

Tesis

by Victor Moreno

Submission date: 07-Dec-2021 04:27PM (UTC-0600)

Submission ID: 1723718783

File name: Tesis_Maestria.pdf (1.98M)

Word count: 28637

Character count: 157026

69

Universidad de Guadalajara



Centro Universitario de Ciencias Exactas e Ingenierías

División de Tecnologías para la Integración Ciber-Humana.

140

Departamento de innovación Basada en la Información y el Conocimiento.

Maestría en Cómputo Aplicado.

25

Aprendizaje automático y análisis de datos para la prevención de complicaciones de la diabetes.

3

Líneas de investigación: Analítica de datos y aprendizaje automático.

Presenta:

Ing. Víctor Ernesto Moreno González
220344075

Director de tesis:

Dr. Omar Avalos Alvarez

Codirector de tesis:

Mtro. José Luis David Bonilla Carranza

Guadalajara Jalisco. Diciembre de 2021.

AGRADECIMIENTOS.

Este proyecto es resultado del apoyo de muchas personas a quienes agradezco por medio del presente.

Por su guía y confianza durante estos años, a mi director de tesis, el Dr. Omar Avalos Álvarez y a mi codirector, el Mtro. José Luis David Bonilla Carranza.

Al Dr. Marco Antonio Pérez Cisneros por todo el apoyo y por motivarme a seguir creciendo en lo profesional.

Al Dr. Jorge de Jesús Gálvez Rodríguez por toda la atención, por su apoyo y dedicación.

A mis profesores por contribuir en mi desarrollo y formación profesional y por todas sus enseñanzas.

A mis padres que siempre han sido mi principal fuente de inspiración y a mis hermanas que me motivan a con su cariño.

A mi esposa e hijo que son mi motor de vida y a quienes les dedico mi ser.

CONTENIDO

| | |
|--------------------------------------------------------------------------|-----------|
| INTRODUCCIÓN..... | 5 |
| 1. MINERÍA DE DATOS..... | 7 |
| 1.1. Análisis exploratorio de los datos..... | 7 |
| 1.2. Representación de los datos..... | 7 |
| 1.3. Variables y observaciones..... | 8 |
| 1.4. Medidas de tendencia central..... | 10 |
| 1.4.1. Moda..... | 10 |
| 1.4.2. Mediana..... | 11 |
| 1.4.3. Media..... | 11 |
| 1.5. Medidas de dispersión..... | 12 |
| 1.5.1. Varianza..... | 12 |
| 1.5.2. Desviación estándar..... | 12 |
| 1.5.3. Rango..... | 13 |
| 1.6. Normalización de los datos..... | 13 |
| 1.7. Redundancia esencial..... | 16 |
| 1.8. Preprocesamiento de los datos..... | 16 |
| 1.9. Limpieza de los datos..... | 16 |
| 1.10. Identificación de clasificación errónea..... | 17 |
| 1.11. Identificación de valores atípicos..... | 18 |
| 1.11.1. Identificación con medidas centrales y de dispersión..... | 20 |
| 1.12. Tratamiento de datos faltantes..... | 22 |
| 1.13. Imputación de la información faltante..... | 23 |
| 1.14. Transformación de datos..... | 25 |
| 2. APRENDIZAJE AUTOMÁTICO..... | 27 |
| 2.1. Tipos de aprendizaje automático..... | 27 |
| 2.1.1. Aprendizaje supervisado..... | 27 |
| 2.1.2. Aprendizaje no supervisado..... | 27 |
| 2.2. Tipos de soluciones en el aprendizaje automático..... | 28 |
| 2.3. Clasificación mediante Regresión Logística..... | 29 |
| 2.3.1. Razón de probabilidades (Odds Ratio)..... | 30 |
| 2.3.2. Función Logit..... | 30 |

| | | |
|----------------------------------|---------------------------------------------------------------------------------------------------------------------|-----------|
| 19 | 3. MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO EN EL CUIDADO DE LA SALUD..... | 31 |
| 3.1. | Ventajas del aprendizaje automático en el cuidado de la salud..... | 31 |
| 3.2. | Las 4 "V" del análisis de datos y el cuidado de la salud..... | 34 |
| 4. DIABETES MELLITUS..... | 35 | |
| 17 | 4.1. Tipos de diabetes..... | 35 |
| 4.1.1. | Diabetes Tipo 1..... | 35 |
| 4.1.2. | Diabetes tipo 2..... | 36 |
| 4.2. | Métodos de diagnóstico..... | 37 |
| 4.3. | Complicaciones agudas de la diabetes..... | 38 |
| 4.3.1. | Hipoglucemias..... | 38 |
| 4.3.2. | Hiperglucemias..... | 39 |
| 4.3.3. | Coma hiperglucémico hiperosmolar no cetósico..... | 40 |
| 4.3.4. | Acidosis láctica..... | 41 |
| 4.4. | Complicaciones crónicas de la diabetes..... | 41 |
| 4.4.1. | Macroangiopatía..... | 42 |
| 4.4.2. | Cardiopatía isquémica..... | 42 |
| 4.4.3. | Cardiopatía isquémica silente..... | 42 |
| 4.4.4. | Insuficiencia cardiaca..... | 43 |
| 4.4.5. | Arteriopatía periférica..... | 43 |
| 4.4.6. | Accidente cerebrovascular..... | 43 |
| 111 | 4.4.7. Estenosis de la arteria renal y aneurisma de la aorta abdominal..... | 44 |
| 4.5. | Otras complicaciones..... | 44 |
| 4.5.1. | Pie diabético..... | 44 |
| 4.5.2. | Alteraciones en la piel..... | 45 |
| 4.5.3. | Alteraciones bucales..... | 45 |
| 4.6. | Datos y cifras de la diabetes en México..... | 45 |
| 29 | 5. MODELO DE REGRESIÓN LOGÍSTICA PARA EL ANÁLISIS DE HÁBITOS Y COMPLICACIONES EN PACIENTES CON DIABETES..... | 48 |
| 206 | 5.1. Encuesta Nacional de Salud y Nutrición 2018..... | 48 |
| 5.2. | Descripción del conjunto de datos de la muestra..... | 48 |
| 5.3. | Estado del arte..... | 50 |
| 5.4. | Métodos y herramientas implementados en el modelo..... | 57 |
| 5.5. | Descripción del modelo..... | 60 |

| | | |
|-------------------------|--------------------------------------------------------------|-----------|
| 5.6. | Resultados para cada una de las variables..... | 60 |
| 5.6.1. | Resultados para la variable “Ulceras en pies o piernas”..... | 60 |
| 5.6.2. | Resultados para la variable “Amputación”..... | 63 |
| 5.6.3. | Resultados para la variable “Perdida de la vista”: | 65 |
| 5.6.4. | Resultados para la variable “Insuficiencia renal aguda”..... | 68 |
| 5.6.5. | Resultados para la variable “Infarto cardiaco”..... | 71 |
| 5.6.6. | Resultados para la variable “Infarto cerebral”..... | 73 |
| 5.6.7. | Resultados para la variable “Coma diabético”..... | 75 |
| 5.7. | Diabetes - Intelligent Notify System (Diabet-INS) | 78 |
| 5.7.1. | Herramientas para el desarrollo del sistema..... | 78 |
| 5.7.2. | Descripción del sistema..... | 79 |
| 5.7.3. | Actividades y notificaciones de prevención..... | 85 |
| CONCLUSIONES. | | 88 |
| REFERENCIAS..... | | 90 |

INTRODUCCIÓN.

⁴⁹ La diabetes es una enfermedad crónica en la que el páncreas, órgano que produce la insulina, una hormona que regula los niveles de azúcar en la sangre no produce suficiente o ⁷⁰ lo utiliza eficazmente, lo que provoca un aumento significativo de los niveles de azúcar en la sangre (hiperglucemia) que, con el tiempo puede dañar gravemente muchos órganos y sistemas de las personas que padecen dicha enfermedad (1).

