se analizan 6 de las 8 variables del conjunto de datos, excluyendo las variables discretas que son el número de embarazos y la edad, recuperando la propuesta de Tarrés y colaboradores (2016), donde se analizan únicamente las variables continuas.

Una vez que se ha expuesto la base de datos con la cual se llevan a cabo los análisis de la tesis, en el siguiente apartado se presentan las técnicas de estimación mediante las cuales se desarrolla el objetivo principal de la tesis: comparar los algoritmos

regresión logística y perceptrón multicapa tomando como criterio de comparación la precisión con la que logran clasificar los casos que padecen diabetes. Entonces, en un primer subapartado se expone parte de los fundamentos del aprendizaje automático dado que este es el principio de las técnicas de estimación que se comparan. Luego se expone la regresión logística que es el primer método de clasificación a contrastar, posteriormente se resumen las bases del perceptrón. Finalmente, se presentan los métodos de contraste entre la regresión logística y el perceptrón.

3.3 El análisis discriminante y sus mejoras

A mediados del siglo XX se empezaron a hacer estudios en distintas áreas del conocimiento que requerían de algoritmos más sofisticados dada la dificultad para resolver los problemas de clasificación, tal es el caso del detector de quiebras de Altman (Hernandez-Ramírez, 2014), el procesador de imágenes de Rosenblatt (Rosenblatt, 1958), entre otros. Sin embargo, se vieron superados por problemas más complejos, ya que todos los algoritmos mencionados previamente solo hacen separaciones o agrupaciones de manera lineal. Es así que surgen nuevos algoritmos que son modificaciones de éstos o que están inspirados en ellos, tal es el caso del análisis discriminante cuadrático (ADC o QDA del inglés Quadratic Discriminant Analysis) y el análisis discriminante regularizado (ADR o RDA por el inglés de Regularized Discriminant Analysis).

A propósito, se sabe que el rendimiento del análisis discriminante lineal y cuadrático depende del tamaño del conjunto de entrenamiento, de la cercanía de la población, de que los datos sigan una distribución normal multivariante y de si las covarianzas en las clases consideradas son iguales o no, lo cual se verifica con la prueba M de Box para la homogeneidad de las matrices de covarianza. Dicha prueba es clave para determinar si el algoritmo es apto para los datos y si se consigue un rendimiento significativo estadísticamente.

Una de las áreas del conocimiento que requirieron de algoritmos más sofisticados fue la química, donde se aplicó las modificaciones del análisis discriminante para determinar el tipo de excipiente químico que se tenía a partir de datos de la región espectral del infrarrojo cercano. El conjunto de datos al que se aplica se divide en 7 subgrupos de excipientes. Los dos primeros y el cuarto conjunto de datos fueron especialmente diseñados con la intención de hacer que la clasificación sea un reto, los demás datos proceden de la industria y, por lo tanto, son representativos de lo que cabría esperar en la práctica. Una vez ejecutados los tres algoritmos la mejor precisión para clasificar la obtiene el ADR. Es poco probable que el ADC produzca resultados satisfactorios, a menos que la proporción del tamaño de las muestras de las clases sea representativa del número de variables planteadas en el problema, y no tiene ninguna ventaja en comparación con el análisis discriminante lineal ADL, excepto cuando las matrices de covarianza de clases son bastante diferentes (Tharwat, 2016).

En situaciones en las que las matrices de covarianza de clases son similares, es probable que proporcione mejores resultados ADL en comparación con el ADC, porque se puede emplear una sola matriz de covarianza en la regla de clasificación, con lo que se reduce el número de estimaciones que deben computarse. Por lo tanto, el ADL debería poder producir mejores resultados, cuando los tamaños de las muestras son más pequeños. El ADR siempre da resultados, equivalentes o mejores que el ADL y el ADC. En algunos casos se reduce automáticamente al ADL y a veces al ADC (Wu et al., 1996), sin embargo, en los escenarios en la vida real generalmente no sucede esto, y por eso es que un algoritmo como el ADC es una herramienta más valiosa.

Otra rama donde se requieren algoritmos de clasificación es en la medicina, empleando una regresión logística binaria aplicada para construir el mejor modelo para datos de hepatitis usando la mejor regresión de subconjuntos y procedimientos escalonados y dependiendo de algunas pruebas de laboratorio como el glutamato-oxalato-transaminasa, glutamato-piruvato-transaminasa, fosfatasa-alcalina y bilirrubina sérica total que representan las variables explicativas o independientes.

Además, la técnica se ha utilizado para clasificar a las personas en dos grupos que están infectados y no infectados con hepatitis en una muestra aleatoria de 200 personas de las cuales 86 son personas no infectadas y 114 están infectadas (Abdulqader, 2017). El modelo logístico ajustado fue fiable usando únicamente 3 variables explicativas, resolviendo el problema de agrupación con una precisión del 98% (Abdulqader, 2017).

En el diagnóstico clínico de enfermedades la forma de medir la efectividad del diagnóstico es mediante la curva ROC[21] la cuál es una métrica que emplea cálculos similares a los que se hacen en el campo del aprendizaje automático, sin embargo la curva ROC es una métrica más profunda ya que evalúa más que la simple clasificación correcta de los casos por parte del diagnóstico.

3.4 El aprendizaje automático

El aprendizaje automático está dominando las investigaciones sobre cómo la analítica avanzada emergente puede proporcionar a las empresas una ventaja competitiva para el negocio. No hay debate sobre el hecho de que los líderes empresariales actuales se enfrentan a competidores nuevos e imprevistos. Estas empresas están buscando nuevas estrategias que puedan prepararlas para el futuro. Mientras que un negocio puede probar diferentes estrategias, todas ellas vuelven a una verdad fundamental que hay que seguir los datos (Langley y Carbonell, 1984).

Si bien el aprendizaje automático existe desde hace mucho tiempo, el enfoque del mundo de los negocios en él puede parecer un desarrollo de la noche a la mañana. La tecnología ha estado creciendo constantemente. El aprendizaje automático representa una evolución clave en los campos de la informática y el análisis de datos, se fundamenta en la base de que las máquinas deben ser capaces de aprender y

[21] La curva ROC (receiver operating characteristic curve) constituye un método estadístico para determinar la exactitud diagnóstica de ciertas pruebas de diagnóstico de enfermedades, siendo utilizada con dos propósitos específicos: determinar el punto de corte de una escala continua en el que se alcanza la sensibilidad y especificidad más alta y evaluar la capacidad discriminativa del test diagnóstico.

adaptarse a través de la experiencia. Los crecientes volúmenes de datos diversos, los avances en el procesamiento informático y el almacenamiento de datos asequibles han contribuido a que en los últimos cinco años haya resurgido el interés por el aprendizaje automático. Las empresas, grandes y pequeñas, de todo el mundo están comenzando a utilizar el aprendizaje automático para transformar procesos críticos, en particular el procesamiento del lenguaje natural, la clasificación y extracción de textos, el análisis emocional/conductual y el reconocimiento de imágenes (Attaran y Deb, 2018).

El aprendizaje automático se ha convertido en uno de los temas más importantes dentro de las organizaciones de desarrollo que están buscando formas de aprovechar las bases de datos para ayudar a la empresa a ganar un nuevo nivel de entendimiento. Con los modelos apropiados de aprendizaje automático, las organizaciones tienen la capacidad de predecir continuamente los cambios en el negocio para que sean los más capaces de predecir lo que sigue. Como los datos son constantemente generados, los modelos de aprendizaje automático se siguen entrenando, así aseguran que la solución es constantemente actualizada. El aprendizaje automático utiliza una variedad de algoritmos que iterativamente aprenden de los datos para mejorar, describir los datos y predecir los resultados; a medida que los algoritmos ingieren datos de entrenamiento, es posible producir modelos más precisos basados en esos datos (Langley y Carbonell, 1984).

El aprendizaje automático, por tanto, se basa en el uso de datos. Esos datos de entrenamiento se utilizan para obtener conclusiones más o menos generales partiendo de casos particulares. A diferencia del razonamiento deductivo, que va de lo general a lo particular, el razonamiento inductivo usado en el aprendizaje automático va de lo particular a lo general: un tránsito de las cosas individuales a los conceptos universales en palabras del propio Aristóteles. El razonamiento deductivo establece reglas que permiten establecer de forma precisa cuándo un razonamiento es válido: partiendo de las premisas, la conclusión se infiere necesariamente. Sin

embargo, en el razonamiento inductivo no existe ningún tipo de garantía que nos indique cuándo un argumento es válido (Berzal, 2018).

El filósofo inglés Francis Bacon, en su Novum Organum de 1620, identificó la inducción como la viga maestra sobre la que se asienta el método científico. En su Tratado sobre la Naturaleza Humana de 1739, el filósofo escocés David Hume formalizó la inducción como un proceso esencial para la adquisición de conocimiento en el mundo real, relegando la deducción al mundo de las ideas y las relaciones abstractas. Posteriormente, en sus Principios de la Ciencia de 1874, sería el lógico y economista inglés William Stanley Jevons el primero que interpretaría la inducción como la aplicación inversa de la deducción (Berzal, 2018).

3.4.1 Tipos de aprendizaje automático

Una primera clasificación de las técnicas de aprendizaje automático puede realizarse atendiendo a la filosofía utilizada en el proceso de adquisición del conocimiento. Estas técnicas se dividen en dos ramas: aprendizaje automático supervisado y aprendizaje automático no supervisado.

3.4.2 Aprendizaje automático supervisado

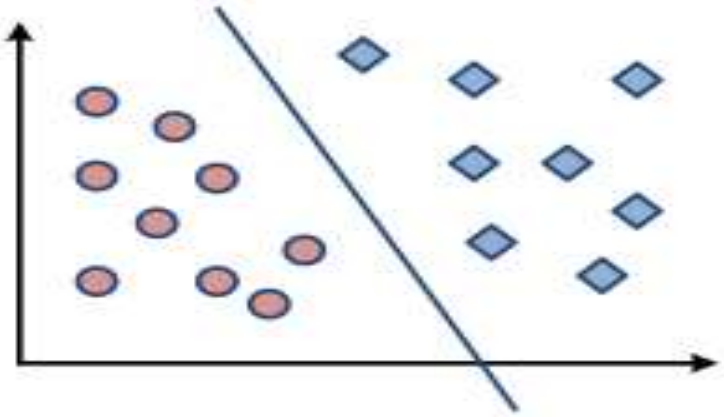
En el aprendizaje supervisado (o aprendizaje a partir de ejemplos, con profesor), los ejemplos de entrenamiento van acompañados de la salida correcta que el sistema debería ser capaz de reproducir. El entrenamiento de un modelo de aprendizaje supervisado consiste en ajustar sus parámetros para que sea capaz de reproducir una salida lo más parecida posible a la deseada. Una vez entrenado el modelo, lo verdaderamente importante es que sea capaz de generalizar correctamente. Esa capacidad de generalización consiste en que el modelo proporcione salidas adecuadas para datos de entrada diferentes a los datos utilizados durante su entrenamiento (Berzal, 2018).

El aprendizaje automático supervisado es habitualmente asociado a los problemas de clasificación, es uno de los problemas más estudiados en Inteligencia Artificial. En un

problema de clasificación, el objetivo de cualquier algoritmo de aprendizaje supervisado es construir un modelo de clasificación a partir de un conjunto de datos de entrada, denominado conjunto de entrenamiento, que contiene algunos ejemplos de cada una de las clases que se pretende modelar. Los casos del conjunto de entrenamiento incluyen, además de la clase a la que corresponde cada uno de ellos, una serie de atributos o características que se utilizarán para construir un modelo abstracto de clasificación (Berzal, 2018).

El objetivo del aprendizaje supervisado es la obtención de una descripción precisa para cada clase utilizando para ello los atributos incluidos en el conjunto de entrenamiento. El modelo que se obtiene durante el proceso de aprendizaje puede utilizarse para clasificar nuevos ejemplos o, simplemente, para comprender mejor los datos de los que disponemos, si nuestro modelo de clasificación es interpretable (Berzal, 2018). En la figura 3.1 se presenta un ejemplo de un algoritmo que se entrena para separar dos clases de manera lineal.

Figura 3.1 Ejemplo de clasificador de aprendizaje supervisado



Fuente: Berzal (2018).

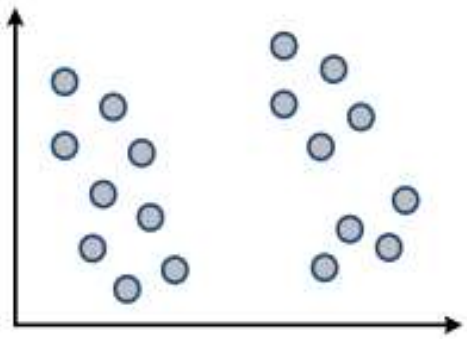
3.4.3 Aprendizaje automático no supervisado

En el aprendizaje no supervisado se construyen descripciones, hipótesis o teorías a partir de un conjunto de hechos u observaciones, sin que exista información adicional acerca de cómo deberían clasificarse los ejemplos del conjunto de entrenamiento.

Será el método de aprendizaje no supervisado el que decida cómo han de agruparse los datos del conjunto de entrenamiento o qué tipo de patrones son más interesantes dentro del conjunto de entrenamiento (Berzal, 2018).

En las técnicas de aprendizaje no supervisado, no existen clases predefinidas (véase figura 3.2). Nuestro objetivo será encontrar patrones en los datos de entrada $x \in X$ que nos permitan construir un modelo de la distribución de probabilidad p(X), el aprendizaje no supervisado nos puede servir como herramienta de análisis exploratorio de datos y para preprocesar los datos antes de utilizar una técnica supervisada (Friedman et al., 2009).

Figura 3.2 Modelo de aprendizaje no supervisado



Fuente: (Berzal, 2018).

Cuando se examina un nuevo conjunto de datos, pero se desconoce cuáles son sus características, las técnicas de aprendizaje no supervisado pueden servir para analizar su estructura. Con ayuda de técnicas no supervisadas, se puede descubrir qué patrones se repiten, cómo se agrupan los datos de forma natural y si existen algunas anomalías, tales como los puntos atípicos (Berzal, 2018).

En marketing, las técnicas no supervisadas pueden ayudar a resolver problemas de segmentación de clientes. En recuperación de información, pueden servir para clasificar documentos sin tener que definir clases de antemano. En el análisis de web

logs, la identificación de patrones de acceso similares puede servir para construir perfiles de los usuarios que acceden a un sitio web.

3.5 Métricas de precisión para algoritmos de aprendizaje automático

Sea cual sea el tipo de técnica de aprendizaje que se decida emplear para resolver los problemas que se enfrentan, todas ellas tienen algo en común, el resultado del proceso de aprendizaje será un modelo, y ese modelo, como todos los modelos, no es más que una simplificación de la realidad, y al ser una simplificación, estará sujeto a errores. Por tanto, se debe evaluar la calidad del modelo obtenido para determinar si realmente puede servir en la práctica.