La diabetes ³implica un fuerte impacto en la economía de los pacientes y sus familias, representa una de las cuatro principales causas de muerte por enfermedades no transmisibles en el mundo y en México es una de las enfermedades crónicas con mayor incidencia en la población y está relacionada con altos índices de morbilidad y mortalidad (1).

³⁰ Existen dos tipos principales de diabetes, la diabetes tipo 1 y la diabetes tipo 2. Esta última es la más común en la población y representa del 85 al 90 por ciento de los casos, presentándose principalmente en adultos (2).

⁹⁹ En este contexto, la diabetes tipo 2 que se caracteriza por la incapacidad del páncreas para utilizar la insulina para la regulación de los niveles de azúcar en la sangre, hoy en día puede prevenirse eficazmente tomando medidas en relación con el estilo de vida de las personas. Algunos de los principales aspectos a tomar en cuenta para su prevención son entre otros, realizar actividad física de forma regular, mantener un peso corporal adecuado y llevar una dieta saludable evitando los azúcares y las grasas saturadas (2).

Gracias a la digitalización de la información, hoy en día podemos contar con grandes cantidades de datos sobre salud, con los que mediante una clasificación adecuada y con el apoyo de diferentes métodos de aprendizaje automático, podemos obtener grandes ventajas en diferentes áreas para el cuidado de la salud. Así pues, gracias al análisis de grandes cantidades de información almacenada en bases de datos históricos, podemos predecir eventos que podrían conducir a una mejor atención de los pacientes con diferentes padecimientos.

²⁷ El aprendizaje automático es una rama de la inteligencia artificial que consiste en la aplicación de programas informáticos que aprenden con base en sus experiencias para realizar tareas de forma más eficiente a medida que evoluciona. Su principal objetivo es el desarrollo de teorías, técnicas y algoritmos que permitan a un sistema modificar su comportamiento a través de la inferencia inductiva que se basa en la observación de datos que representan información sobre un proceso o fenómeno estadístico. En la actualidad, el aprendizaje automático supone una gran ventaja para la aplicación de sistemas que dan soporte a diferentes áreas dedicadas a la atención sanitaria (3).

77

Por otro lado, la minería de datos consiste en detectar, extraer y procesar la información de grandes conjuntos de datos. Con la ayuda de diferentes recursos informáticos y mediante el análisis estadístico, esta herramienta sirve para encontrar patrones o tendencias que los datos presentan y los cuales resultarían imposibles de detectar mediante un análisis tradicional ya que la relación contenida en los datos resulta demasiado compleja debido a la gran cantidad de información disponible (4).

A través de la minería de datos, las organizaciones sanitarias pueden obtener grandes beneficios como una gestión eficaz de la salud de una población, la detección de fraudes en la atención médica, la gestión de los recursos hospitalarios y tratamientos para distintas enfermedades, entre otros. Uno de los principales es la detección temprana de enfermedades, lo que permite un tratamiento más eficaz que también se traduce en la prevención de complicaciones que podrían presentarse si la enfermedad no se detecta o atiende a tiempo (5).

Con el avance de la tecnología, hoy en día podemos recopilar grandes cantidades de datos (Big Data) relacionados con la diabetes. A través de estas herramientas de aprendizaje automático podemos utilizar dichos datos para ¹⁶²entrar programas informáticos y lograr detectar marcadores que nos indiquen el riesgo de presentar complicaciones durante el tratamiento y control de la enfermedad. Dicho de otra manera, podemos encontrar patrones que nos permitan alertar a los pacientes sobre la probabilidad de que la enfermedad se complique derivado del análisis de sus patrones de comportamiento para así influir en ellos e incentivar a que estos modifiquen sus hábitos (5).

58

1. MINERÍA DE DATOS.

La minería de datos consiste en detectar, extraer y procesar la información de grandes conjuntos de datos. Con la ayuda de diferentes recursos informáticos y mediante el análisis estadístico, esta herramienta sirve para encontrar patrones o tendencias que los datos presentan y los cuales resultarían imposibles de detectar mediante un análisis tradicional ya que la relación contenida en los datos resulta demasiado compleja debido a la gran cantidad de información disponible (4).

Hoy en día, la minería de datos se aplica en una gran cantidad de disciplinas, ya sea en instituciones, empresas u otras organizaciones, estas técnicas les permiten analizar patrones o tendencias en sus bases de datos existentes lo que representa una gran ventaja al momento de competir con aquellas que no la practican, ya que estas últimas desaprovechan información valiosa oculta a simple vista en sus repositorios de bases de datos (6).

131

En síntesis, la minería de datos es un proceso de análisis de grandes cantidades de datos con la finalidad de encontrar patrones o tendencias.

1.1. Análisis exploratorio de los datos.

87

Previo a la aplicación de cualquier técnica de minería de datos, resulta necesario realizar un análisis profundo de la información y las bases de datos con las que se pretende trabajar. Es necesario examinar de manera minuciosa cada una de las variables, las relaciones existentes entre sí y evaluar el contenido en general para solucionar problemas en el diseño de la investigación y la recopilación de los datos. En este proceso resulta indispensable contar con herramientas que aportan técnicas para el análisis de los datos como algunas que permiten la exploración de las características de la distribución de las variables implicadas en el análisis, sus relaciones y sus diferencias. Con esto, se logra averiguar de forma exhaustiva en la información y se consigue detectar posibles anomalías (7).

1.2. Representación de los datos.

36

El punto de partida para el análisis de datos consiste principalmente en una tabla de datos denominada comúnmente conjunto de datos. Esta contiene los valores de los datos medidos o recogidos que pueden ser representados como números o texto. Estos datos, antes de ser transformados o modificados son denominados datos crudos. Un conjunto de datos enumera los diferentes elementos sobre los que se han recogido o medido los datos. En este conjunto de datos, la información que se considera importante se muestra como diferentes atributos. Los elementos individuales suelen mostrarse como filas en una tabla de datos y los diferentes atributos son mostrados como columnas (8).

Las anotaciones de las variables representan otro nivel de detalle para tener en cuenta. Estas proporcionan información adicional importante que permite conocer

el contexto de los datos. Las unidades de medida son útiles a la hora de presentar los resultados y son fundamentales para la interpretación de los datos. Con ello se puede entender cómo deben alinearse las unidades o que transformaciones sufrirán cuando se fusionan tablas de datos de diferentes fuentes (9).

1.3. Variables y observaciones.

Todas las disciplinas recopilan datos sobre elementos que les resultan importantes para determinado campo de estudio. Estos elementos son organizados en una tabla para el análisis de datos en la que cada fila, denominada observación, contiene información sobre el elemento específico que representa la fila. Estas tablas también contienen atributos acerca de los elementos de la tabla. Se consideran atributos a un conjunto de valores que describen algún aspecto para todas las observaciones a las que se les denomina variables. En concreto, cada fila de la tabla describe una observación y cada columna describe una variable (8).

La forma más común de observar los datos es mediante hojas de cálculo, donde los datos brutos se muestran como filas de observaciones y columnas de variables. Este tipo de visualización resulta útil para revisar los datos brutos, sin embargo, la tabla puede resultar abrumadora cuando contiene una cantidad muy grande de observaciones o variables. En este contexto, clasificar la tabla en función de una variable o varias variables es útil para organizar los datos, sin embargo, es difícil identificar tendencias o relaciones observando solamente los datos brutos (8).

Antes del análisis o extracción de datos, resulta esencial entender el contenido del conjunto de datos y el primer paso consiste en comprender a detalle las variables individuales (8).

Algunas técnicas para el análisis de datos contienen restricciones en cuanto a los tipos de variables que lograran procesar. Por lo tanto, conocer los tipos de variables **1** nos permite realizar una selección adecuada de la técnica que utilizaremos para realizar el análisis de los datos. También podemos transformar los datos en una forma que se aadecue a la técnica que deseamos implementar para el análisis. Además, **186** ciertas características de las variables pueden tener implicaciones en cuanto a la interpretación de los resultados del análisis (8).