3.5.1 La matriz de confusión

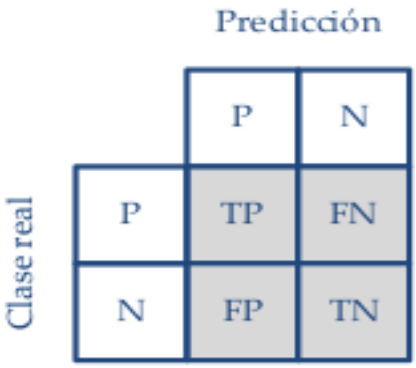
Partiendo del caso más habitual, un problema de clasificación binaria, se desea construir un modelo de clasificación que permita discriminar entre dos clases diferentes. A la primera de ellas, que corresponde a los ejemplos que se desea ser capaz de identificar, se denominará clase positiva (P). Los demás ejemplos se asignan a la clase negativa (N). Cuando se aplica un modelo de clasificación a un conjunto de datos previamente etiquetado, se puede ver fácilmente cómo etiqueta los diferentes ejemplos el clasificador. Sólo existen cuatro posibilidades, que se representan en una matriz de contingencia (véase la figura 3.3), también llamada matriz de confusión (Berzal, 2018).

a) Verdaderos positivos (TP: True Positive): Los ejemplos de la clase positiva que el clasificador es capaz de clasificar correctamente.
b) Falsos positivos (FP: False Positive): Los ejemplos que, aun no siendo de la clase positiva, el clasificador predice que sí lo son.
c) Falsos negativos (FN: False Negative): Los errores que comete el clasificador en sentido contrario, indicando que no son de la clase positiva cuando en realidad no lo son.
d) Verdaderos negativos (TN: True Negative): Los ejemplos de la clase negativa que nuestro el clasifica correctamente.

Los cuatro casos posibles se recogen en una sencilla matriz de confusión de tamaño 2 × 2 en el caso de los problemas de clasificación binaria. Si se enfrenta a un problema de clasificación con múltiples clases, la matriz de confusión tendría una dimensión de K × K, donde K es el número de clases del problema.

Analizando las filas y columnas de esa matriz se puede apreciar dónde acierta y dónde se equivoca el modelo de clasificación, lo que puede indicar posibles problemas y sugerir qué es lo que se debe intentar cambiar. Ahora bien, aparte de poder analizar con detalle cómo trabaja el modelo de clasificación los ejemplos de las diferentes clases, es necesario disponer de medidas numéricas que ayudan a resumir el contenido de la matriz de contingencia y comparar modelos de clasificación alternativos (Berzal, 2018).

Figura 3.3 Matriz de confusión para un problema de clasificación binaria



Fuente: (Berzal, 2018).

3.5.2 La precisión y el error

La métrica más utilizada para resumir el rendimiento de un modelo de aprendizaje supervisado es su precisión. La precisión indica la proporción de ejemplos que un clasificador es capaz de clasificar correctamente, indicada habitualmente en forma de tanto por ciento. En la ecuación 3.1 se muestra cómo se calcula la precisión con la que se compararán los algoritmos, en el numerador se tienen los casos positivos (TP)

que el modelo clasifica como positivos, así como los casos negativos que el modelo clasifica como negativos (TN), dividido entre el total de casos.

3.5.3 Otras medidas de un modelo de aprendizaje automático

Cuando se entrena un modelo de aprendizaje automático usualmente se busca conseguir modelos de aprendizaje lo más precisos posible, la precisión de un modelo no siempre será el único criterio que se utiliza para evaluar la calidad de los modelos construidos.

$$Precisión = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.1)$$

Fuente: Berzal (2018).

Cobertura, robustez, estabilidad, interpretabilidad o, incluso, serendipia[22] pueden ser relevantes en el problema particular que pretendamos resolver (Berzal, 2018).

a) Cobertura: la cobertura de un modelo hace referencia a su capacidad de aplicación sobre los datos que se puedan presentar. En ocasiones, un modelo puede que sea incapaz de tomar una decisión que ofrezca un mínimo de garantías, por lo que no proporcionará una respuesta que se pueda utilizar en esos casos. Si lo que se pretende hacer es automatizar un proceso que se venía realizando de forma manual, la cobertura nos indica el grado de automatización que se ha alcanzado con la ayuda de las técnicas de aprendizaje automático. Por otra parte, si lo que se hacer es diseñar un sistema de recomendación, la cobertura indica cuántos ítems es capaz el sistema de incluir en sus recomendaciones.
b) Robustez: dado que en la práctica los datos con los que ha de trabajar un modelo siempre contienen errores, vienen acompañados de ruido y pueden incluir valores nulos, es importante que la técnica de aprendizaje que se utilice

[22] Serendipia es el término empleado cuando se encuentra algo sin buscarlo o por accidente.

para construir un modelo sea capaz de funcionar correctamente en situaciones como las que se encontrará en el mundo real.

c) Estabilidad: La estabilidad de los modelos puede resultar fundamental en sistemas que interactúan con seres humanos. Por ejemplo, las recomendaciones ofrecidas por un sistema o los resultados de un sistema de recuperación de información no deberían verse demasiado afectados porque un atacante pretenda modificarlos. En ocasiones, existen incentivos económicos que pueden empujar a determinadas personas a intentar manipular el funcionamiento normal de un sistema construido con la ayuda de técnicas de aprendizaje automático, por lo que es otra faceta del sistema que se debe tener en cuenta.

   La interpretabilidad de un modelo puede ser otro aspecto relevante en la práctica, especialmente si el puesto de trabajo que se ocupa depende de la misma después de que el sistema haya cometido un error, aparentemente, garrafal. En ciertas aplicaciones, dará igual si el sistema se comporta como una caja negra, siempre y cuando ofrezca resultados aceptables en cuanto a su precisión, tasa de cobertura, robustez o escalabilidad. Sin embargo, en determinados ámbitos, la interpretabilidad del modelo puede ser un requisito para que se pueda utilizarlo en la práctica, puede que incluso por obligaciones legales.

d) Eficiencia: desde un punto de vista práctico, la eficiencia de los algoritmos necesarios para construir un modelo es vital para que una solución pueda ser factible. El tipo de técnica de aprendizaje suele determinar el tiempo necesario, en primer lugar, para construir el modelo a partir de un conjunto de datos y, posteriormente, para utilizar el modelo en la práctica. Si la construcción del modelo es algo que sólo se necesita hacer una vez, tal vez sea posible emplear una técnica de aprendizaje que resulte especialmente costosa. Si el sistema debe tomar decisiones en tiempo real una vez puesto en marcha, tal vez se imposibilite el uso de ciertos métodos de aprendizaje cuyos requisitos técnicos impiden su funcionamiento en tiempo real utilizando el hardware que se tenga disponible.

e) Escalabilidad: relacionada con la eficiencia, la escalabilidad de una técnica concreta de aprendizaje la cual determina si es posible aplicarla sobre los conjuntos de datos enormes con los que se trabaja en big data[23]. Aunque el coste computacional asociado a un algoritmo de aprendizaje sea lineal con respecto al tamaño del conjunto de datos de entrenamiento; este algoritmo puede que no sea viable en la práctica si no es posible paralelizarlo de alguna forma (Berzal, 2018).

3.6 Métodos de evaluación

A la hora de evaluar el comportamiento de un sistema, independientemente de las métricas concretas que se hayan seleccionado para hacerlo, se debe seguir ciertas directrices. Al construir un modelo, se intenta comprender algo que nunca se puede conocer del todo. No hay más remedio que estimar cómo de bueno es el modelo que se está construyendo. Para realizar esa estimación, se recurre a técnicas de tipo estadístico. En aprendizaje automático, el modelo se construye a partir de un conjunto de datos, ese conjunto de datos es todo lo que se tiene para estimar la bondad del modelo y, dado que no se dispone de otros datos, ese conjunto de datos se tendrá que utilizar para, además de construir el modelo, estimar su calidad (Berzal, 2018).

El objetivo es que esa estimación sirva para evaluar, en media, cuál será el comportamiento del modelo una vez que se utilice. Para ello se reserva una parte de los datos disponibles, en vez de utilizar todos los datos disponibles para entrenar el modelo, una parte no la utilizaremos en el entrenamiento, esta parte será la que sirva para estimar cómo se comportará el modelo con datos diferentes a los datos con los que se ha construido. Es el llamado conjunto de prueba. Utilizar un conjunto de prueba separado del conjunto de entrenamiento puede servir para realizar una estimación puntual de forma rápida y sencilla (Beasley y Rodgers, 2009).

[23] Se denomina *big data* al procesamiento de datos en gran cantidad, es decir datos cuya extensión generalmente supera los 6 gigabytes.

Algunas técnicas estadísticas pueden ayudar a hacer una idea de cuánto pueden variar las estimaciones puntuales obtenidas en cada experimento, si es que se hacen varios experimentos. La media de esas estimaciones puede servir para determinar cómo se comportará el modelo construido con conjuntos de datos distintos a los de su entrenamiento. La varianza o la desviación de las estimaciones puntuales derivadas de cada experimento servirá para descubrir si el rendimiento del sistema será siempre similar (varianza reducida) o si esperamos que se produzcan fluctuaciones (varianza elevada) (Beasley y Rodgers, 2009).

Las técnicas estadísticas de remuestreo[24] serán las que permitan realizar una estimación de algo inicialmente desconocido: la precisión de muestras medidas. Esas técnicas seleccionan subconjuntos de datos (jackknife[25]) del conjunto de datos disponible para realizar una batería de experimentos que proporcionará estimaciones de medias y desviaciones para las métricas que se estén empleando para evaluar la calidad de los modelos (Beasley y Rodgers, 2009).

3.6.1 Muestreo sin reemplazo: validación cruzada

La estimación del error mediante un conjunto de prueba independiente se puede hacer más fiable si se repite el proceso con diferentes conjuntos de prueba. En cada prueba, se puede seleccionar aleatoriamente el conjunto de datos de entrenamiento y evaluar las tasas de error, promediando los resultados obtenidos en cada experimento, los diferentes conjuntos de entrenamiento y de prueba se solaparon en los distintos experimentos. Además, siempre existiría la posibilidad de que algunos ejemplos no se usasen nunca como parte del conjunto de entrenamiento de un modelo. Para solventar ambas limitaciones, lo habitual es recurrir a la validación cruzada. La validación cruzada crea y evalúa múltiples modelos de forma sistemática, cada uno de ellos sobre diferentes subconjuntos del conjunto de datos disponible. La validación

[24] El remuestreo es un enfoque estadístico que se basa en el análisis empírico, basado en los datos observados. El objetivo del remuestreo es tomar una decisión inferencial, que es el mismo objetivo que el de una prueba estadística paramétrica como el análisis convencional de la varianza (ANOVA). La diferencia está en cómo se logra el objetivo (Beasley y Rodgers, 2009).

[25] Jackknife es una técnica especialmente útil para corregir el sesgo de estimación.

cruzada de k iteraciones o k-CV (k-fold Cross-Validation[26]) se realiza de la siguiente manera:

a) Se divide aleatoriamente el conjunto de datos en k subconjuntos o pliegues de igual dimensión o tamaño (Friedman et al., 2009). Generalmente se hace la división del conjunto en 10 pliegues. ¿Por qué 10? Oficialmente, porque se han realizado numerosos experimentos que muestran que la mejor forma de obtener una buena estimación es una validación cruzada estratificada con k=10, aun cuando se disponga de capacidad de cálculo para aumentar el número de pliegues (Berzal, 2018).
b) Posteriormente se seleccionan aleatoriamente los pliegues o subconjuntos en proporción de 70% para entrenamiento y 30% para prueba (Friedman et al., 2009).

Para afinar aún más en las estimaciones conseguidas por validación cruzada, se puede ejecutar la validación cruzada de forma repetida, es decir, repetir 10 veces una validación cruzada. Los modelos construidos, 10 por cada una de las 10 validaciones cruzadas, servirán para obtener una estimación más precisa de la calidad del modelo.

La validación cruzada, obviamente, sólo proporcionará estimaciones fiables si el conjunto de datos es representativo para el problema que estemos resolviendo. Además, los distintos subconjuntos deben seleccionarse para ser también representativos. Si en el problema que se está estudiando, el sistema evoluciona con el tiempo, el conjunto de datos de prueba utilizado podría incluir sesgos que afecten a la calidad de las estimaciones (Efron y Tibshirani, 1993).

3.6.2 Muestreo con reemplazo: Bootstrapping

La validación cruzada utiliza técnicas de muestreo sin reemplazo: el mismo ejemplo, una vez seleccionado, no puede volver a utilizarse. Gracias a ello, no aparecen ejemplos duplicados en el conjunto de entrenamiento y en el de prueba. El

[26] Término en inglés para referirse a la validación cruzada de k pliegues.

bootstrapping es una técnica alternativa que recurre a técnicas de muestreo con reemplazo; esto es, una vez que se escoge un ejemplo, se vuelve a dejar en el conjunto de datos y puede que se vuelva a escoger posteriormente (Efron y Tibshirani, 1993).

El método bootstrap fue propuesto por Bradley Efron en 1979. Se muestrea un conjunto de datos con n ejemplos para formar un nuevo conjunto de datos de n ejemplos. Al realizar el muestreo con reemplazo, algunos de los ejemplos aparecerán más de una vez en la muestra seleccionada. Esa muestra, del mismo tamaño que el conjunto de datos original, es la que se utiliza para entrenar un modelo. El modelo se evalúa recurriendo a los ejemplos del conjunto de datos original que no han sido incluidos en el conjunto de datos de entrenamiento (Efron y Tibshirani, 1993).

Este método también se conoce como 0.632-bootstrap. Cuando se escoge una muestra, se escoge siempre con probabilidad 1/n. La probabilidad de que una muestra no sea escogida será, obviamente, 1 − 1/n. Dado que se construye un conjunto de entrenamiento con n muestras, el proceso hay que repetirlo n veces. Por tanto, la probabilidad de que una muestra no sea escogida para formar parte del conjunto de entrenamiento será $(1 - 1/n)^n \approx e^{-1} = 0.368$. En otras palabras, el 36.8 % de las muestras no se escogerá nunca y pasará el conjunto de prueba. El 63.2 % restante formará parte del conjunto de entrenamiento, de ahí el nombre (Efron y Tibshirani, 1993).

3.6.3 Curva ROC

Una curva ROC es un gráfico bidimensional que ilustra lo bien que funciona un sistema clasificador cuando el valor de corte de la discriminación se cambia en el rango de la variable predictora. El eje x o variable independiente es la tasa de falsos positivos para la prueba de predicción. El eje y o variable dependiente es la verdadera tasa positiva para la prueba de predicción. Cada punto en el espacio ROC es un par de datos positivo verdadero/falso positivo para un valor de corte de discriminación de la prueba predictiva. Si se conocen las distribuciones de probabilidad del verdadero y falso

positivo, se puede trazar una curva ROC a partir de la función de distribución acumulativa. En la mayoría de las aplicaciones reales, una muestra de datos dará un único punto en el espacio ROC para cada elección de corte de discriminación. Un resultado perfecto sería el punto (0, 1) que indica 0% de falsos positivos y 100% de verdaderos positivos (Yang y Berdine, 2017).

3.7 La regresión logística

En la regresión logística, hay una respuesta (binaria) de interés, y se utilizan variables predictivas para modelar la probabilidad de esa respuesta. En términos más generales, en un cuadro de recuentos, el interés primario suele centrarse en un factor que constituye una variable de respuesta (dependiente). Los demás factores de la tabla sólo son de interés por su capacidad de ayudar a explicar la variable de respuesta. Se han desarrollado tipos especiales de modelos para manejar estas situaciones. En particular, en lugar de modelar el registro, se espera que los recuentos de células o el registro probabilidades (como en los modelos log-lineales), cuando hay una variable de respuesta, se modelan varias probabilidades de registro relacionadas con la variable de respuesta (Casella et al., 2006).

Los modelos de regresión logística suelen ajustarse por máxima verosimilitud, usando la probabilidad condicional de *Y* dada *X*. Como *Pr(Y|X=x)* especifica completamente la distribución condicional, la distribución multinomial es apropiada. La probabilidad de esa clase era o bien *p, si yi = 1, o 1 - p, si yi = 0*. La verosimilitud es entonces, la ecuación 3.2.

$$L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i)^{1-y_i} \tag{3.2}$$

Fuente: (Friedman et al., 2009)

Donde: la verosimilitud (L) es igual al producto desde que i es igual con uno hasta n de la probabilidad de que suceda el evento x_i, elevado al exponente y_i que es el resultado de nuestra variable dependiente, es decir, toma valores cero o uno.

El caso especial en el que la variable de respuesta tiene sólo dos categorías es de particular interés y se presta a un tratamiento especialmente agradable. Esto es porque, con sólo dos categorías, hay esencialmente sólo una manera de definir las probabilidades.

Si p_1 es la probabilidad en la primera categoría y p_2 es la probabilidad en la segunda categoría, luego las probabilidades de obtener la categoría uno son p_1 / p_2. Las probabilidades de conseguir la categoría dos son p_2 / p_1. El punto importante es que cualquiera de estos números, junto con el hecho de que $p_1 + p_2 = 1$, determinan completamente tanto p_1 como p_2, por lo tanto, con dos categorías, las dos opciones para las probabilidades conducen a los mismos resultados (Casella *et al.,* 2006; Friedman *et al.,* 2009).

La transformación logit toma un número *p* entre 0 y 1, lo transforma en $log[p / (1 - p)]$. La transformación logística toma un número *x* en la línea real y lo transforma en $e^x /(1 + e^x)$. En otras palabras, la transformación logística aplicada a $log[p/(1 - p)]$ da *p* y la transformación logit aplicada a $e^x /(1 + e^x)$ da *x*. Hacer el análisis de los datos requiere ambas transformaciones. Es en gran medida una cuestión de preferencia personal en cuanto a qué nombre se asocia con el modelo. La situación cuando hay más de dos categorías en la respuesta variable es considerablemente más complicada porque no está claro qué conjuntos de probabilidades para modelar (Casella et al., 2006).

Para solucionar las ecuaciones y obtener el valor de los parámetros beta de la regresión se emplea un método de optimización, que es muy parecido a los métodos de optimización de la familia Newton llamado mínimos cuadrados reponderados iterativamente, el cual se muestra en la ecuación 3.3.

$$\begin{aligned}\beta^{\text{new}} &= \beta^{\text{old}} + (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y}-\mathbf{p}) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\left(\mathbf{X}\beta^{\text{old}} + \mathbf{W}^{-1}(\mathbf{y}-\mathbf{p})\right) \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{z}.\end{aligned} \tag{3.3}$$

Fuente: (Friedman et al., 2009)

3.8 Perceptrón multicapa

El Perceptrón multicapa es un conjunto de *k* perceptrones independientes que se combinan para construir una única red neuronal artificial con *k* salidas. Como los distintos perceptrones son completamente independientes, se puede entrenar por separado cada uno de ellos o en paralelo. La estrategia es la habitual para construir clasificadores multicapa a partir de clasificadores binarios y se denomina *1 vs. all.* En este caso, la entrada del Perceptrón multicapa será la misma que para el Perceptrón individual: un vector de características *x*. Su salida será un vector de bits de tamaño *k* en el que sólo un valor estará cercano a 1 y todos los demás serán 0 (Friedman et al., 2009).

Capas del perceptrón

Las capas del Perceptrón pueden clasificarse en tres tipos:

a) Capa de entrada: Constituida por aquellos nodos que introducen los patrones de entrada en la red, que son las variables independientes. En estos nodos no se produce procesamiento.
b) Capas ocultas: Formada por aquellos nodos cuyas entradas provienen de capas anteriores y cuyas salidas pasan a nodos de capas posteriores.
c) Capa de salida: Nodos cuyos valores de salida se corresponden con las salidas de toda la red, para este caso es la predicción que se hace de la variable dependiente.

Las redes multicapa de tipo *feed-forward*[27] en ocasiones se denominan *backpropagation*[28] *networks* porque el algoritmo de entrenamiento que se suele utilizar con ellas está basado en la propagación hacia atrás del error, es decir que los pesos de las sumas ponderadas se calculan a partir de la disminución de la función de error a partir de la búsqueda de mínimos globales.

Las redes multicapa usadas en *Deep Learning*[29] tienen una capa de entrada, múltiples capas ocultas y una capa de salida, sin embargo en problemas de clasificación con una capa oculta es suficiente, ya que las herramientas de *Deep Learning* son usadas para clasificaciones y procesamiento de imágenes. Desde un punto de vista formal, podemos verlas como una función matemática *f* que, a partir de un vector de entrada *x*, obtiene un vector de salida $y = f(x)$. Dado un conjunto de entrenamiento en forma de pares (*x, y*), el objetivo del algoritmo de entrenamiento es ser capaz de aproximar la función f de forma que para cada posible entrada x se obtenga una salida $\hat{y} = f(x)$ lo más similar posible a la observada en el conjunto de entrenamiento.

Si los datos de entrada no incluyesen ruido, nos bastaría con utilizar una red lo suficientemente grande como para que funcionase como una simple tabla de consulta. Sin embargo, estaríamos sobreaprendiendo[30] y la red neuronal no sería capaz de generalizar correctamente. Sería incapaz de proporcionar salidas apropiadas para entradas con las que no se hubiese encontrado en el conjunto de entrenamiento. Para que una red neuronal sea capaz de generalizar correctamente, tendremos que ajustar su capacidad hasta un nivel que resulte adecuado para la función que pretendemos modelar.

---

27 Una red neuronal prealimentada (feed-forward en inglés) es una red neuronal artificial donde las conexiones entre las unidades no forman un ciclo. Estas son diferentes de las redes neuronales recurrentes (Berzal, 2018).

28 La propagación hacia atrás de errores o retro propagación (del inglés backpropagation) es un método de cálculo del gradiente utilizado en algoritmos de aprendizaje supervisado utilizados para entrenar redes neuronales artificiales (Berzal 2018).

29 Término empleado para la rama de redes neuronales con más de una capa oculta.

30 Término para cuando se sobreajusta el algoritmo al conjunto de entrenamiento y por ende el algoritmo no puede generalizar de manera apropiada.

Este ajuste de la capacidad se puede realizar, o bien ajustando los parámetros que determinan la topología de la red (habitualmente denominados hiperparámetros[31], para distinguirlos de los parámetros ajustados durante el entrenamiento de la red) o bien utilizando técnicas de regularización (cualquier técnica que nos permita reducir el error sobre el conjunto de prueba sin aumentar demasiado el error sobre el conjunto de entrenamiento) (Berzal, 2018).

Función de activación

En las redes neuronales artificiales, la función de activación de un nodo define la salida de ese nodo dada una entrada o conjunto de entradas. Esto es similar al comportamiento del Perceptrón lineal en las redes neuronales, y a la regresión logística, ocupando la curva logística para que el valor de la suma ponderada tome valores entre 0 y 1. Sin embargo, sólo las funciones de activación no lineales permiten a estas redes calcular problemas no triviales utilizando sólo un pequeño número de nodos, y estas funciones de activación se denominan no linealidades (Berzal, 2018).

Dado que para entrenar una red neuronal utilizamos el gradiente del error y en el cálculo de dicho gradiente interviene la derivada de la función de activación, se suelen emplear funciones que sean diferenciables. Sin embargo, no es necesario que las funciones sean estrictamente diferenciables en todos los puntos. Suele bastar con que las funciones tengan definidas sus derivadas por la izquierda y por la derecha, aunque en algunos puntos no coinciden. Son habituales funciones en las que su derivada no está definida en algunos puntos, como la función rectificadora de las unidades lineales rectificadas, ReLU.

Entrenamiento de una red neuronal

Desde el punto de vista formal, una red neuronal multicapa es un aproximador universal. En principio, una red de este tipo es capaz de aprender cualquier cosa, siempre que la red sea lo suficientemente grande como para representar las

[31] Se denomina hiperparámetros a aquellos que se obtienen al entrenar un algoritmo de aprendizaje automático.

peculiaridades de la función que se pretende aprender. Si se dispone de un número suficiente de ejemplos de entrenamiento, una red multicapa de la capacidad suficiente será capaz de construir un modelo de los datos con los que se entrena. Sin embargo, no existe una definición formal de lo que resulta “suficiente” (Berzal, 2018).

A lo máximo que se puede aspirar en la práctica es a descubrir una serie de criterios heurísticos que ayuden a tomar decisiones con respecto al diseño de una red neuronal y de su proceso de entrenamiento. Como todas las heurísticas, normalmente suelen ofrecer buenos resultados, pero no garantizan nada desde un punto de vista formal. A la hora de entrenar una red neuronal multicapa, se debe de tener en cuenta múltiples aspectos. Hemos de tomar decisiones con respecto tanto a los parámetros de diseño de la red como a los parámetros del algoritmo de entrenamiento que decidamos utilizar. Colectivamente, estos parámetros reciben el nombre de hiperparámetros para distinguirlos de los propios parámetros de la red, los pesos de sus sinapsis y los sesgos de sus neuronas que se ajustan durante el proceso de entrenamiento de la red (Berzal, 2018).

Aspectos del diseño y entrenamiento de la red neuronal artificial

Los aspectos a considerarse para el entrenamiento de una red neuronal artificial se enumeran a continuación:

a) Topología: en este rubro se incluye el número de capas ocultas, cuántos nodos por capa, así como la función de activación.
b) Optimización: inicialización de los pesos, tasa de aprendizaje, ajuste de los pesos, algoritmo optimizador.
c) Invariabilidad: conseguir robustez en la red frente a cambios comunes en los datos de entrada.
d) Generalización: lograr que la red procese datos distintos a los del proceso de entrenamiento.

Topología

Aunque la mayor parte de las redes neuronales que se utilizan en la práctica son simples redes multicapa, existe una gran variedad de bloques que se pueden utilizar en su construcción. Se pueden escoger capas de salida especializadas para determinados tipos de problemas, como las capas softmax[32] utilizadas para problemas de clasificación. Se pueden seleccionar capas cuyos patrones de conectividad resulten adecuados para determinados tipos de aplicaciones, como las capas convolutivas utilizadas en procesamiento de imágenes. Una vez establecida la arquitectura general de la red, el diseñador debe decidir el número de capas de la red, sus patrones de interconexión y el tamaño de cada una de las capas (esto es, el número de neuronas que forman parte de cada capa) (Berzal, 2018).

Procedimiento de evaluación de desempeño de los algoritmos

En un problema de clasificación binaria se desea construir un modelo de clasificación que permita discriminar entre dos clases diferentes. A la primera de ellas, se denomina clase positiva (P). Los demás ejemplos pertenecen a la clase negativa (N). Cuando se aplica un modelo de clasificación a un conjunto de datos previamente etiquetado, se puede ver fácilmente cómo etiqueta los diferentes ejemplos el modelo clasificador. Sólo existen cuatro posibilidades, que podemos representar en una matriz de contingencia, también llamada matriz de confusión:

- Verdaderos positivos [TP: True Positive]: Los ejemplos de la clase positiva que el clasificador es capaz de clasificar correctamente.
- Falsos positivos [FP: False Positive]: Los ejemplos que, aun no siendo de la clase positiva, el clasificador predice que sí lo son.
- Falsos negativos [FN: False Negative]: Los errores que comete el clasificador en sentido contrario, indicando que no son de la clase positiva cuando en realidad sí lo son.
- Verdaderos negativos [TN: True Negative]: Los ejemplos de la clase negativa que el clasificador clasifica correctamente.

[32] Softmax es una función de activación, pero en este caso solo es aplicado a una capa en específico.

La métrica más utilizada para resumir el rendimiento de un modelo de aprendizaje supervisado[33] es, sin duda, su precisión. La precisión nos indica la proporción de ejemplos que un clasificador es capaz de clasificar correctamente, indicada habitualmente en forma de tanto por ciento: *precisión = (verdaderos positivos + verdaderos negativos) / total de casos.*

Esto es, el total de aciertos dividido entre el total de casos observados en el conjunto de datos. Para los casos que tienen más de dos clases, se generaliza el principio de la matriz de confusión, donde la diagonal principal representa los aciertos. Ingenuamente, se podría pensar que el objetivo será conseguir una tasa de acierto del 100%, lo que equivale a reducir el error al cero absoluto. Sin embargo, generalmente el objetivo es algo mucho más modesto. Nunca se conseguirá eliminar por completo el error en los modelos entre otras cosas, porque son modelos o simplificaciones de la realidad (Berzal, 2018).

Tomando en cuenta las definiciones anteriores el objetivo de la presente tesis es comparar la capacidad clasificatoria de la regresión logística y el perceptrón multicapa, empleando la base de datos “Pima Indian Diabetes”, siendo la precisión promedio y los criterios de robustez, estabilidad etc. los que permitan decidir qué modelo es mejor clasificando el padecimiento de la diabetes.