Cada una de las variables en una tabla de datos puede examinarse de diferentes maneras. Es útil iniciar por definir cada variable basándose en el tipo de valores que contiene. Por ejemplo, las variables que contienen un número fijo de valores distintos se denominan variable discreta, esta puede adoptar un numero finito de valores y una variable que puede tomar cualquier valor numérico se denomina variable continua, estas últimas pueden tener un numero infinito de valores (8).

Las variables pueden clasificarse también según la escala en la que se miden. Las escalas nos ayudan a comprender la **72** posición de una variable individual y se utilizan para la visualización de los datos tanto para la toma de decisiones como la selección

de los métodos para su análisis. A continuación, una descripción de dichas escalas (8):

- **Escala nominal:** Describe una variable con un número limitado de valores diferentes que no pueden ser ordenados. Dado que los valores en una variable nominal asignan simplemente una observación a una categoría en particular, el orden de estos valores no tiene ningún significado. Por ejemplo, una variable “Industria” que contiene valores categóricos como “Financiero”, “Ingeniería”, “Comercio”, etc. (8).
- **Escala ordinal:** Describe una variable cuyos valores pueden ordenarse o clasificarse. Al igual que en la escala nominal, los valores se asignan a un número fijo de categorías. En una variable ordinal, aunque los valores estén ordenados, resulta imposible determinar la magnitud de la diferencia entre los valores. Por ejemplo, una escala en la que los valores son “bajo”, “medio” y “alto”, podemos determinar que “alto” es mayor que “medio” y este a su vez es mayor que “bajo”. Sin embargo, no se puede comparar la diferencia entre “alto” y “medio” con la diferencia entre “medio” y “bajo” (8).
- **Escala de intervalo:** Describe valores en los que se puede comparar el intervalo entre valores. En esta, los intervalos entre los valores de la escala comparten la misma unidad de medida por lo que pueden compararse de forma significativa. Sin embargo, como la escala carece de un cero significativo, las proporciones de los valores no pueden compararse. Duplicar un valor no implica duplicar la medida real. Por ejemplo, 10 °F no es el doble de calor que 5 °F (8).
- **Escala de proporción:** Describe variables en las que pueden compararse tanto los intervalos entre valores como las razones de estos. Las escalas para las que es posible tomar proporciones de valores se definen como que tienen un cero natural. Un ejemplo para estas es el saldo de una cuenta bancaria cuyos valores son 5\$, 10\$ y 15\$. La diferencia entre cada una es de 5\$ y 10\$ si corresponde el doble de 5\$ (8).
- **Escala dicotómica:** Se denomina así, si solo puede contener dos valores. Por ejemplo, los valores de una variable “Genero” solo pueden ser “Hombre” o “Mujer”. Una variable binaria es una variable dicotómica ampliamente utilizada con valores 0 o 1. Estas variables son utilizadas a menudo para el análisis de datos, dado que proporcionan una representación numérica conveniente para muchos tipos de diferentes datos discretos (8).

Algunos tipos de variables no se utilizan directamente para el análisis de datos, sin embargo, estas pueden ser útiles para preparar las tablas de datos o para interpretar los resultados del análisis. A veces se utiliza una variable para identificar cada

observación en una tabla de datos, la cual tendrá un valor único para cada observación. La inclusión de un identificador único puede proporcionar una referencia a la información detallada de cada observación (8).

Las anotaciones de las variables representan otro nivel de detalle para tener en cuenta. Estas proporcionan información adicional importante que permite conocer el contexto de los datos. Las unidades de medida son útiles a la hora de presentar los resultados y son fundamentales para la interpretación de los datos. Con ello se puede entender cómo deben alinearse las unidades o qué transformaciones sufrirán cuando se fusionan tablas de datos de diferentes fuentes (8).

9

1.4. Medidas de tendencia central.

De las diversas formas en que se puede resumir una variable, una de las más importantes es el valor utilizado para caracterizar el centro del conjunto de valores que contiene. Es bastante útil cuantificar el punto medio o central de una variable, como su media, en torno a la cual se sitúan muchos de los valores de las observaciones de dicha variable. Existen varios enfoques para calcular este valor y el que se utilizara puede depender de la clasificación de la variable (9).

1.4.1. Moda.

Representa comúnmente el valor más reportado para una variable en particular y es usualmente representado como M_o . El cálculo de esta medida se muestra en el siguiente ejemplo (9):

$$X = [70 \ 50 \ 40 \ 70 \ 80 \ 70 \ 60 \ 90 \ 70 \ 70 \ 100 \ 80 \ 60 \ 70 \ 80 \ 60]$$

Para visualizar mejor los datos, los acomodamos de menor a mayor, de la siguiente manera:

$$X = [40 \ 50 \ 60 \ 60 \ 60 \ 70 \ 70 \ 70 \ 70 \ 70 \ 80 \ 80 \ 80 \ 90 \ 100]$$

En este ejemplo podemos observar que el valor que más se repite es el "70" por lo que, $M_o = 70$.

Cuando nos encontramos con más de un valor con el mayor y mismo número de ocurrencias, se pueden reportar todos los valores o se selecciona un punto medio. Por ejemplo, para los siguientes valores:

$$X = [3 \ 4 \ 5 \ 6 \ 7 \ 7 \ 7 \ 8 \ 8 \ 8 \ 9]$$

Tanto el valor 7 como el valor 8 se informan tres veces. En este caso, la moda puede ser reportada como $M_o = \{7,8\} = 7.5$.

67

La moda proporciona la única medida de tendencia central para las variables medidas en una escala nominal, sin embargo, esta también puede calcularse para las variables medidas en las escalas ordinal, de intervalo y de razón.

1.4.2. Mediana.

9

La mediana representa el valor medio de una variable, una vez ordenada de menor a mayor. Si la variable tiene un numero impar de valores la mediana se calcula de la siguiente manera (9):

$$M_e = X \frac{n}{2}$$

44

Si la variable tiene un numero par de valores la mediana se calcula como sigue (9):

$$M_e = \frac{X_{n/2} + X_{n+1/2}}{2}$$

En el siguiente ejemplo, utilizamos el siguiente conjunto de valores de una variable para calcular la mediana:

$$X = [40 50 60 60 60 70 70 70 70 70 70 80 80 80 90 100]$$

$$M_e = \frac{X_{n/2} + X_{n+1/2}}{2} = \frac{70 + 70}{2} = 70$$

1.4.3. Media.

La media, comúnmente denominada promedio, representa el resumen de tendencia central más utilizado para las variables medidas en las escalas de intervalo de razón. Esta se define como la suma de todos los valores dividida por el número de valores (9).

Para una variable que representa un subconjunto de todas las observaciones posibles de (X), la mediana suele denominarse \bar{X} . La fórmula para calcular la media donde n es el número de observaciones y X_i representa los valores individuales, suele escribirse con la siguiente formula (9):

$$\bar{X} = \frac{\sum X}{n}$$

En el siguiente ejemplo, utilizamos el conjunto de valores de una variable X para calcular la media:

$$X = [40 50 60 60 60 70 70 70 70 70 70 80 80 80 90 100]$$

$$n = 16$$

$$\bar{X} = \frac{\sum X}{n} = \frac{1120}{16} = 70$$

1.5. Medidas de dispersión.

Existen ocasiones en las que una población o muestra que son muy distinta pueden presentar medidas de tendencia central muy parecidas. Por ello, en ocasiones para describir las características principales de una distribución estadística no es suficiente con las medidas de centralización. Las medidas de centralización no detectan ciertas circunstancias de la distribución que pueden ser bastante importantes y las cuales deben tomarse en cuenta para realizar una descripción acertada de su distribución. Es ¹⁵⁷ aquí donde se utilizan las medidas de dispersión pues estas nos dan indicios de si los datos están relativamente agrupados respecto a las medidas de tendencia central (10).

1.5.1. Varianza.

191

Este parámetro nos permite detectar las variaciones de cada valor respecto a la media aritmética. Para ello, primero se elevan las diferencias de los valores al cuadrado, con lo que se evitan posibles compensaciones. Por último ⁵⁷ considera el promedio de dichas diferencias (varianza). La varianza es, junto a la desviación estándar, la medida de dispersión más utilizada en estadística. En términos matemáticos, la varianza es igual a la media de los cuadrados menos el cuadrado de la media y se representa mediante la siguiente fórmula (10).

$$S^2 = \frac{\sum(X - \bar{X})^2}{N}$$

Sin embargo, puede haber ocasiones en que por el tamaño o características de la población resulte más adecuado realizar el cálculo a una muestra. Para este supuesto, la fórmula que utilizaremos será la siguiente (10):

$$\sigma^2 = \frac{\sum(X - \bar{X})^2}{n - 1}$$

En el siguiente ejemplo, utilizamos el conjunto de valores de una variable X para calcular la varianza:

$$X = [40 50 60 60 60 70 70 70 70 70 80 80 80 90 100]$$

$$S^2 = \frac{900 + 400 + 100 + 100 + 100 + 100 + 100 + 100 + 400 + 900}{16} = 200$$

1.5.2. Desviación estandar.

El cálculo de la desviación estandar tiene el mismo objetivo que el cálculo de la varianza ⁶ con la ventaja de que las unidades en las que se mide son las mismas que las de los datos de la distribución. Esta es considerada la medida de dispersión por excelencia. La desviación estandar también podemos encontrarla con el término de

⁵⁷ desviación típica y se define como la raíz cuadrada positiva de la varianza. En términos matemáticos, la desviación estándar es la raíz cuadrada de la media de los cuadrados menos el cuadrado de la media. De acuerdo con la definición anterior, la fórmula para la obtención de la desviación estándar ya sea para una muestra o para una población es la siguiente, respectivamente (10):

$$S = \sqrt{\frac{\sum(X - \bar{X})^2}{N}} \quad \sigma = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

Utilizando el ejemplo anterior, el cálculo de la desviación estándar se obtiene de la siguiente manera:

$$S = \sqrt{\frac{3200}{16}} = \sqrt{200} = 14.1421$$

1.5.3. Rango.

⁶ El rango de una variable estadística se define como la diferencia entre el mayor y el menor valor de la variable. Este nos indica la longitud del intervalo en el que se encuentran todos los datos de la distribución. El rango supone una medida de dispersión importante aun que resulta insuficiente para la valoración de la variabilidad de los datos. La forma para expresar el rango mediante una fórmula es la siguiente (10):

$$R = \text{Max}(X) - \text{Min}(X)$$

En el siguiente ejemplo, utilizamos el conjunto de valores de una variable X para conocer su rango:

$$X = [40 \ 50 \ 60 \ 60 \ 60 \ 70 \ 70 \ 70 \ 70 \ 70 \ 80 \ 80 \ 80 \ 90 \ 100]$$

$$R = 100 - 40 = 60$$

1.6. Normalización de los datos.

¹⁸⁴ La preparación de los datos es ¹⁷⁵ de las tareas que implica más tiempo en un proyecto de análisis o minería ¹³⁶ de datos. La forma en que se recogen y preparan los datos resulta fundamental para el análisis y la toma de decisiones de forma confiable. Los datos deben recolectarse en una tabla y ello puede implicar la integración de datos de múltiples fuentes con diferentes formatos o escalas. Por ello, los datos deben limpiarse, resolviendo las ambigüedades y los errores, eliminando los datos redundantes o problemáticos y descartando aquellos que resultan irrelevantes para el análisis que se pretende realizar (8).

Dependiendo de cómo se diseñe una base de datos relacional, esta puede ser susceptible a múltiples tipos de problemas. Ya sea que contenga datos duplicados

o irrelevantes. Esto puede no solo ocupar espacio de almacenamiento, sino que también provoca que la actualización de todos estos datos resulte en una tarea que consume demasiado tiempo. En terminología de bases de datos, a este tipo de problemas se les denomina anomalías o problemas (11).

11

La normalización es un proceso que consiste en reordenar la base de datos para convertirla en una forma estándar (normal) que evite las anomalías. Existen diferentes niveles de normalización agrupados de forma ascendente y en cada uno de ellos se incluyen los niveles anteriores. Por ejemplo, una base de datos se encuentra en tercera forma normal si se ha cumplido primero con la segunda y primera forma normal y cumple algunas propiedades adicionales. Esto significa que, si una base de datos se encuentra en un nivel de normalización, entonces por definición obtiene las ventajas de los niveles anteriores (11).

A continuación, se muestran los diferentes niveles de normalización en orden ascendentes o del más débil al más fuerte:

11

- **Primera forma normal 1NF:** La primera forma normal consiste básicamente en que los datos se encuentren en una base de datos. La mayoría de las propiedades necesarias para que se encuentre en 1NF se cumple de forma automática en cualquier base de datos relacional razonable (11).

200

Las características principales de una base de datos en este nivel son las siguientes (11):

- Cada columna debe tener un único nombre.
- No importa el orden de las columnas y las filas.
- Cada columna debe tener un único tipo de datos.
- No puede haber dos filas con valores idénticos.
- Cada columna debe contener un único valor.
- Las columnas no pueden contener grupos repetidos.

- **Segunda forma normal 2NF:** Una base de datos se encuentra en 2NF si cumple con la forma 1NF y si todos los campos no clave dependen de todos los campos clave (11).
- **Tercera forma normal 3NF:** En este punto la base de datos se considera que se encuentra en forma 3NF si esta se encuentre en forma 2NF y no contiene dependencias transitivas. Una dependencia transitiva se da cuando el valor de un campo no clave depende del valor de otro campo no clave (11).

3

Muchos diseñadores de bases de datos dejan de normalizar una vez llegando a la forma 3NF debido a que esta proporciona comúnmente una mejor relación costo beneficio. Resulta bastante fácil convertir una base de datos a forma 3NF y este nivel de normalización evita las anomalías de datos más

comunes. Almacena los datos por separado de forma que es posible añadir y eliminar piezas de información sin destruir datos no relacionados. También elimina los datos redundantes para que la base de datos no esté llena de una gran cantidad de copias de la misma información que ocuparía bastante espacio y dificultaría la actualización de los valores. Sin embargo, la base de datos en este nivel puede seguir siendo vulnerable a algunas anomalías menos comunes que se evitarían con una normalización más completa (11).

- **Forma normal de Boyce-Codd BCNF:** Se considera que una base de datos se encuentra en BCNF cuando esta se encuentra en forma 3NF y cada determinante es una llave candidata. Un determinante es un campo que determina, al menos en parte, el valor de otro campo. Este punto resulta un tanto técnico por lo que debemos tomar en cuenta que una superclave es un conjunto de campos que contienen valores únicos. Se puede usar una superclave para identificar de forma exclusiva los registros de una tabla. También debemos considerar que una clave candidata es una superclave mínima (11).

En otras palabras, si elimina alguno de los campos de la clave candidata, esta dejará de ser una superclave. En la definición de 3NF se hace referencia a los campos que dependen de otro campo que no forma parte de la clave primaria. Ahora estamos hablando de campos que en esencia si pueden depender de otros campos que forman parte de la clave primaria o de cualquier clave candidata (11).

- **Cuarta forma normal 4NF:** Una base de datos se encuentra en forma 4NF cuando cumple con la forma BCNF y además no contiene una dependencia multivaluada no relacionada. Un ejemplo de dependencia multivaluada no relacionada sería una tabla en la que la variable "Empleado" implica "Habilidad" y a la vez implica "Herramienta" pero "Habilidad" y "Herramienta" son independientes (11).
- **Quinta forma normal 5NF:** Una base de datos se encuentra en forma normal de unión de proyectos o 5NF si esta se encuentra principalmente en 4NF y además no contiene dependencias multivaluadas relacionadas (11).
- **Forma normal de Dominio/Clave DKNF:** Podemos decir que la base de datos se encuentra en forma normal de dominio/clave o DKNF si esta no contiene ninguna restricción, excepto las restricciones de dominio y las restricciones de clave (11).