3.9 Comparativa de algoritmos

Se eligen la regresión logística y el perceptrón multicapa porque tienen presencia en los problemas de clasificación dado que son muy capaces en lo que a propiedades clasificatorias refiere. La regresión logística destaca de entre los algoritmos convencionales porque combina la precisión junto con la interpretabilidad, si bien no es el algoritmo que podría brindar la mejor precisión (dado que hace separaciones lineales únicamente), es fácil de interpretar, en muchos casos a parte de buscar

[33] Los modelos de aprendizaje supervisado son aquellos donde existe una variable dependiente cuyos valores se conocen para entrenar el modelo.

clasificar con la mayor precisión posible también se busca obtener información a partir de los coeficientes, sobre todo saber cuáles son las variables que más influyen para determinar en qué categoría podría pertenecer una observación. Por otro lado el perceptrón multicapa es parte de el apogeo de las redes neuronales y uno de los algoritmos que más aplicaciones tienen dada su naturaleza, puede ser usado para la clasificación así como para regresión, en áreas como el procesamiento de imágenes y la prevención de fraudes financieros las redes neuronales artificiales son herramientas capaces, sin embargo carecen de interpretación.

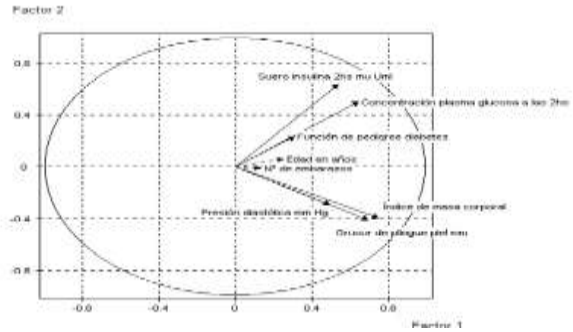
CAPÍTULO 4. RESULTADOS DE LA INVESTIGACIÓN

El objetivo de este capítulo es mostrar la comparación que se realiza entre el perceptrón multicapa y la regresión logística, así como el análisis, selección de los datos y balanceo previo a la aplicación de los algoritmos. Una vez que se tienen los datos preparados se procede a calibrar los algoritmos, obtener sus parámetros y comparar sus propiedades de clasificación. Como se comentó, estos análisis se desarrollan a partir de los *datos Pima Indian Diabetes*, se ha empleado esta base de datos debido a la alta prevalencia de diabetes tipo 2 en dicha etnia, así como el fácil acceso a la base de datos.

4.1 Análisis y selección de las variables independientes

Retomando la propuesta del análisis multidimensional (Tarrés et al., 2016), como variables independientes se seleccionan únicamente las variables continuas, excluyendo a las variables discretas del conjunto de datos, denominándolas como variables ilustrativas, que en este caso son la edad y el número de embarazos, debido a que el análisis de los datos se hace bajo el enfoque de la escuela francesa la cual solo contempla el uso de las variables continuas. En el estudio que se retoma se hizo un análisis de componentes principales cuyos resultados se muestran en la figura 4.1.

Figura 4.1. Variables activas e ilustrativas en los ejes factoriales



Fuente: (Tarrés et al., 2016).

4.2 Análisis descriptivo de las variables seleccionadas

4.2.1 Distribución de las variables

Es importante hacer un análisis descriptivo previo a la ingesta de datos en el modelo, dado que en la estadística frecuentista e incluso en el análisis multivariado es de suma importancia conocer la distribución de los datos, dado que para muchos procedimientos es necesario que los datos sigan una distribución normal. Para probar que los datos sigan una distribución normal se realizó una prueba no paramétrica de Kolmogorov-Smirnov, cuyos resultados arrojan que ninguna de las variables sigue una distribución normal, sumado a lo anterior, el hecho de que las variables estén en diferentes escalas implica tener que nivelar la distribución de los datos de tal manera que se encuentren en la misma escala, por lo que se procederá a estandarizar.

A continuación, se muestra una tabla con medidas de tendencia central de las variables independientes.

Tabla 4.1 Estadísticas descriptivas de las variables seleccionadas

	Media	Desviación	Mínimo	Percentil 25	Percentil 50	Percentil 75	Máximo
Glucosa mg/dl	120.8	31.97	0	99	117	140.25	199
Presión mmHg	72.37	12.1	24	64	72	80	122
GP mm	20.53	15.95	0	0	23	32	99
Insulina mU/ml	79.79	115.24	0	0	30.5	127.25	846
IMC kg/m2	31.99	7.88	0	27.3	32	36.6	67.1
FPD	0.472	0.331	0.078	0.244	0.373	0.626	2.42

Notas. IMC = Índice de Masa Corporal; FPD = Función Pedigrí de Diabetes; GP = Grosor de piel;

Fuente: Elaboración propia con base en el conjunto de datos PID

4.2.2 Análisis gráfico de las variables

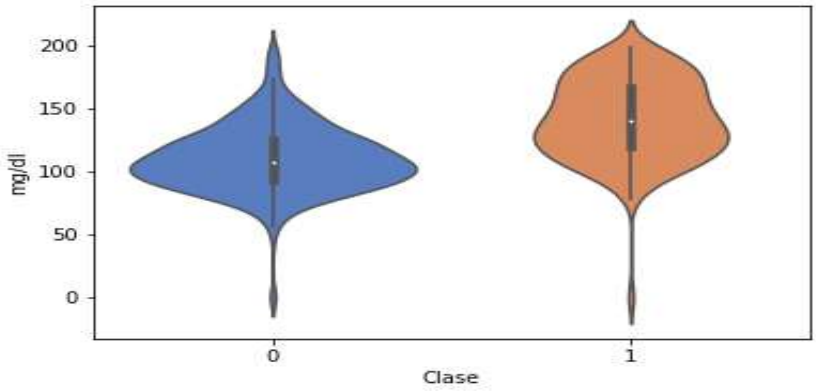
Como se mencionó con anterioridad, es importante conocer la distribución de las variables, para lo cual se emplea una herramienta llamada gráfico de violín[34]. La primera variable que se grafica es la de concentración de glucosa (mg/dl). Se concluye que el valor de cero que contiene el conjunto de datos es un valor faltante, dado que como lo establece Savona-Ventura et al (2010), el nivel más bajo para esta prueba es de 75 mg/dl, esto en condiciones de embarazo, sin embarazo los niveles son más altos a los 75 mg/dl.

En la figura 4.2 se aprecia cómo es que los ceros alteran la distribución de los datos a pesar de no ser significativo. Cabe resaltar que los datos faltantes tienen mayor presencia en los casos donde no se padece diabetes o, visto desde la perspectiva de la rutina, la clase cero. Una característica del gráfico es el rango de las clases: la clase cero, toma valores desde 60mg/dl hasta 175 mg/dl aproximadamente, para la clase uno el rango de los datos es de 80mg/dl hasta 200 mg/dl aproximadamente. Para la clase cero, la mediana de los datos toma un valor de 110 mg/dl aproximadamente, mientras que para la clase uno el valor de la mediana es de aproximadamente 140 mg/dl.

La siguiente variable que se grafica es la insulina (mU/ml), el valor mínimo que se tiene registrado después de dos horas de realizar la prueba es de 16 mU/ml (Crowther, 2004), y esta variable contiene valores iguales a cero que no coinciden con el rango establecido para la prueba, por lo que se consideran valores faltantes. Esta variable es una de las que más valores faltantes registrados tiene, siendo 374 casillas que se reemplazaron con el valor de cero, esto indica que aproximadamente 50% de los datos tienen valor cero.

[34] Un diagrama de violín se utiliza para visualizar la distribución de los datos y su densidad de probabilidad. Este gráfico es una combinación de un diagrama de cajas y bigotes y un diagrama de densidad girado y colocado a cada lado, para mostrar la forma de distribución de los datos.

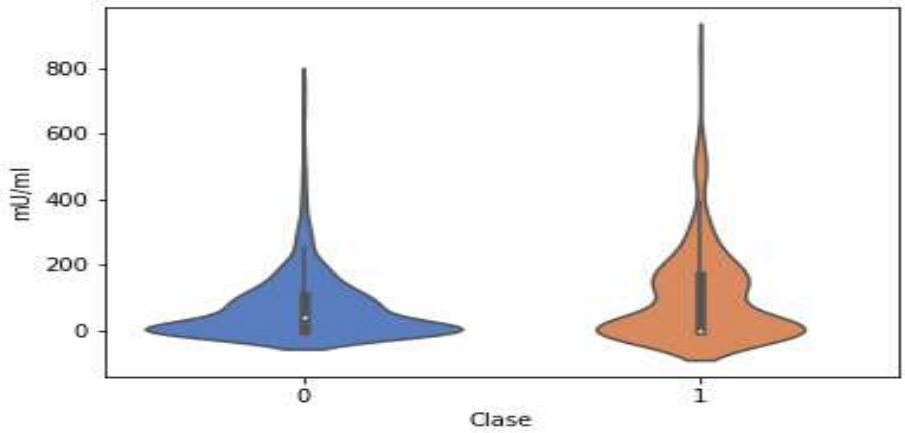
Figura 4.2. Gráfico de violín de la variable glucosa



Fuente: Elaboración propia con base en el conjunto PID.

Entonces, en el gráfico de violín la densidad será más amplia la base, donde se encuentran los valores de cero, es posible apreciar el sesgo que estos valores generan y que por consiguiente no sigue una distribución normal. De los 374 casos con valor de cero, 236 son de la clase cero o los casos que no padecen de diabetes, mientras que 137 son para los casos que padecen de diabetes o que pertenecen a la clase uno (véase la figura 4.3).

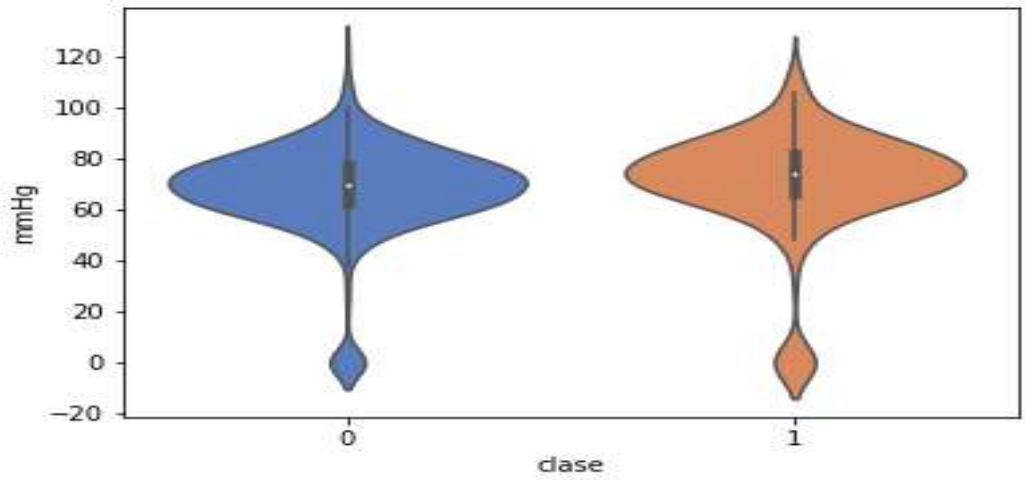
Figura 4.3. Gráfico de violín de la variable insulina



Fuente: Elaboración propia con base en el conjunto PID.

La siguiente variable que se grafica es la presión arterial (mmHg), que como establecen los estándares médicos el rango para la presión normal[35] va de los 80 mmHg hasta los 120 mmHg (Green L, 2003), por lo que se puede concluir que en esta variable los datos están completos. Analizando la gráfica (véase la figura 4.4), los valores registrados oscilan entre 50 mmHg y 100 mmHg para la clase cero o los casos que no padecen diabetes, con una mediana de 70 mmHg aproximadamente, mientras que para la clase uno los valores registrados oscilan entre 50 - 110 mmHg, con una mediana de 75 mmHg, aproximadamente.

Figura 4.4 Gráfico de violín para la presión arterial

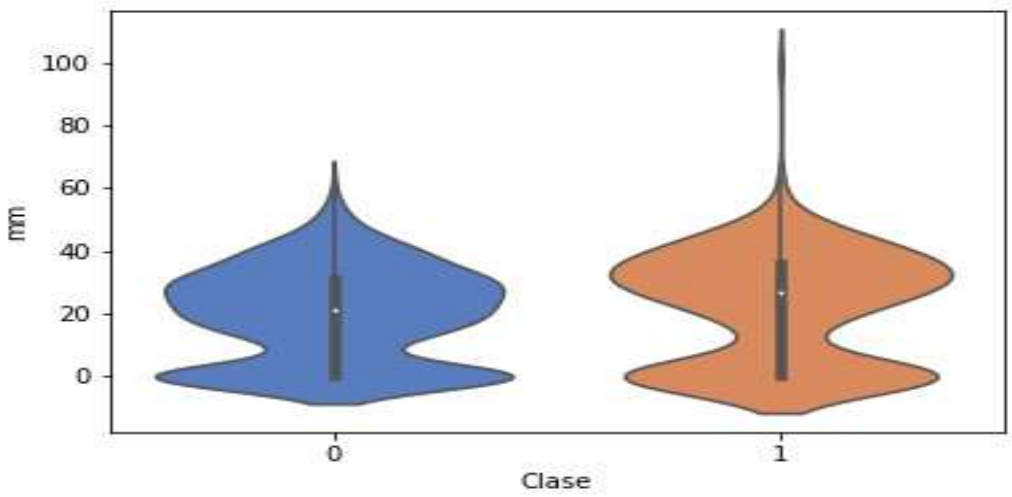


Fuente: Elaboración propia con base en el conjunto PID.

La siguiente variable es el grosor de la piel (GPPT mm), la cual al igual que algunas de las variables previas, contiene 227 valores faltantes, de los cuales 139 pertenecen a la clase cero y 88 a la clase 1, estos valores son los que desvían la gráfica haciendo lucir la densidad como una distribución bimodal, sin embargo, no es así. Para la clase 0 la mediana de grosor del pliegue es de 25 milímetros aproximadamente, mientras que para la clase 1 la mediana es cercana a 30 milímetros (véase la figura 4.5).

[35] Presión normal se refiere al intervalo en el cual debería estar situada una persona que no tiene afecciones cardíacas.