Otras palabras, una base de datos se encuentra en DKNF si cada restricción es una consecuencia de las restricciones de dominio y clave. El

dominio de un campo consta de sus valores permitidos. Una restricción de dominio significa simplemente que un campo tiene un valor que se encuentra en su dominio. Es fácil de comprobar si una restricción de dominio se satisface simplemente examinando todos los valores del campo. Una restricción de clave significa que los valores de los campos que componen una clave son únicos. Por lo tanto, la base de datos está en DKNF, con la finalidad⁴⁴ de validar todas las restricciones de los datos y para ello basta con validar las restricciones de dominio y las restricciones de clave (11).

Una base de datos que se encuentra en DKNF está fuertemente protegida contra las anomalías (11).

1.7. Redundancia esencial.

Una de las principales anomalías en los datos es la redundancia. Si una tabla contiene muchos datos redundantes, es bastante probable que sea vulnerable a las anomalías de datos, especialmente a las anomalías de modificación. Sin embargo, esto no siempre resulta cierto, siempre y cuando los datos redundantes se encuentren en las claves. Por ejemplo, en estos casos, la repetición de esos datos en diferentes tablas resulta necesaria para representar los datos contenidos en ellas (11).

1.8. Preprocesamiento de los datos.

La mayoría⁴⁵ las veces, derivado de la gran cantidad de registros con que se cuenta en las bases de datos, en la minería de datos podemos encontrar información que podría generar ruido al momento del análisis. Por ejemplo, podríamos encontrarnos con datos incompletos, campos obsoletos o redundantes, valores atípicos o incoherentes o incluso información que no necesitamos para el análisis que pretendemos realizar (6).

En este contexto, en la minería de datos el primer paso a realizar es un preprocesamiento de los datos ya que podemos encontrar información que podría generar ruido al momento de llevar a cabo un análisis. Además, dependiendo del modelo y las técnicas que pretendemos utilizar y la información que intentamos obtener, muchas veces resulta necesario modificar el formato de los datos para un análisis correcto.

El objetivo principal del preprocesamiento de datos consiste en minimizar la cantidad de ruido que entra en nuestro modelo para obtener un análisis más preciso (6).

1.9. Limpieza de los datos.

En ocasiones podemos encontrar registros con diferentes tipos de datos que no necesariamente son incorrectos y que eliminarlos podría suponer un error. Por ello, para este tipo de información inusual es preferente buscar la forma de convertir ese

registro en un formato más adecuado para el análisis. Para exemplificar, en la Tabla 1 tenemos una base de datos con registros de información de la temperatura corporal de pacientes con resultado positivo y negativo en la prueba de PCR para Covid19 en diferentes regiones por lo que en algunas partes la unidad de medición está registrada en grados Celsius y en otras en Fahrenheit (6).

| ID | Sexo | PCR Covid19 | Temperatura | Lugar | Fecha |
|-----|------|-------------|-------------|---------|------------|
| 001 | M | Positivo | 37.5°C | México | 20/10/2021 |
| 002 | F | Positivo | 100.8°F | EE. UU. | 20/10/2021 |
| 003 | F | Negativo | 36.5°C | México | 20/10/2021 |
| 004 | M | Positivo | 38.0°C | México | 20/10/2021 |
| 005 | M | Negativo | 98.5°F | EE. UU. | 20/10/2021 |
| 006 | F | Positivo | 38.5°C | México | 20/10/2021 |
| 007 | M | Positivo | 37.5°C | México | 20/10/2021 |

Tabla 1. Ejemplo de diferentes escalas de temperatura registradas (Limpieza de datos).

La mejor opción para este caso es realizar una conversión en lugar de eliminar registros que podrían resultar valiosos para el análisis.

| ID | Sexo | PCR Covid19 | Temperatura | Lugar | Fecha |
|-----|------|-------------|-------------|---------|------------|
| 001 | M | Positivo | 37.5°C | México | 20/10/2021 |
| 002 | F | Positivo | 38.0°C | EE. UU. | 20/10/2021 |
| 003 | F | Negativo | 36.5°C | México | 20/10/2021 |
| 004 | M | Positivo | 38.0°C | México | 20/10/2021 |
| 005 | M | Negativo | 37.0°C | EE. UU. | 20/10/2021 |
| 006 | F | Positivo | 38.5°C | México | 20/10/2021 |
| 007 | M | positivo | 37.5°C | México | 20/10/2021 |

Tabla 2. Ejemplo de conversión de la escala de temperatura de la Tabla 1 (Limpieza de datos).

Así también, utilizando para este ejemplo las tablas 1 y 2, suponiendo que estamos analizando una comparación de la temperatura y el género de pacientes positivos y negativos en la prueba PCR de Covid19, sin importar el lugar o la fecha en que se hizo el registro, podríamos excluir las columnas “Lugar” y “Fecha” para con ello ahorrarnos tiempo y trabajo de procesamiento computacional en el análisis.

En esta parte de la minería de datos, el principal objetivo es, por un lado, identificar información o datos que podrían parecer un error pero que no lo son y buscar la forma de corregirlos o transformarlos de manera que puedan seguir siendo utilizados en el modelo para su análisis y por otro lado determinar según el estudio que se está realizando, la información o los datos que aportan un valor en la investigación (6)..

1.10. Identificación de clasificación errónea.

Otro punto importante en el preprocesamiento de los datos consiste en identificar los errores en las etiquetas de clasificación de las variables categóricas (6)..

Por ejemplo, en la tabla 3 se muestra una distribución de frecuencias con siete clases: “Jalisco”, “Ciudad de México”, “Michoacán de Ocampo”, “Mórenlos”, “Nayarit”, “Nuevo León” y “Guadalajara”. Sin embargo, la clase “Guadalajara” corresponde a un dato de clasificación de clase erróneo ya que la etiqueta entidad se refiere a los estados del país y Guadalajara es un municipio que corresponde al

estado de Jalisco por lo que ese registro no debería figurar en nuestra base de datos.

| Entidad | Población | Porcentaje | Pacientes | Enfermedad |
|---------------------|-----------|------------|-----------|------------|
| Jalisco | 5338306 | 7.59 | 405299 | Diabetes |
| México | 11938264 | 8.96 | 1069493 | Diabetes |
| Michoacán de Ocampo | 2962822 | 9.93 | 294199 | Diabetes |
| Morelos | 1346952 | 11.99 | 161497 | Diabetes |
| Nayarit | 860475 | 9.55 | 82191 | Diabetes |
| Nuevo León | 3719881 | 12.63 | 469656 | Diabetes |
| Guadalajara | 1435315 | 9.5 | 136354 | Diabetes |

Tabla 3. Ejemplo de clasificación errónea.

1.11. Identificación de valores atípicos.

Los valores atípicos pueden representar errores en la introducción de información en una base de datos. Estos se identifican como valores extremos que van en contra de la tendencia de los datos restantes. Incluso puede haber datos que no necesariamente correspondan a un error pero que si presenta una tendencia diferente a la de los demás datos por lo que ciertos métodos estadísticos pueden ser sensibles a la presencia de dichos valores atípicos y podrían ofrecer resultados poco fiables en los resultados del análisis (6).

Los valores que se definen concretamente como observaciones aisladas cuya conducta se diferencia claramente del comportamiento medio del resto de observaciones, afectan fuertemente al análisis, sobre todo en muestras pequeñas (7).

Durante el análisis exploratorio de los datos podemos encontrar casos atípicos derivados de observaciones que provienen de un error de procedimiento, como errores en la codificación o en la entrada de los datos. Otro tipo de casos contempla aquellas observaciones consecuentes de un acontecimiento extraordinario por lo que existe una explicación valida de su presencia en la muestra y dependiendo del valor que pueda aportar para el análisis, podría optarse por conservarlos (7).

Por otro lado, podemos encontrar otro tipo de casos que comprende observaciones extraordinarias para las que el investigador no encuentra explicación alguna por lo que comúnmente resulta conveniente eliminar dichos datos. También podemos encontrar observaciones que se sitúan fuera del rango ordinario de valores en las variables. A estos se les denomina valores extremos y lo más común es eliminarlos al observar que no sean elementos significativos en la investigación. En conclusión, las características propias de cada valor atípico serán las que determinen si se mantiene o se elimina dicha información (7).