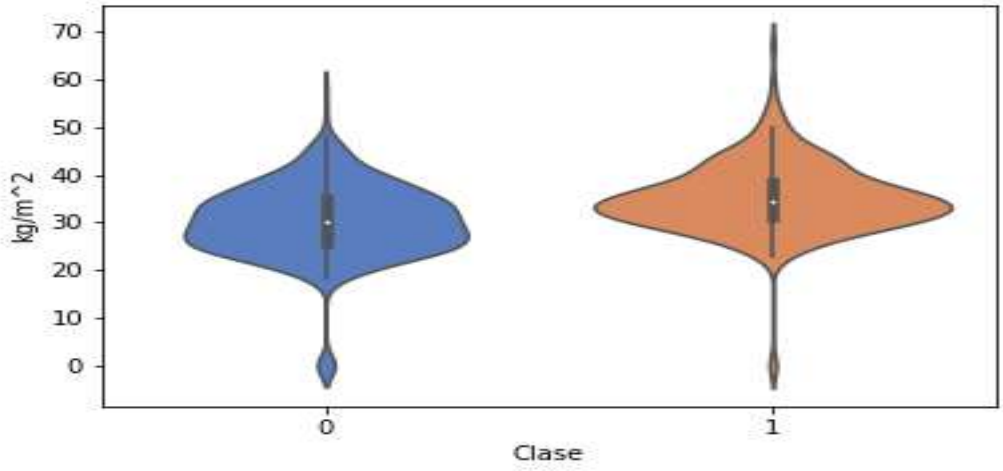
Figura 4.5 Gráfico de violín para el grosor de la piel (GPPT)



Fuente: Elaboración propia con base en el conjunto PID.

La siguiente variable es el índice de masa corporal (IMC), la forma de calcularlo es el cociente del peso en kilogramos y la estatura en metros elevada al cuadrado (peso (kgs) / altura (mts^2)). Entonces, los valores iguales con cero son valores faltantes, debido a que la única forma de que el cociente sea igual a cero se da si el numerador es igual a cero, en este caso que el peso sea igual a cero lo cual es una contradicción. En el conjunto de datos se presentan 11 casillas con el valor de cero, de los cuales 9 son para la clase 0 y 2 para la clase 1, sin embargo, por ser una cantidad poco significativa, la gráfica de la densidad no se ve tan alterada como en casos anteriores donde se asemeja a una distribución bimodal. Para la clase 0 el rango de valores que toma van de las 15 unidades a las 45, mientras que para la clase 1 el rango va desde las 20 unidades hasta las 55 unidades (véase la figura 4.6).

Figura 4.6 Gráfico de violín para el IMC



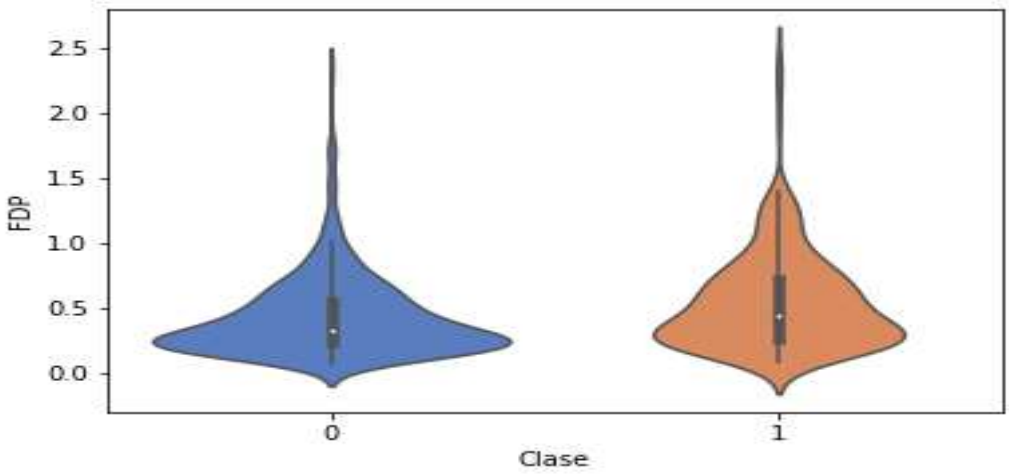
Fuente: Elaboración propia con base en el conjunto PID.

La última variable que se grafica es la función de pedigrí de la diabetes (FPD) la cual fue desarrollada para proporcionar una síntesis de la historia de la diabetes en los parientes y la relación genética de esos parientes con el sujeto. El FDP utiliza la información de los padres, abuelos, hermanos, tíos y primos hermanos. Proporciona una medida de la influencia genética esperada de los parientes afectados y no afectados en el eventual riesgo de diabetes del sujeto (Shanker y Hu, 1999), por lo cual se llega a la conclusión de que esta variable no contiene datos faltantes.

El mínimo valor registrado en esta variable es de 0.078, el rango para la clase cero es de 0.078 unidades hasta 1.75 unidades, mientras que para la clase 1 el rango es de 0.08 unidades hasta 2.42 unidades. Para la clase cero la mediana es de aproximadamente 0.25 unidades, mientras que para la clase uno la mediana es de 0.4 unidades (véase la figura 4.7). De las variables que se analizaron ésta en particular es la que tentativamente podría tener un comportamiento de una distribución sesgada, dado el comportamiento de su gráfica, lo cual a su vez tentativamente explicaría los

puntos que se encuentran en los extremos, ya que en lugar de catalogarlos como puntos atípicos podrían ser característicos de una distribución de cola pesada[36].

Figura 4.7 Gráfico de violín para el IMC



Fuente: Elaboración propia con base en el conjunto PID.

4.3 Imputación de valores faltantes

Las rutinas de los paquetes estadísticos asumen que se trabaja con datos completos e incorporan opciones para imputar observaciones sin que el usuario se perciba de ello. Existe evidencia que la aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio (Aljuaid y Sasi, 2016).

Durante las últimas décadas se han propuesto distintas metodologías para sustituir datos faltantes; sin embargo, es frecuente que estos procedimientos se apliquen sin tener en cuenta sus fundamentos teóricos y sus limitaciones prácticas (Medina y Galván, 2007). El código "." que comúnmente se asocia con información faltante, se debe reservar para situaciones en que no fue posible recabar datos, mientras que el dígito "0" (cero) se preserva para variables que puedan asumir ese valor (Medina y

[36] Las distribuciones con alta probabilidad en las colas comparadas con una normal, se denominan distribuciones de colas pesadas (Mora Valencia, 2011).

Galván, 2007), como lo es el caso del conjunto de datos con el que se trabaja en esta tesis.

Por lo tanto, los valores de cero en las variables de concentración de glucosa plasmática a las 2 horas de una prueba de tolerancia oral a la glucosa, la concentración de insulina sérica a las 2 horas de una prueba de tolerancia oral a la glucosa, el grosor del pliegue de la piel del tríceps y el índice de masa corporal serán reemplazados con la mediana de cada clase, es decir, los ceros que contenga la clase cero serán imputados con la mediana de la misma, dado que las variables no se distribuyen normal. En dado caso de que se distribuyeran normal, sería más apropiado reemplazar los ceros con la media (Aljuaid y Sesi, 2016).

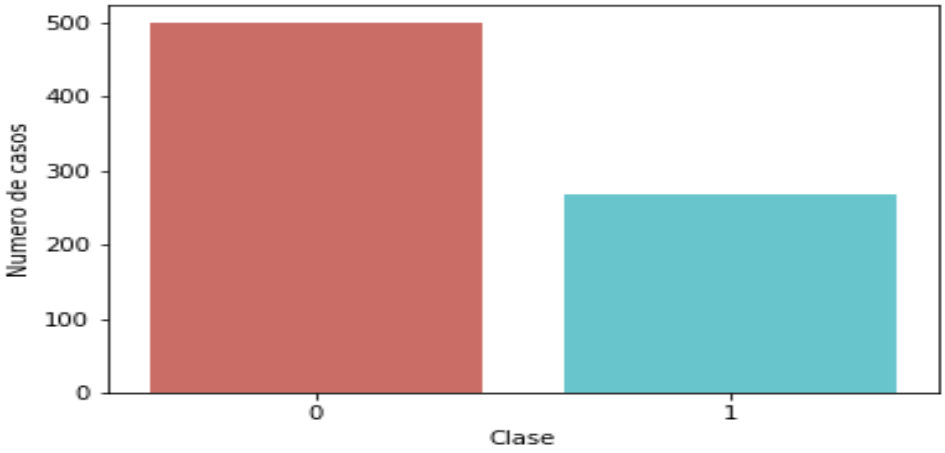
4.4 Balanceo del conjunto de datos

En el aprendizaje supervisado, el problema de la representación desigual de clases, también conocido como el problema de clases desbalanceadas (véase la figura 4.8), se presenta cuando en el conjunto de datos de entrenamiento no hay un número aproximadamente igual de muestras de cada clase. Este problema es particularmente importante en aquellos dominios de aplicación en los que clasificar erróneamente un objeto de la clase minoritaria tiene un costo medio muy elevado, debido a que las ponderaciones de los algoritmos están sesgadas. Durante los últimos años se han realizado muchos esfuerzos para proporcionar soluciones a este problema:

a) Desarrollo de algoritmos de muestreo para cambiar las distribuciones de las clases.
b) Propuestas de técnicas de aprendizaje basadas en costos que castigan la clasificación errónea de los objetos en la clase minoritaria.
c) Desarrollo de técnicas de clasificación de una clase para modelar solo los objetos de clase minoritaria.
d) Desarrollo de técnicas ensambles de clasificadores con las que se busca centrar la atención de cada miembro del ensamble en los objetos de la clase minoritaria.

Los métodos de muestreo más simples son el sobremuestreo y el submuestreo aleatorios.

Figura 4.8 Conjunto de datos desbalanceado



Fuente: Elaboración propia con base en el conjunto PID.

En el primero se seleccionan al azar muestras de la clase minoritaria, las cuales se duplican y se agregan al conjunto de datos original, mientras que en el segundo caso se eliminan muestras al azar escogidas de la clase mayoritaria. Estas estrategias tienen algunas desventajas: el sobremuestreo aleatorio puede llevar a problemas de sobreajuste del clasificador, mientras que el submuestreo aleatorio puede eliminar información importante para la definición de las fronteras de decisión (Mera y Arrieta Ramos, 2015).

Otras estrategias consisten en hacer un muestreo informativo en el que se agregan o eliminan muestras de forma más inteligente. Un método representativo del muestreo informativo es el SMOTE (*Synthetic Minority Over Sampling Technique*), el cual crea datos sintéticos a partir de los segmentos de línea que unen dos muestras de la clase minoritaria (Mera y Arrieta Ramos, 2015). Este último algoritmo será empleado para balancear el conjunto de datos, dado que hay diferencia de casos entre las clases (véase la figura 4.8). Antes de aplicar el muestreo informativo, se hace la división del conjunto entrenamiento/prueba en relación 70/30.

Una vez aplicada la técnica de muestreo informativo, ambas clases cuentan con 354 observaciones cada una (véase la figura 4.9), de tal forma que el conjunto de datos está preparado para el procesamiento en los modelos.

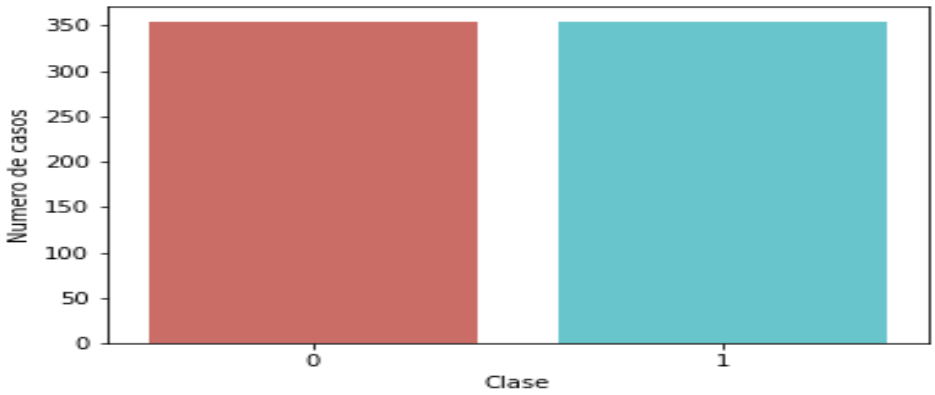
4.5 Estandarización de los datos

Esencialmente, el propósito de la estandarización es hacer que los datos sean más comprensibles para los cálculos computacionales y sus interpretaciones. Algunas razones comunes para estandarizar son:

a) La presencia de un efecto de nivel de dispersión en los lotes de muestras.
b) La distribución de una variable está sesgada.
c) Los residuos de un modelo ajustado exhiben un patrón sistemático.
d) Los datos no satisfacen los supuestos de un procedimiento estadístico.

Así, la principal dificultad que surge en estas situaciones es la presencia de no linealidad que puede aumentar sustancialmente la complejidad del análisis estadístico (Muralidharan, 2010). Antes de la ingesta de los datos al modelo las variables han sido estandarizadas previamente para que como se menciona anteriormente, tengan la misma escala las variables, el estandarizar las variables podría suponer un impedimento en la interpretación, sin embargo, los coeficientes de las ecuaciones son los que podrán ser interpretados posteriormente, a pesar de haber cambiado la escala.

Figura 4.9 Conjunto de datos balanceado



Fuente: Elaboración propia con base en el conjunto PID.

4.6 Obtención de los parámetros de la regresión logística

Una vez que se han logrado los procedimientos previamente descritos a), b), se procede a elegir el optimizador que nos permita maximizar la verosimilitud de la ecuación de la regresión logística, así como las variables de entrada. Para el caso de la regresión logística únicamente se hace la selección del optimizador L-BFGS, dado que las variables de entrada ya estaban definidas desde el principio. Se ocupan las funciones *logitmodel* de la paquetería statsmodels (Seabold y Perktold, 2010) y *LogisticRegression* de la paqueteria Scikit-learn (Li y Phung, 2014) disponibles en el lenguaje python.

4.7 Obtención de los parámetros del perceptrón

De los parámetros descritos en capítulos anteriores, a continuación, se hace mención de los métodos con los cuales se obtuvieron.

4.7.1 Tasa de aprendizaje

El rango de valores a considerar para la tasa de aprendizaje es menor de 1.0 y mayor de 10^-6 de acuerdo a Bengio (2012). La tasa de aprendizaje interactúa con muchos otros aspectos del proceso de optimización, y las interacciones pueden ser no lineales. Sin embargo, en general, las tasas de aprendizaje más pequeñas requerirán más periodos de entrenamiento. Por el contrario, las tasas de aprendizaje más grandes requerirán menos periodos de entrenamiento, cabe mencionar que los tamaños de lote más pequeños se adaptan mejor a las tasas de aprendizaje más pequeñas.

Se pueden utilizar gráficos de diagnóstico para investigar cómo la tasa de aprendizaje afecta el grado de aprendizaje y a la dinámica de aprendizaje del modelo. Un ejemplo es crear un gráfico lineal de las pérdidas durante los periodos de entrenamiento. El gráfico lineal puede mostrar muchas propiedades, por ejemplo:

a) El proceso de aprendizaje durante los periodos de entrenamiento, se comporta de manera rápida o lenta.
b) Si el modelo ha aprendido demasiado rápido (aumento brusco y meseta) o está aprendiendo demasiado despacio (poco o ningún cambio).

c) Si la tasa de aprendizaje puede ser demasiado grande a través de las oscilaciones en la pérdida.
d) La elección del valor de la tasa de aprendizaje puede ser bastante crítica, ya que si es demasiado pequeña la reducción del error será muy lenta, mientras que si es demasiado grande pueden producirse oscilaciones divergentes.

4.7.2 Función de activación

La función de activación es la encargada de que la suma ponderada de las variables de entrada tome valores entre cero y uno, en la paquetería de python que se emplea que es scikit learn tiene como opciones la sigmoide, tangente hiperbólica, ReLU y softmax.

Sigmoide. La función de activación sigmoide es a veces referida como la función logística. Los resultados de la investigación de la función sigmoide han generado tres variantes de la función de activación sigmoide, los cuales pueden ser consultados en el artículo *Activation Functions: Comparison of trends in Practice and Research for Deep Learning.* La función sigmoide se puede expresar en la ecuación 4.1).

$$f(x) = \left(\frac{1}{(1 + exp^{-x})} \right) \tag{4.1}$$

Fuente: Nwankpa et al., (2018).

Tangente hiperbólica. La función de la tangente hiperbólica es otro tipo de función de activación utilizada en el *Deep Learning*. La función tangente hiperbólica (función *tanh*), es una función más suave que la sigmoide, la cual está centrada en cero, y cuyo rango se encuentra entre -1 y 1, por lo que la salida de la función tanh viene dada por la ecuación 4.2

$$f(x) = \left(\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \right) \tag{4.2}$$

Fuente: Nwankpa et al., 2018).

ReLU (Rectified Linear Activation Function). La función de activación de la unidad lineal (ReLU) es una función de activación de aprendizaje más rápido, que ha demostrado ser la función más exitosa y ampliamente utilizada, ofrece el mejor rendimiento y generalización en el aprendizaje profundo en comparación con las funciones de activación del sigmoide y el tanh. La función ReLU representa una función casi lineal y, por lo tanto, conserva las propiedades de los modelos lineales que los hacían fáciles de optimizar (Nwankpa et al., 2018). La función de activación ReLU viene dada por la siguiente función (véase la ecuación 4.3)

$$f(x) = max(0, x) = \begin{cases} x_i, & if\ x_i \geq 0 \\ 0, & if\ x_i < 0 \end{cases} \qquad (4.3)$$

Fuente: Nwankpa et al., (2018).

Softmax. La función softmax se usa en modelos multiclase donde devuelve las probabilidades de cada clase, teniendo la clase objetivo la mayor probabilidad (véase la ecuación 4.4). La función softmax aparece mayormente en casi todas las capas de salida de las arquitecturas de aprendizaje profundo, donde se utilizan. La principal diferencia entre la función de activación sigmoide y softmax es que la sigmoide se utiliza en clasificación binaria, mientras que la Softmax se utiliza para tareas de clasificación multiclase (Nwankpa et al., 2018).

$$f(x_i) = \frac{exp\ (x_i)}{\sum_j exp\ (x_j)} \qquad (4.4)$$

Fuente: Nwankpa et al., (2018).

4.7.3 Optimizadores

Descenso estocástico del gradiente

El descenso estocástico del gradiente (SGD) es un enfoque simple pero muy eficiente para ajustar clasificadores y regresores lineales bajo funciones de pérdida convexa como las máquinas de soporte vectorial (lineales) y la Regresión Logística. Aunque el

SGD ha existido en la comunidad del aprendizaje automático durante mucho tiempo, los optimizadores actuales parten de éste (Li y Phung, 2014).

Algoritmo Broyden-Fletcher-Goldfarb-Shanno de baja memoria

El BFGS de memoria limitada (L-BGFS) es un algoritmo de optimización de la familia de métodos cuasinewton que se aproxima al algoritmo Broyden-Fletcher-Goldfarb-Shanno (BFGS) utilizando una cantidad limitada de memoria de computadora. Es un algoritmo popular para la estimación de parámetros en el aprendizaje automático, el problema objetivo del algoritmo es minimizar *f(x)* sobre los valores no restringidos del vector real *x,* donde *f* es una función escalar diferenciable. Al igual que el algoritmo original BFGS, L-BFGS utiliza una estimación de la matriz inversa del Hessiano para dirigir su búsqueda a través del espacio variable, pero donde el BFGS almacena una densa aproximación al Hessiano inverso, L-BFGS almacena sólo unos pocos vectores que representan la aproximación implícitamente. Debido a su requisito de memoria lineal resultante, el método L-BFGS es particularmente adecuado para problemas de optimización con muchas variables (Ketkar, 2017).

Adam (Adaptative Moment Estimation). Adam es un método de optimización estocástica eficiente que sólo requiere gradientes de primer orden con poca necesidad de memoria. El método calcula las tasas de aprendizaje adaptativo individual para diferentes parámetros de las estimaciones de los primeros y segundos momentos de los gradientes; el nombre Adam se deriva de la estimación del momento de adaptación. El método está diseñado para combinar las ventajas de dos métodos: AdaGrad, que funciona bien con gradientes escasos y RMSProp, que funciona bien en línea y no posee ajustes estacionarios. Algunas de las ventajas de Adam son que las magnitudes de las actualizaciones de los parámetros son invariables en la reescalada del gradiente, sus tamaños escalonados están aproximadamente limitados por el hiperparámetro de tamaño escalonado, no requiere un objetivo estacionario, trabaja con gradientes escasos, y naturalmente realiza una forma de recocido de tamaño escalonado (Kingma y Ba, 2015).

Existen muchos más algoritmos de optimización que pueden ser empleados para la resolución del problema de clasificación, sin embargo, se limita a dar una breve introducción de éstos tres dado que son los que emplea el módulo scikit-learn en la función *MLPClassifier* (Li y Phung, 2014).

4.7.4 Número de nodos en la capa oculta

Sobre cómo determinar el número de nodos en la capa oculta, (Nababan, 2020) sugiere lo siguiente:

a) El número de nodos en la capa oculta debe estar entre el número de nodos de entrada y los nodos de salida.
b) El número de nodos en la capa oculta debe ser de aproximadamente ⅔ partes el tamaño de los nodos de entrada más el tamaño de los nodos de salida.
c) El número de nodos debe ser menor al doble del tamaño de los nodos de entrada.

Por otro lado, Castro y Salgado (2014) proponen una metodología para la determinación del modelo neuronal artificial con mayor parsimonia, es decir, generar varias redes neuronales y tomar la que tenga la mejor combinación entre precisión de clasificación (lo más alta posible y el número de nodos en la capa oculta (lo más bajo posible). En este caso el número óptimo de nodos en la capa oculta debería estar entre 5 y 7 nodos aproximadamente.

4.8 Resultados de la regresión logística

Ya que se definieron los parámetros para la función, se procede a alimentar la función con nuestros datos listos para la ingesta. En el primer entrenamiento se elaboró un resumen, donde se evalúa el primer conjunto de prueba (véase la tabla 4.1) teniendo como resultado que las variables son significativas al 95% excepto la variable de la presión arterial diastólica, la cual resulta ser significativa en 90% únicamente, sin embargo, dado que se retoma la propuesta de Tarrés et al., (2016), se decide mantener la variable dentro del modelo. Posteriormente, se procede a hacer la validación cruzada del conjunto total en 10 pliegues o subconjuntos, obteniendo una

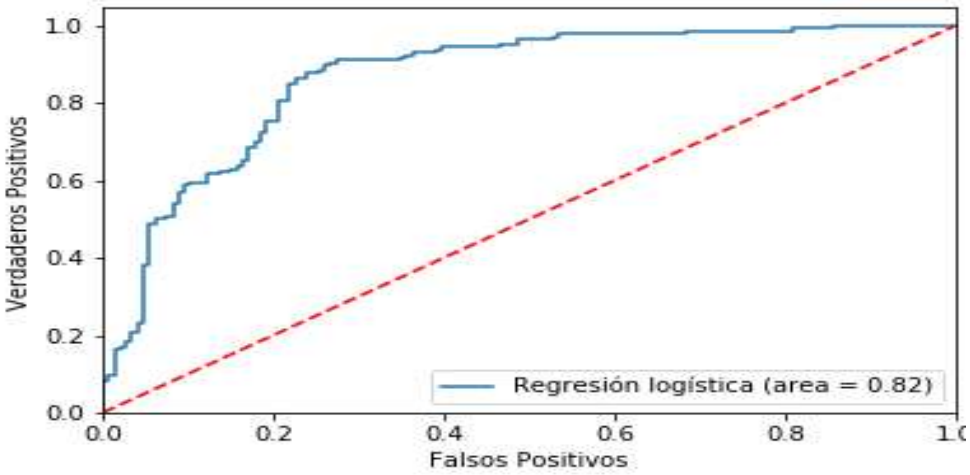
precisión promedio de 76.1%, es decir que clasifica con apropiadamente el 76.1% de los casos del conjunto de datos. La precisión no es el único criterio para establecer que un modelo sea bueno, tal y como se mencionó en el capítulo 3, también se debe evaluar la variación de las estimaciones, para que un modelo sea robusto y tenga la capacidad de generalizar la variación de las estimaciones debe ser mínima, por lo que se procede a calcular la variación de la regresión logística, dando como resultado una variación de 4.8% en cada estimación, lo cual es una cifra relativamente baja. Posteriormente se evalúa la curva ROC de la regresión logística (véase la figura 4.10), obteniendo como resultado un área debajo la curva de 82%, lo que nos indica que el modelo tiene la capacidad de discriminar a los pacientes que padecen y no padecen diabetes en 82% de los posibles cortes a lo largo de la curva. Dado que éste número es superior al 75%, se considera una clasificación buena, dado que el 75% representa una clasificación medianamente buena.

Tabla 4.1 Resumen del modelo de regresión logística inicial

```
         Current function value: 0.480304
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  708
Model:                          Logit   Df Residuals:                      702
Method:                           MLE   Df Model:                            5
Date:                Fri, 20 Nov 2020   Pseudo R-squ.:                  0.3071
Time:                        12:35:32   Log-Likelihood:                -340.06
converged:                       True   LL-Null:                       -490.75
Covariance Type:            nonrobust   LLR p-value:                 5.045e-63
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Glucosa        1.4417      0.129     11.176      0.000       1.189       1.695
Presion       -0.1968      0.114     -1.719      0.086      -0.421       0.028
Grosorp        0.4513      0.126      3.569      0.000       0.203       0.699
Insulina      -0.5380      0.108     -4.989      0.000      -0.749      -0.327
IMC            0.4639      0.129      3.592      0.000       0.211       0.717
Antecedentes   0.4594      0.106      4.315      0.000       0.251       0.668
==============================================================================
```

Fuente: resultados obtenidos a partir de la ejecución propuesta en Python.

Figura 4.10 Curva ROC de la regresión logística



Fuente: Elaboración propia con base en el conjunto PID.

4.9 Resultados del perceptrón multicapa

4.9.1 Tasa de aprendizaje

Considerando los criterios expuestos en el capítulo 4 respecto a la obtención de los parámetros, el primer parámetro que se obtiene es la tasa de aprendizaje del modelo neuronal, cuyo valor para éste tipo de modelos debe oscilar entre 10e-6 y 1, por lo que en la rutina se ejecuta un ciclo *for* con el cuál se evalúa en la función tasas de aprendizaje que van desde 0.001 hasta 1, dicho criterio puede ser modificado fácilmente, ya que como criterio inicial se toma 1/1000 por cuestiones de tiempo de cómputo. Como resultado la tasa de aprendizaje que arroja el mejor resultado es 0.466.

4.9.2 Función de activación

Como función de activación se seleccionó la ReLU, dado que es la más recomendable para problemas como éste, de acuerdo a lo que mencionan los autores citados previamente en el capítulo 4.

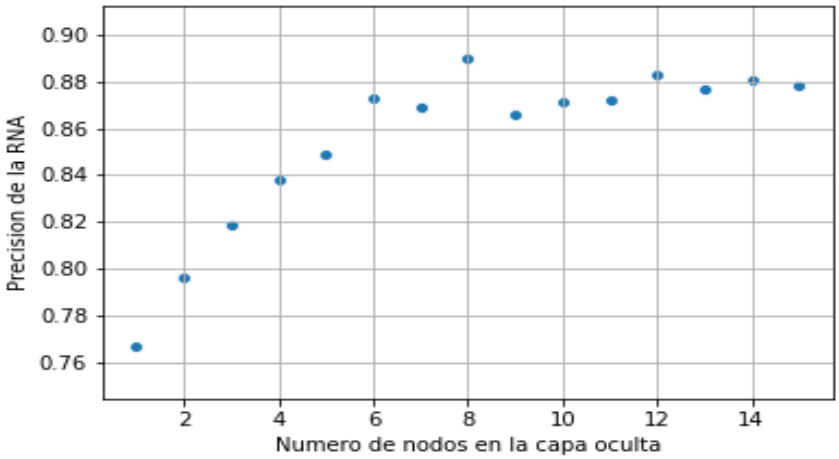
4.9.3 Optimizador

Como optimizador se eligió el Adam, dado que es de los optimizadores que mejor desempeño tiene de entre las opciones disponibles en la función.

4.9.4 Nodos en la capa oculta

Evaluando las propuestas expuestas en el capítulo 4, el número de nodos se seleccionó con un ciclo *for* evaluando perceptrones con 1 nodo en la capa oculta hasta 15 nodos en la capa oculta (véase la figura 4.11), siendo el MLP(6-6-2) el mejor modelo encontrado. Una vez definidos todos los parámetros del perceptrón (6-6-2), se procedió a evaluar la precisión del modelo con el conjunto de datos, obteniendo una precisión promedio de 86.8%, con una variación de 3.62% de las estimaciones de la validación cruzada.

Figura 4.11 Selección del modelo neuronal con mayor parsimonia

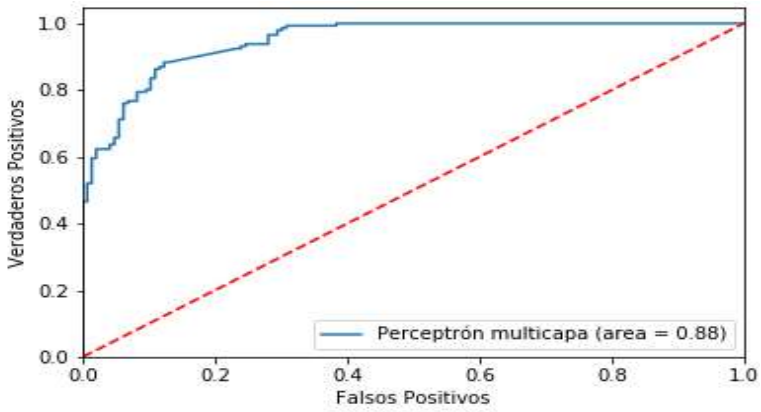


Fuente: Elaboración propia con base en el conjunto PID.