El método principal para identificar valores atípicos consta de herramientas gráficas que permiten visualizar mejor este tipo de errores, por ejemplo, en la siguiente gráfica se muestra la cantidad de pacientes por determinados rangos de edad y podemos observar a la derecha una barra con valores atípicos y en la cual, al realizar un análisis de la información, nos podemos dar cuenta de que se trata de

un duplicado del registro referente al rango de edad “20 – 30” con valores atípicos (6).

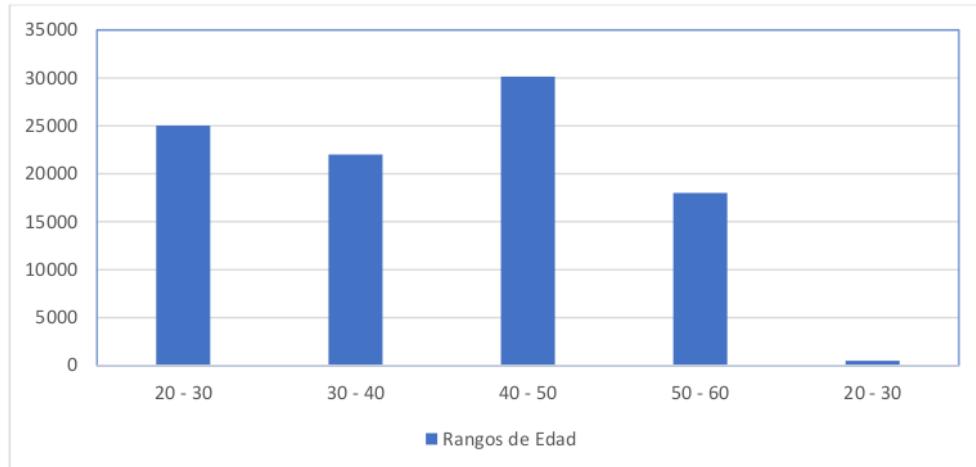


Diagrama 1. Ejemplo de grafica con valores atípicos en la muestra.

Otra manera de detectar los valores atípicos mediante herramientas visuales es a través de los diagramas de control que consisten en una representación gráfica con una línea central que señala el valor medio de la variable y dos líneas más que señalan el límite superior y el límite inferior de control. Con esto, todos los valores que se encuentren dentro de los límites de control son valores válidos y aquellos que se encuentren fuera de estos límites, se consideran atípicos por lo que requieren de un análisis más puntual para determinar las correcciones pertinentes (7).

En la siguiente grafica se muestra un ejemplo de diagrama de control en el que se analiza la variable de “Edad” de diferentes pacientes y se logra observar tanto las dos líneas que señalan los límites inferior y superior en la gráfica como un valor atípico que sobresale de los límites de control y corresponde a un paciente de edad bastante alejada del promedio de la de los demás pacientes en la muestra.

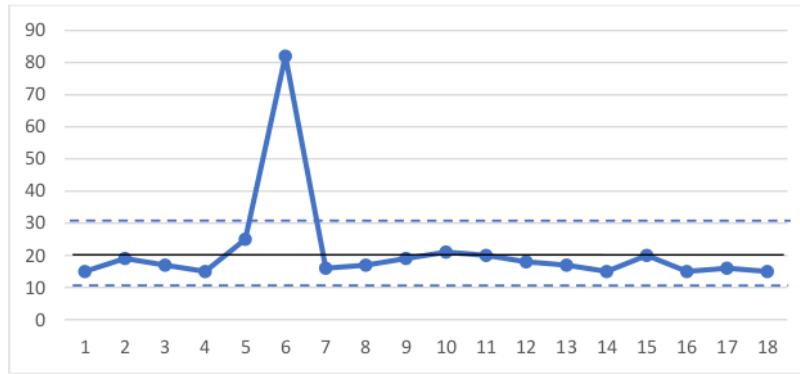


Diagrama 2. Ejemplo de diagrama de control con valores atípicos.

Otro ejemplo son los gráficos de caja de bigotes en los que los valores atípicos se visualizan como puntos aislados en los extremos ¹ de los bigotes. En algunas herramientas para análisis estadísticos estos valores suelen aparecer tachados con una X e indicar de manera habitual el número de observaciones con valores atípicos. En la siguiente grafica de bigotes se presenta un ejemplo similar al anterior, en el que se analiza la variable de "Edad" con el registro de diferentes pacientes. En esta se logran observar dos valores atípicos señalados en ambos extremos de la gráfica, mediante dos puntos (7).

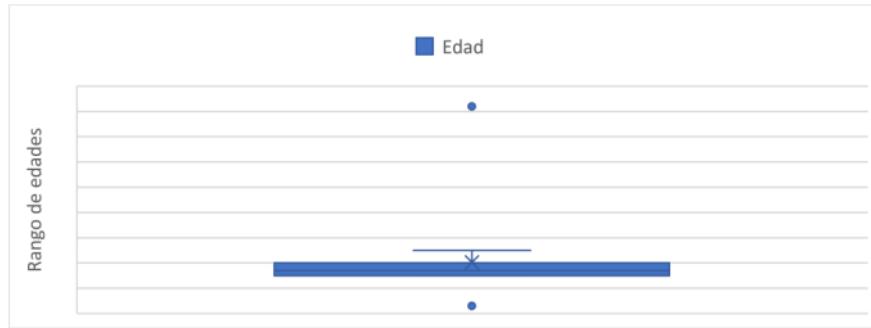


Diagrama 3. Ejemplo de Caja de Bigotes con valores atípicos.

1.11.1. Identificación con medidas centrales y de dispersión.

Existen diferentes métodos para identificar valores atípicos o fuera de rango mediante herramientas de análisis estadístico. Por ejemplo, para estimar el centro de los valores en c⁴⁶ la variable podemos utilizar el cálculo de las medidas centrales como la media, la moda y la mediana. Otras son las medidas de dispersión, como la varianza, la desviación estándar, el coeficiente de variación y el rango, las cuales indican en ¹¹⁷ qué lugar del espacio se encuentran determinadas características de una variable. A continuación, se muestra una breve descripción de cada una de estas técnicas de análisis estadístico (6):

- **Media:** Corresponde a la medida media de la suma de los datos. Para el cálculo de esta medida, basta con sumar los valores de todos los datos y dividir el resultado por el tamaño de la muestra. La media de la muestra está señalada por la siguiente expresión matemática (6).

$$\bar{X} = \frac{\sum X}{n}$$

Donde n representa al tamaño de la muestra.

- **Mediana:** Corresponde al valor de campo que se encuentra en el centro cuando estos se clasifican en orden ascendente. A diferencia de la media, la mediana se comporta estable ante la presencia de valores atípicos, razón por

la cual, dependiendo del comportamiento de los datos, esta puede ser una alternativa para análisis central de los datos (6).

- **Moda:** Figura el valor de campo que se repite con más frecuencia y puede utilizarse tanto para datos numéricos como para categóricos. Esta puede ser otra alternativa para análisis central de los datos, aunque no siempre se asocia al centro de la variable (6).
- **Rango:** Es el intervalo que existe entre el valor mínimo y el valor máximo de la distribución de frecuencias. Es la medida de dispersión de menor uso, ya que cualquiera de los dos límites puede encontrarse muy distante de un grupo compacto de valores. Su expresión matemática es la siguiente (6):

$$R = \text{Max} - \text{Min}$$

171

- **Varianza:** Representa que tanto se alejan los valores de la muestra respecto de la media. Tiene aplicación en el estudio de la aceptación o rechazo de la hipótesis. Esta denotada por la siguiente expresión matemática (6):

$$\sigma^2 = \frac{\sum(X - \bar{X})^2}{N}$$

Donde N corresponde al tamaño de la muestra.