4.9.5 Resultados para la Curva ROC

Aplicando la curva ROC al perceptrón multicapa (véase la figura 4.12), se obtuvo un área bajo la curva de 88%, lo cual significa que el modelo clasifica correctamente a quienes padecen diabetes y quienes no la padecen con un 88% de probabilidad, en los estándares de la curva es una clasificación catalogada como buena/muy buena.

Figura 4.12 Curva ROC del modelo neuronal propuesto



Fuente: Elaboración propia con base en el conjunto PID.

CAPÍTULO 5. CONCLUSIONES

Es importante aplicar varios algoritmos de aprendizaje automático dado que los modelos se adaptan a los datos y no al revés. Hay algoritmos que se pueden presumir como superiores debido a que son más sofisticados que otros, sin embargo, en la práctica siempre hay que buscar la parsimonia, por lo que lo mejor es explorar con distintos algoritmos y elegir el que tenga mejor desempeño.

La comparación de ambos algoritmos se llevó a cabo empleando el conjunto de datos *Pima Indian Diabetes*, los algoritmos fueron evaluados acorde a la precisión y capacidad clasificatoria que permitiera diferenciar entre pacientes que padecen diabetes tipo 2 y pacientes sanos. Con el conjunto de datos preparado para ser introducido a ambos modelos, evaluando bajo las mismas métricas, por lo que estando en equilibrio como el que se establece, es apropiada la comparación realizada.

Analizando los resultados la regresión logística obtuvo un 76% de acierto, lo cual es un resultado moderado aceptable, ese 76% resulta de la matriz de confusión, que indica el número de clasificaciones correctas entre el número total de clasificaciones. El perceptrón que obtuvo un 86.8% de precisión bajo la misma métrica de la matriz de confusión. Una diferencia de 10.8% en términos del conjunto de datos original son entre 78 y 80 diagnósticos acertados por encima de la regresión logística, en términos relativos 14% mejor.

Posteriormente se hace el análisis de la curva ROC, medida que en el ramo de la medicina es más importante que la matriz de confusión como tal dado que el cálculo de ésta tiene como base los criterios de sensibilidad, que se calcula con el cociente de los casos verdaderos positivos y los casos que padecen la enfermedad, y la especificidad que es el cociente de los verdaderos negativos y los casos que no padecen la enfermedad. El gráfico de la curva ROC tiene como ordenadas la sensibilidad y como abscisas la tasa de 1 – especificidad, lo cual representa la tasa de los casos positivos sobre la tasa de negativos.

Como resultado de la curva ROC, el área bajo la curva de la regresión logística es de 82%, lo que significa que el modelo de regresión logística tiene un 82% de probabilidad de identificar si una persona padece diabetes o no de forma correcta, mientras que el área bajo la curva del perceptrón es de 88%, es decir que el perceptrón tiene un 88% de probabilidad de identificar si una persona tiene diabetes o no de forma correcta.

Después de la ejecución, los resultados de la comparativa evidencian la superioridad de clasificación del perceptrón sobre la regresión logística. Este aspecto es de suma importancia ya que la diferencia de la precisión entre ambos algoritmos se puede traducir en situaciones como:

- Diagnósticos certeros en el caso de enfermedades como el caso del análisis desarrollado para la diabetes.

Una vez discutidas las ventajas que tiene el perceptrón, también se discuten las desventajas, que a pesar de parecer “superficiales”, dependiendo de la situación puede ser que no sea tan superior a la regresión logística, que a pesar de ser un tanto obsoleta sigue vigente. Al ser un algoritmo sofisticado el perceptrón, es también exigente, en los siguientes sentidos:

a) Tiempo de computo: comparado con las 7 iteraciones promedio en las que converge una regresión logística, está claramente en desventaja el algoritmo, considerando que la rutina que se ejecuta del perceptrón dura aproximadamente un minuto, incluyendo la obtención de parámetros para el perceptrón que es la parte que requiere más tiempo de procesamiento. Quizá el tiempo no sea de tanta relevancia, después de todo vale la pena esperar un poco esa precisión extra.
b) Infraestructura: el caso que se resuelve el presente trabajo es relativamente sencillo, ya que son poco más de 700 observaciones aproximadamente, sin embargo, en un conjunto de datos grande es necesario contar con un ordenador capaz de procesar toda esa información.

La diferencia entre las precisiones de los algoritmos aplicados en contextos distintos de la diabetes puede resumirse en los siguientes puntos:

- Otorgar o no un crédito a un cliente moroso, que tiene que ver con el área de riesgo de crédito de una institución financiera.
- Análisis de la productividad de la empresa u organización que requiere de análisis como los realizados.
- Selección de candidatos a ingresar a un programa de estudios, contemplando sus antecedentes como calificaciones y capacitaciones previas al examen.

A propósito, se comparte la programación realizada en esta tesis con la cual sus resultados son replicables y pueden servir como base para que otras personas puedan desarrollarlo.

Futuras líneas de investigación

El campo del aprendizaje automático está experimentando un crecimiento exponencial hoy en día, especialmente en el tema de la visión por computadora. Hoy en día, la tasa de error es sólo de 3% para las computadoras. Esto significa que las computadoras son superiores a los humanos en reconocimiento de imágenes, cuyo error ronda el 8% aproximadamente.

Aplicación de la visión computarizada

Una aplicación por excelencia para la visión por computadora es la retinopatía diabética, que es una complicación de la diabetes se la que se habla en los primeros capítulos. Para diagnosticarla, se requiere un examen ocular extenso. En los países del tercer mundo y las aldeas rurales donde hay escasez de médicos, un modelo de aprendizaje automático que utilice la visión por ordenador para hacer un diagnóstico será extremadamente beneficioso. Al igual que con todos los campos de imágenes médicas, esta visión computarizada también puede ser una segunda opinión para los expertos en dominio, asegurando la credibilidad de su diagnóstico. Generalmente, el propósito de la visión computarizada en el campo médico es replicar la experiencia de los especialistas y desplegarla en los lugares donde las personas más lo necesitan.

Procesamiento natural del lenguaje

El modelo BERT, el último modelo de procesamiento de lenguaje natural (NLP) que Google anunció se ha puesto en uso en sus algoritmos de clasificación de búsqueda. Esto ayudó a mejorar los resultados de búsqueda para innumerables tipos diferentes de consultas que antes eran muy difíciles. En otras palabras, el sistema de búsqueda ahora puede entender mejor diferentes tipos de búsquedas realizadas por los usuarios y ayudar a proporcionar respuestas mejores y más precisas. Estos modelos basados en Transformer para la traducción están mostrando ganancias espectaculares en la puntuación BLEU, que es una medida de la calidad de la traducción. Por lo tanto, las arquitecturas de Machine Learning que utilizan transformadores como BERT están aumentando en popularidad y funcionalidad.

REFERENCIAS

Abdulqader, Q. M. (2017). Applying the Binary Logistic Regression Analysis on The Medical Data. *Science Journal of University of Zakho*, *5*(4), 330. https://doi.org/10.25271/2017.5.4.388

Aljuaid, T., y Sasi, S. (2016). *Proper imputation techniques for missing values in data sets*. https://doi.org/10.1109/ICDSE.2016.7823957

American Diabetes Association. (2020). Introduction: Standards of Medical Care in Diabetes 2020. *Diabetes Care*, *43*(Supplement 1the), S1 LP-S2. https://doi.org/10.2337/dc20-Sint

Andersson, T., Ahlbom, A., y Carlsson, S. (2015). Diabetes Prevalence in Sweden at Present and Projections for Year 2050. *PLOS ONE*, *10*, e0143084. https://doi.org/10.1371/journal.pone.0143084

Attaran, M., y Deb, P. (2018). *Machine Learning: The New "Big Thing" for Competitive Advantage*. *5*, 277–305. https://doi.org/10.1504/IJKEDM.2018.10015621

Baan, C. A., Bonneux, L., Ruwaard, D., y Feskens, E. J. M. (1998). The prevalence of diabetes mellitus in the Netherlands. A quantitative review. *European Journal of Public Health*, *8*(3), 210–216. https://doi.org/10.1093/eurpub/8.3.210

Basto-Abreu, A., Barrientos-Gutiérrez, T., Rojas-Martínez, R., Aguilar-Salinas, C. A., López-Olmedo, N., De la Cruz-Góngora, V., Rivera-Dommarco, J., Shamah-Levy, T., Romero-Martínez, M., Barquera, S., López-Ridaura, R., Hernández-Ávila, M., y Villalpando, S. (2020). Prevalencia de diabetes y descontrol glucemico en Mexico: Resultados de la Ensanut 2016. *Salud Publica de Mexico*, *62*(1), 50–59. https://doi.org/10.21149/10752

Beasley, W., y Rodgers, J. (2009). *Resampling methods*.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7700 LECTU*, 437–478. https://doi.org/10.1007/978-3-642-35289-8-26

Berzal, F. (2018). Redes Neuronales y Deep Learning. *Departamento de Ciencias de La Computacion e IA*, 753.

Casella, G., Fienberg, S., y Olkin, I. (2006). Log-Linear Models and Logistic Regression. In *Design* (Vol. 102). https://doi.org/10.1016/j.peva.2007.06.006

Castro, A. M., y Salgado, O. G. (2014). *Empresas exitosas y no exitosas que cotizan en la BMV del Sector Comercial: Una clasificación con Análisis Discriminante Múltiple, Modelos Logit y Redes …*. 33–62. http://148.206.79.158/handle/11191/4187

Crowther, N. J. (2004). The clinical relevance of fasting serum insulin levels in obese subjects. *South African Medical Journal = Suid-Afrikaanse Tydskrif Vir Geneeskunde*, *94*, 519–520. https://doi.org/10.1080/16089677.2004.11073591

Efron, B., y Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Macmillan Publishers Limited. All rights reserved.

Escobedo-De La Peña, J., Buitrón-Granados, L. V., Ramírez-Martínez, J. C., Chavira-Mejía, R., Schargrodsky, H., y Champagne, B. M. (2011). Diabetes en México. Estudio CARMELA. *Cirugia y Cirujanos*, *79*(5), 424–431.

Forsham, P. (1982). Diabetes mellitus. A rational plan for management. *Postgraduate Medicine*, *71*, 139-144,148. https://doi.org/10.1080/00325481.1982.11716019

Friedman, J., Hastie, T., y Tibshirani, R. (2009). *The Elements of Statistical Learning Data Mining, Interface and Prediction Preface to the Second Edition*. 809. http://gen.lib.rus.ec/book/bibtex.php?md5=0161E6689920ACB72E562A5B8D726F4D

Gerdtham, U.-G., Clarke, P., Hayes, A., y Gudbjornsdottir, S. (2009). Estimating the Cost of Diabetes Mellitus-Related Events from Inpatient Admissions in Sweden Using Administrative Hospitalization Data. *PharmacoEconomics*, *27*, 81–90. https://doi.org/10.2165/00019053-200927010-00008

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hernández-Ávila, M., Gutiérrez, J. P., y Reynoso-Noverón, N. (2013). Diabetes mellitus en México: El estado de la epidemia . *Salud Pública de México* (Vol. 55, pp. s129–s136). scielomx .

Hernandez-Ramírez, M. (2014). Modelo financiero para la detección de quiebras con el uso de análisis discriminante múltiple. *InterSedes: Revista de Las Sedes Regionales*, *XV*(32), 4–19.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science y Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jiménez Navarrete, M. F. (2000). Diabetes mellitus: actualización . In *Acta Médica Costarricense* (Vol. 42, pp. 53–65). scielo .

Joshi, S. (2015). Diabetes Care in India. *Annals of Global Health*, *81*, 830–838. https://doi.org/10.1016/j.aogh.2016.01.002

Ketkar, N. (2017). Deep Learning with Python. In *Deep Learning with Python*. https://doi.org/10.1007/978-1-4842-2766-4

Kingma, D. P., y Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Langley, P., y Carbonell, J. G. (1984). Approaches to machine learning. In *Journal of the American Society for Information Science* (Vol. 35, Issue 5). https://doi.org/10.1002/asi.4630350509

Lemaitre, G., Nogueira, F., y Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*, 1–5. http://jmlr.org/papers/v18/16-365.html

Li, H., y Phung, D. (2014). Journal of Machine Learning Research: Preface. *Journal of Machine Learning Research*, *39*(2014), i–ii.

Luo, Z., Fabre, G., y Rodwin, V. G. (2020). Meeting the challenge of diabetes in China. *International Journal of Health Policy and Management*, *9*(2), 47–52. https://doi.org/10.15171/ijhpm.2019.80

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, *1*(Scipy), 56–61. https://doi.org/10.25080/majora-92bf1922-00a

Medina, F., y Galván, M. (2007). *estudios estadísticos y prospectivos*.

Mera, C., y Arrieta Ramos, J. M. (2015). *Estudio Comparativo de Técnicas de Balanceo de Datos en el Aprendizaje de Múltiples Instancias*.

Michael Waskom and the seaborn development team}. (2020). *waskom2020seaborn* (lastest). Zenodo. https://doi.org/10.5281/zenodo.592845

Nababan, E. (2020). *How to decide the number of hidden layers and nodes in a hidden layer?*

Naik, P., y Oza, K. (2019). *Python with Spyder: An Experiential Learning Perspective*.

Nwankpa, C., Ijomah, W., Gachagan, A., y Marshall, S. (2018). *Activation Functions: Comparison of trends in Practice and Research for Deep Learning*.

O'Connell, J. M., y Manson, S. M. (2019). Understanding the Economic Costs of Diabetes and Prediabetes and What We May Learn About Reducing the Health and Economic Burden of These Conditions. *Diabetes Care*, *42*(9), 1609 LP – 1611. https://doi.org/10.2337/dci19-0017

Organizacion Mundial de la Salud. (2016). *Informe Mundial Sobre la diabetes*.

Papatheodorou, K., Banach, M., Bekiari, E., Rizzo, M., y Edmonds, M. (2018). Complications of Diabetes 2017. *Journal of Diabetes Research*, *2018*, 1–4. https://doi.org/10.1155/2018/3086167

Ramachandran, A. (2005). Epidemiology of diabetes in India—Three decades of research. *The Journal of the Association of Physicians of India*, *53*, 34–38.

Rojas-martínez, R., Basto-abreu, A., Aguilar-salinas, C. A., y Zárate-rojas, E. (2018). *Rojas R, Basto A,Aguilar C. Prevalencia de diabetes por diagnóstico médico previo en México. Salud Pública de México. [revista en Internet] 2018 ; 60(3): 224. Disponible en: https://www.medigraphic.com/cg*. *60*(3).

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Savona-Ventura, C., Craus, J., Vella, K., y Grima, S. (2010). Lowest threshold values for the 75g oral glucose tolerance test in pregnancy. *Malta Medical Journal*, *22*, 18–20.