- **Desviación estándar:** Se interpreta como la distancia típica entre un valor de campo y la media. La mayoría de los valores se encuentran en dos desviaciones⁴¹ estándar de la media. Esta medida se obtiene mediante el cálculo de la raíz cuadrada de la varianza. Es sensible a la presencia de valores atípicos: La desviación estándar se define, ya sea para una muestra o para una población, mediante la siguiente expresión matemática respectivamente (6):

$$S = \sqrt{\frac{\sum(X - \bar{X})^2}{N}} \quad \sigma = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}}$$

Donde N corresponde al tamaño muestral y n se refiere al tamaño poblacional.

142

- **Coeficiente de variación:** Mide la variabilidad relativa entre la desviación estándar y la media. Esta medida resulta útil ya que relaciona las variaciones con la media de los valores que estudia y permite comparar conjuntos de datos pertenecientes a distintas poblaciones. Esta medida se expresa matemáticamente de la siguiente manera (6):

$$CV = \frac{S_x}{|\bar{X}|}$$

Donde S_X corresponde a la desviación estándar del conjunto de datos y $|\bar{X}|$ es el valor absoluto de la media del conjunto de datos $(X_1, X_2, X_3, \dots, X_n)$ cuando $\bar{X} \neq 0$.

1.12. Tratamiento de datos faltantes.

En muchas ocasiones nos encontramos con bases de datos que contienen algunos registros con campos faltantes o sin información. El tratamiento de información faltante constituye una de las principales tareas previas al análisis de los datos. En muchas ocasiones la presencia de registros con información faltante puede ser el resultado de un registro defectuoso, la ausencia natural de la información o a una falta de respuesta ya sea total o parcial (7).

En estas situaciones, la primera prueba consiste en comprobar si la distribución de estas incidencias ocurre de manera aleatoria y en todo el conjunto de datos. Para ello en primera instancia se puede para el caso de una única variable "Y" consiste en determinar dos grupos de valores para dicha variable, uno de ellos con los datos con campos ausentes y otro para los no ausentes. Después, para cada variable "X" distinta de "Y", se realiza una prueba para determinar si existen diferencias marcadas entre los dos grupos de valores determinados por la variable "Y" sobre los valores de "X" (7).

Otro de los métodos más empleados para determinar la aleatoriedad de los datos faltantes es a través de la prueba de correlaciones dicotomizadas. La forma para realizar esta prueba es mediante la creación de una nueva variable dicotomizada para cada una de las variables "Y", asignando en ellas el valor de cero a los datos faltantes y el valor de 1 a los espacios con presencia de datos. El siguiente paso es generar la matriz de correlación junto a los contrastes de significancia de cada coeficiente de correlación. Con esto se pretende determinar el grado de asociación entre los valores faltantes con cada par de variables para con ello inferir si los elementos de la matriz de correlaciones son significativos o no. Cuando no hay significancia de correlación en los datos faltantes se puede deducir que estos son en su totalidad aleatorios (7).

Así, como los ejemplos anteriores, existen una variedad de otros métodos estadísticos más que nos permiten determinar la aleatoriedad o el comportamiento de los datos. Muchas veces dependerá del propio comportamiento, que herramienta será la más adecuada para su tratamiento.

Un procedimiento común para el tratamiento de valores perdidos es simplemente omitir en el análisis los registros o campos con valores faltantes, a esto se le conoce como el método de aproximación de casos completos. Sin embargo, en ocasiones esto podría ser un error ya que el patrón de valores en los campos faltantes podría ser sistemático, conduciendo a un subconjunto de datos sesgado. Además, muchas

veces la información contenida en los demás campos del registro podría ser muy valiosa como para eliminarla solo por tener otro campo vacío (6).

46

Con la finalidad de rescatar la mayor cantidad de información posible sin eliminar registros con campos vacíos, se puede optar por sustituir el valor faltante por alguna constante para no afectar los demás campos del registro. En el caso de algunas variables podríamos elegir sustituir los campos faltantes con valores estadísticos como la moda o la media, obtenidos de los demás registros en los que si se cuenta con algún valor o sustituirlos por un valor generado de forma aleatoria a partir de la distribución observada en la variable. También podemos analizar las características de los demás campos de las variables y formular un juicio para darle algún valor a los campos faltantes.

En la tabla 4 se presenta un ejemplo en el que falta uno de los registros referentes a la temperatura del paciente, si en este caso el principal motivo de nuestro análisis es la temperatura, podríamos realizar un examen de los demás datos para determinar que podemos hacer para resolver el problema. Por un lado, dependiendo de la cantidad de registros faltantes podríamos optar por eliminar el registro completo y con ello ahorrar el tiempo que implicaría realizar un análisis más profundo.

| ID | Sexo | PCR Covid19 | Temperatura | Lugar | Fecha |
|-----|------|-------------|-------------|---------|------------|
| 001 | M | Positivo | 37.5°C | México | 20/10/2021 |
| 002 | F | Positivo | 38.0°C | EE. UU. | 20/10/2021 |
| 003 | F | Negativo | 36.5°C | México | 20/10/2021 |
| 004 | M | Positivo | 38.0°C | México | 20/10/2021 |
| 005 | M | Negativo | | EE. UU. | 20/10/2021 |
| 006 | F | Positivo | 38.5°C | México | 20/10/2021 |
| 007 | M | positivo | 37.5°C | México | 20/10/2021 |

Tabla 4. Ejemplo de datos faltantes.

181

Sin embargo, suponiendo que en nuestra base de datos una gran cantidad de registros presentan el mismo problema, podríamos optar por resolverlo realizando un análisis estadístico para así completar los datos faltantes con el valor de la media obtenida de los registros que presentan el mismo patrón.¹⁰⁵ En el ejemplo anterior, el registro faltante corresponde a un paciente con resultado negativo en la prueba PCR de Covid19. En este contexto, podríamos realizar el análisis estadístico de los demás pacientes con resultado negativo en la prueba PCR para con ello determinar alguna técnica de análisis estadístico para completar los registros de temperatura faltantes.

1

1.13. Imputación de la información faltante.

¹⁸⁸ Método de aproximación de casos completos suele ser la técnica más común en el tratamiento de los datos faltantes y es apropiado siempre y cuando no haya demasiados valores perdidos ya que de lo contrario esto provocaría una reducción bastante significativa de la muestra y perdería la representatividad de la información completa (7)..

La imputación es una alternativa en el tratamiento de los datos faltantes. Es un proceso en el que se realiza una estimación de valores ausentes con base en valores validos obtenidos de otras variables u otros argumentos de la muestra. En este proceso se pueden realizar, por un lado, diferentes técnicas de análisis estadísticos y por otro, metodologías como la sustitución ¹el caso, la sustitución por un valor constante, la i¹⁹⁶polación lineal, la estimación por regresión y el método de imputación múltiple. A continuación, se presenta una breve descripción de dichas técnicas (7):

- **Enfoque de disponibilidad completa:** En este proceso no se reemplazan los datos ausentes si no que se utilizan las características de la distribución o las correlaciones de los demás valores validos disponibles de la muestra (7)..
- **Sustitución por la media:** Consiste en obtener la media de los demás valores en determinada variable y reemplazar los datos faltantes con valor resultante. Este método proporciona información importante y es fácil de implementar. Su principal desventaja radica en que modifica las correlaciones e invalida las estimaciones de la varianza. Una variante de este método, para el caso en que existen valores extremos en la variable y consiste en sustituir los valores por la mediana en lugar de la media (7)..

En este método, cuando existe demasiada variabilidad ¹en los datos, resulta conveniente sustituir en valor de los datos faltantes con la media o la mediana de cierto número de observaciones adyacentes.

- **Sustitución por interpolación:** Este método consiste en sustituir cada valor ausente de la variable por el valor resultante de la interpolación con valores adyacentes (7).
- **Sustitución por una constante:** Para este método es necesario realizar una investigación previa. Consiste en sustituir los valores ausentes por una constante obtenida de una fuente externa, pues derivado de la investigación, se determina que dicha constante resulta más valida que el valor de la media o la mediana de los datos validos en la muestra (7).
- **Imputación por regresión:** Consiste en predecir los valores ausentes de una variable, a través de su correlación con otras variables de la muestra, mediante la ecuación de regresión que las relaciona. Para optar por este método es necesario que la correlación de la variable con datos ausentes y las demás variables sea sustancial. Este método desestima la varianza de la distribución y refuerza la relación existente en los datos, lo que supone una desventaja ya que entre más se utiliza, los datos resultantes son menos generalizables en el análisis (7).