Schulz, L. O., Bennett, P. H., Ravussin, E., Kidd, J. R., Kidd, K. K., Esparza, J., y Valencia, M. E. (2006). Effects of traditional and western environments on prevalence of type 2 diabetes in Pima Indians in Mexico and the U.S. *Diabetes Care*, *29*(8), 1866–1871. https://doi.org/10.2337/dc06-0138

Seabold, S., y Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*, *Scipy*, 92–96. https://doi.org/10.25080/majora-92bf1922-011

Shanker, M., y Hu, M. Y. (1999). *Estimating Probabilities of Diabetes Mellitus*.

Silink, M. (2009). Problemas globales de salud de impacto local II. The economic and social consequences of type 2 diabetes. *Gac Méd Méx*, *145*(4), 290–294. https://www.anmm.org.mx/GMM/2009/n4/26_vol_145_n4.pdf

Singh, U. (2016). Prevalence of diabetes and other health related problems across India and worldwide: An overview. *Journal of Applied and Natural Science*, *8*(1), 500–505. https://doi.org/10.31018/jans.v8i1.825

Tarrés, M. C., Moscoloni, N., Navone, H., y D'Ottavio, A. E. (2016). Análisis Multidimensional De Una Base De Datos De Mujeres Pima/ Multidimensional Analysis From a Database of Pima Women. *Biotecnia*, *18*(3), 14–19. https://doi.org/10.18633/biotecnia.v18i3.330

Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, *3*(2), 145. https://doi.org/10.1504/ijapr.2016.079050

Toelsie, J., Bipat, R., Algoe, M., y Mans, D. (2013). Diabetes mellitus: historical background, global aspects, and impact in Suriname. *Acad J Sur*, *4*, 365–371.

Urquidez-Romero, R., Esparza-Romero, J., y Valencia, M. E. (2015). Interacción Entre Genética Y Estilo De Vida En El Desarrollo De La Diabetes Mellitus Tipo 2: El Estudio En Los Indios Pima. *BIOtecnia*, *17*(1), 40. https://doi.org/10.18633/bt.v17i1.17

van Rossum, G. (1995). Python tutorial, May 1995. *CWI Report CS-R9526*, *CS-R9526*, 1–65. http://oai.cwi.nl/oai/asset/5007/05007D.pdf

Wahdan, M. (1996). The epidemiological transition. *La Revue de Santé de La Mediterranee Orientale*, *2*.

Welch, J. R. (2009). Diabetes among the pima: stories of survival. *Cadernos de SaÃ\textordmasculinede PÃ\textordmasculineblica*, *25*, 1429–1430. http://www.scielo.br/scielo.php?script=sci_arttextypid=S0102-311X2009000600026ynrm=iso

Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, D. L., Heuerding, S., y Erni, F. (1996). Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data. *Analytica Chimica Acta*, *329*(3), 257–265. https://doi.org/10.1016/0003-2670(96)00142-0

Yang, S., y Berdine, G. (2017). The receiver operating characteristic (ROC) curve. *The Southwest Respiratory and Critical Care Chronicles*, *5*, 34. https://doi.org/10.12746/swrccc.v5i19.391

Yu, Y., Ouyang, X., Lou, Q.-L., Gu, L.-B., Mo, Y.-Z., Ko, G., Chow, F., So, W.-Y., Ma, R., Kong, A., Brown, N., Nan, H., Chan, J., y Bian, R.-W. (2012). Validity of Glycated Hemoglobin in Screening and Diagnosing Type 2 Diabetes Mellitus in Chinese Subjects. *The Korean Journal of Internal Medicine*, *27*, 41–46. https://doi.org/10.3904/kjim.2012.27.1.41

Anexo 1: ejecución de la rutina en python

A continuación, se muestra el script o rutina empleada en esta investigación, la cual emplea distintos módulos. Para los gráficos se emplearon seaborn (Michael Waskom and the seaborn development team, 2020) y matplotlib (Hunter, 2007), para la ingesta de datos se empleó pandas (McKinney, 2010), para los cálculos se emplearon statsmodels (Seabold y Perktold, 2010), scikit-learn (Li y Phung, 2014) y NumPy (Harris et al., 2020), y para el balanceo de los datos se empleó imbalanced-learn (Lemaitre et al., 2017). La ejecución de la rutina se realizó en Spyder (Naik y Oza, 2019). Todos los módulos forman parte del lenguaje de programación python (van Rossum, 1995).

Importación de los módulos empleados

```
#modulos
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.neural_network import MLPClassifier
from imblearn.over_sampling import SMOTE
import numpy as np
from sklearn.linear_model import LogisticRegression
import statsmodels.api as sm
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score

#lectura de los datos
datos = pd.read_csv('diabe.csv')
datos.columns = ['Embarazos','Glucosa','Presion','Grosorp','Insulina',
            'IMC','Antecedentes','Edad','Salida']
datos['Salida'].value_counts()
```

Apartado para la creación de gráficos de violín y de conteo

```
plt.figure()
sb.countplot(x='Salida',data=datos,palette='hls')
plt.xlabel('Clase')
plt.ylabel('Numero de casos')
#plt.savefig('Numerocasos.png')

datos.groupby('Salida').mean()
datos['Glucosa'].describe()
datos['Presion'].describe()
datos['Grosorp'].describe()
datos['Insulina'].describe()
datos['IMC'].describe()
datos['Antecedentes'].describe()
#graficos de violin
plt.figure()
sb.violinplot(x = 'Salida', y = 'Glucosa', data = datos,
        palette = 'muted', split = True)
plt.xlabel('Clase')
plt.ylabel('mg/dl')
#plt.savefig('Glucosa.png')

plt.figure()
sb.violinplot(x = 'Salida', y = 'Presion', data = datos,
        palette = 'muted', split = True)
plt.xlabel('clase')
plt.ylabel('mmHg')
#plt.savefig('Presion.png')

plt.figure()
sb.violinplot(x = 'Salida', y = 'Grosorp', data = datos,
        palette = 'muted', split = True)
plt.xlabel('Clase')
plt.ylabel('mm')
#plt.savefig('Grosorp.png')

plt.figure()
sb.violinplot(x = 'Salida', y = 'Insulina', data = datos,
```

```
        palette = 'muted', split = True)
plt.xlabel('Clase')
plt.ylabel('mU/ml')
#plt.savefig('Insulina.png')

plt.figure()
sb.violinplot(x = 'Salida', y = 'IMC', data = datos,
        palette = 'muted', split = True)
plt.xlabel('Clase')
plt.ylabel('kg/m^2')
#plt.savefig('IMC.png')

plt.figure()
sb.violinplot(x = 'Salida', y = 'Antecedentes', data = datos,
        palette = 'muted', split = True)
plt.xlabel('Clase')
plt.ylabel('FDP')
#plt.savefig('Antecedentes.png')
#reemplazar los valores de cero
#Glucosa, insulina, IMC, Grosorp
marco_1 = datos.loc[datos['Salida'] == 1]
marco_2 = datos.loc[datos['Salida'] == 0]
```

En éste apartado se imputan los valores de cero por la mediana de cada variable, separando por clase

```
marco_1 = marco_1.replace({'Glucosa':0},
                np.median(marco_1['Glucosa']))
marco_1 = marco_1.replace({'Insulina':0},
                np.median(marco_1['Insulina']))
marco_1 = marco_1.replace({'IMC':0},
                np.median(marco_1['IMC']))
marco_1 = marco_1.replace({'Grosorp':0},
                np.median(marco_1['Grosorp']))
marco_2 = marco_2.replace({'Glucosa':0},
                np.median(marco_2['Glucosa']))
marco_2 = marco_2.replace({'Insulina':0},
                np.median(marco_2['Insulina']))
marco_2 = marco_2.replace({'IMC':0},
```

```
                    np.median(marco_2['IMC']))
marco_2 = marco_2.replace({'Grosorp':0},
                    np.median(marco_2['Grosorp']))
datos = pd.concat([marco_1, marco_2])
features = ['Embarazos','Glucosa','Presion','Grosorp','Insulina',
            'IMC','Antecedentes','Edad']
#variables
var_ind = datos.loc[:,features].values
```

En éste apartado se hace la estandarización de las variables así como el balanceo del conjunto de datos

```
#estandarizacion
var_dep = datos.loc[:,['Salida']].values
var_ind_t = StandardScaler().fit_transform(var_ind)
pca = PCA(n_components=var_ind.shape[1])
pca.fit(var_ind_t)
x_pca = pca.transform(var_ind_t)
var_exp = pca.explained_variance_ratio_
#variables seleccionadas
features_selected = ['Glucosa','Presion','Grosorp','Insulina',
            'IMC','Antecedentes']
var_ind_t = pd.DataFrame(var_ind_t)
var_ind_t.columns = features
X = var_ind_t.loc[:,features_selected]
##oversampling
over_samp = SMOTE(random_state=0)
X_train,X_test,y_train,y_test = train_test_split(X,var_dep,test_size = 0.3,
                                random_state = 0)

os_dx,os_dy = over_samp.fit_sample(X_train,y_train)
os_dxp,os_dyp = over_samp.fit_sample(X_test,y_test)
x_comp, y_comp = over_samp.fit_sample(X, var_dep)

plt.figure()
sb.countplot(x = os_dy, palette = 'hls')
plt.ylabel('Numero de casos')
plt.xlabel('Clase')
#plt.savefig('Balance.png')
```

```
#regresion logistica
```

Aquí se entrena la regresión logística con el conjunto de datos, así como la elaboración del gráfico de la curva ROC

```
reg_log = LogisticRegression()
reg_log.fit(os_dx, os_dy)
reg_log.score(os_dxp, os_dyp)
reg_log2 = sm.Logit(exog = os_dx, endog = os_dy).fit()
print(reg_log2.summary())
#curva ROC regresion logistica
rl_rocauc = roc_auc_score(y_true = os_dyp,
                          y_score = reg_log.predict(os_dxp))
fpr, tpr, thresholds = roc_curve(os_dyp,
                        reg_log.predict_proba(os_dxp)[:,1])
plt.figure()
plt.plot(fpr, tpr, label = 'Regresión logística (area = %0.2f)'%
       rl_rocauc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('Falsos Positivos')
plt.ylabel('Verdaderos Positivos ')
plt.legend(loc="lower right")
#plt.savefig('rocrl')
#validacion cruzada para la regresion logistica
```

En esta parte se calcula la robustez y estabilidad de la regresión logística, a través de los resultados de la validación cruzada.

```
val_cru_rl = cross_val_score(estimator = reg_log, X = x_comp,
                             y = y_comp, cv = 10)

prec_prom_rl = np.mean(val_cru_rl)
variacion_rl = np.std(val_cru_rl)
```

A partir de aquí se obtienen los parámetros para el perceptrón, comenzando por el más importante que es la tasa de aprendizaje, para obtenerla se inicia un ciclo for el cuál va a introducir valores a la función que van desde 0.50001 acorde a la variable itera, que nos da el valor de cuantas veces se va a recorrer el ciclo for. El

procedimiento puede comenzar desde cero, sin embargo para ahorrar tiempo de computo se reduce el numero de iteraciones previo a que ya se había obtenido el valor ideal para la función

```
##selección de la tasa de aprendizaje
itera = 500
cont = np.zeros(itera)
precision = np.zeros(itera)
for i in range(itera):
    cont[i] = i
    np.random.seed(0)
    mlp=MLPClassifier(hidden_layer_sizes=5,activation='relu',
                      learning_rate_init = (0.5+(i/10000)) ,max_iter=1700)
    mlp.fit(os_dx,os_dy)
    precision[i] = mlp.score(os_dxp,os_dyp)

lr_opt = pd.DataFrame({'preci':precision})
lr = lr_opt[lr_opt.eq(lr_opt.preci.max()).any(1)]
lear_rate = 0.0001+(744/10000)
##Entrenando la RNA
```

Aquí se crea una función que permite crear un perceptrón y devolver la precisión junto con los resultados de la validación cruzada, teniendo como entradas las variables del número de nodos en la capa oculta, el conjunto de entrenamiento y de prueba, y finalmente el conjunto original.

```
def red_neuronal(num_neu,X_ent,y_ent,X_pru,y_pru,X_comp,y_comp):
    np.random.seed(0)
    mlp=MLPClassifier(hidden_layer_sizes=num_neu,activation='relu',
                      learning_rate_init = lear_rate ,max_iter = 1700)
    mlp.fit(X_ent,y_ent)
    val_cruz = cross_val_score(mlp,X_comp,y_comp,cv = 10)
    return val_cruz
#Precisión de la red
```

En este apartado nuevamente se desarrolla un ciclo for el cuál tiene la finalidad de emplear la función previamente construida, la cual arroja los resultados de 15 perceptrones donde la variable de conteo es el numero de nodos en la capa oculta,

es decir, en la iteración 1 es un perceptrón con un nodo en la capa oculta, y en la iteración 15 es un perceptrón con 15 nodos en la capa oculta.

```
num_iter = 15
precision = {}
prec_prom = np.zeros(num_iter)
variacion = np.zeros(num_iter)
num_neu = np.zeros(num_iter)
for i in range(num_iter):
  print(i)
  precision[i] = red_neuronal(num_neu = i+1,X_ent = os_dx,y_ent = os_dy,
      X_pru = os_dxp,y_pru = os_dyp,X_comp = x_comp,
      y_comp = y_comp)
  prec_prom[i] = np.mean(precision[i])
  variacion[i] = np.std(precision[i])
  num_neu[i] = i+1

plt.figure()
plt.xlabel('Numero de nodos en la capa oculta')
plt.ylabel('Precision de la RNA')
plt.grid()
sb.scatterplot(x = num_neu,y = prec_prom)
#plt.savefig('Precision.png')
#ROC perceptron
```

Aquí se define el modelo ya con los parámetros previamente obtenidos, y se crea el gráfico de la curva ROC.

```
mlp = MLPClassifier(hidden_layer_sizes = 6, activation = 'relu',
              learning_rate_init = lear_rate, max_iter = 1700)
mlp.fit(X = os_dx, y = os_dy)
mlp_rocauc = roc_auc_score(y_true = os_dyp,
                    y_score = mlp.predict(X = os_dxp))
fpr_mlp, tpr_mlp, thresholds = roc_curve(os_dyp,
                  mlp.predict_proba(os_dxp)[:,1])
plt.figure()
plt.plot(fpr_mlp, tpr_mlp, label = 'Perceptrón multicapa (area = %0.2f)'%
      mlp_rocauc)
plt.plot([0, 1], [0, 1],'r--')
```

```
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('Falsos Positivos')
plt.ylabel('Verdaderos Positivos')
plt.legend(loc="lower right")
#plt.savefig('rocmlp')
```