- **Imputación múltiple:** Dependiendo del comportamiento de los datos en determinadas variables resulta conveniente utilizar una combinación de varios de los métodos anteriores (7).

En esta parte de la minería de datos, el objetivo principal es rescatar la mayor cantidad de información posible. Para ello, es necesario formularse un criterio y encontrar el método adecuado para reemplazar el campo vacío con determinado valor.

1.14. Transformación de datos.

En ocasiones, dependiendo del análisis que se realiza, puede ser necesario realizar una transformación a los datos, ya sea con la intención de obtener una mejor respuesta al momento de aplicar determinado modelo, para modificar su dimensionalidad y evitar problemas que podrían derivar de bases de datos demasiado extensas, entre otras. Durante el proceso de minería de los datos podemos encontrarnos con un gran número de situaciones que podrían obligarnos a realizarle ciertas modificaciones o transformaciones a los datos con la finalidad de aprovechar ciertas características que podrían aportar una mejora al momento de análisis (7).

Para este proceso, podemos clasificar la transformación de los datos en las siguientes categorías: (7).

- **Transformaciones lógicas:** Podemos reducir la amplitud o modificar el campo de definición de las variables para con ello eliminar categorías sin respuestas o convertir variables de intervalo en nominales mediante la creación de variables ficticias a las que se les califica como variables dummy (7).
- **Transformaciones ¹²⁹lineales:** Se obtienen mediante la aplicación de operaciones lógicas como la suma, la resta, la multiplicación o la división para con ello mejorar su interpretación. Este tipo de transformaciones no modifican la forma de la distribución ni el orden de las variables (7).
- **Transformaciones algebraicas:** Se obtienen mediante la aplicación de operaciones lógicas no lineales como la raíz cuadrada y los logaritmos, entre otras, para con ello ¹²¹mejorar la interpretación de los datos. Esta transformación cambia la forma de la distribución de los datos, pero mantiene el orden de los registros (7).
- **Transformaciones no lineales:** En esta se cambia la distancia y el orden entre los valores. Por ello se debe ser cuidadoso y evaluar el beneficio que se obtiene al realizarlas ya que pueden llegar a modificar demasiado la información original (7).

- **Transposición, fusión, adición, segmentación y orden:** Dependiendo del análisis, es posible realizar transformaciones mediante modificaciones a la base de datos. Podemos transponer las filas y las columnas, fusionar archivos entre diferentes bases de datos ya sea con las mismas variables, pero diferentes casos o variables diferentes aplicadas a los mismos casos, también agregar datos, por ejemplo, datos obtenidos mediante el análisis estadístico de ciertas variables o mediante la agrupación de dos o mas variables, de igual forma podemos dividir o segmentar un archivo en distintos grupos (7).
- **Reducción de la dimensión:** Se trata de eliminar la redundancia en la información, derivada de un tamaño excesivo en la dimensión de la muestra. Mediante estos métodos se combinan muchas variables para obtener unas cuantas variables ficticias que proporcionen una tendencia real en el comportamiento de los datos (7).

2. APRENDIZAJE AUTOMÁTICO.

161

El aprendizaje automático radica en un conjunto de métodos que permiten detectar de manera automática patrones en los datos. Estas técnicas se encuentran estrechamente relacionadas con otras ciencias como la minería de datos y las estadísticas. Su principal objetivo consiste en el desarrollo de modelos que permiten la obtención de patrones mediante el análisis de grandes cantidades de datos, para con ello lograr predecir datos futuros o entender el comportamiento de dicha información y como es que está se relaciona entre sí, y así poder tomar decisiones en condiciones de incertidumbre (12).

2.1. Tipos de aprendizaje automático.

El aprendizaje automático se clasifica en dos enfoques principales, uno de aprendizaje predictivo o supervisado y el otro de aprendizaje descriptivo o no supervisado (12):

2.1.1. Aprendizaje supervisado.

En este enfoque, el objetivo consiste en realizar un análisis de las entradas (X) respecto de las salidas (Y) a través de un conjunto de datos de entrenamiento. Cada entrada de entrenamiento X_i puede ser desde un vector de n dimensiones u objetos estructurados más complejos como imágenes, frases, formas moleculares, gráficos, mensajes de correo electrónico, etc. Así también, la forma de la variable de salida puede ser cualquier cosa, sin embargo, se asume que Y_i podría ser una variable categórica nominal o un escalar de valor real. Cuando la variable de salida Y_i es categórica se trata de un problema de clasificación o reconocimiento de patrones y cuando se trata de un escalar de valor real el problema es de regresión (12). En otras palabras, en el aprendizaje supervisado encontramos un conjunto de datos con etiquetas que proporcionan información que podemos utilizar para predecir datos no etiquetados (13).

2.1.2. Aprendizaje no supervisado.

El principal objetivo en este enfoque es encontrar patrones interesantes en los datos. A diferencia del aprendizaje supervisado, en este no sabemos cuál es el resultado deseado para cada entrada. Uno de los métodos principales en el aprendizaje no supervisado se encuentra el agrupamiento de datos (12).

128 Los algoritmos de aprendizaje no supervisado se utilizan para datos no etiquetados con la finalidad de encontrar relaciones entre ellos (13).

2.2. Tipos de soluciones en el aprendizaje automático.

El aprendizaje automático consta de un conjunto de algoritmos y técnicas con una base matemática y estadística sólida para el diseño de sistemas computacionales que aprenden de los datos y que ofrecen distintas soluciones a diferentes tipos de problemas, los cuales se pueden clasificar en tres tipos principales (13):

- **Clasificación:** Consiste en identificar a qué categoría pertenece una nueva observación basándose en el conjunto de datos de entrenamiento (13).

Un problema de clasificación puede ser de dos clases o multiclase y su resultado puede ser un valor discreto que indica la clase predicha en la que se encuentra cierta observación o un valor continuo que indica la probabilidad de una observación pertenezca a una clase en concreto (13).

En este tipo de aprendizaje, el objetivo es realizar un análisis de las variables X con respecto a las salidas Y cuando $Y \in \{1, \dots, C\}$, donde C corresponde al numero de clases. Si $C = 2$, el tipo de clasificación es binaria en cuyo caso podemos asumir que $Y \in \{0,1\}$. Cuando $C > 2$ se trata de una clasificación multiclase (12).

Una manera de visualizar este tipo de problemas es mediante una aproximación de funciones. Suponiendo que $Y = f(X)$, el objetivo del aprendizaje consiste en estimar la función f , mediante un conjunto de datos de entrenamiento para después realizar predicciones utilizando $\hat{Y} = \hat{f}(X)$. A esto se le llama generalización (12).

- **Regresión:** Consiste en predecir de forma estimada el futuro mediante la relación entre las variables (13).

Al igual que en la clasificación, en el aprendizaje por regresión, el objetivo es realizar un análisis de las variables X con respecto a las salidas Y , solo que en este caso la variable de respuesta Y es continua (12).

- **Agrupamiento de datos:** Permite agrupar puntos de datos similares de forma intuitiva y ayuda a descubrir cómo están organizados mediante su agrupación en conjuntos naturales. Es muy útil cuando se desea encontrar un patrón específico en el comportamiento de los datos (13).

10

La esencia del agrupamiento consiste en dividir un conjunto de datos en grupos que sean lo más parecido posible. Dependiendo del modelo específico utilizado, las variaciones en la definición del problema a resolver mediante agrupamiento pueden ser bastante significativas. Por ejemplo, un modelo generativo puede definir la similitud sobre la base de un mecanismo generativo probabilístico, mientras que un enfoque basado en la distancia

utilizara una función de distancia tradicional para la cuantificación. Además, el tipo de datos específico también tiene un impacto significativo en la definición del problema (9).

2.3. Clasificación mediante Regresión Logística.

123

Esta técnica se utiliza para resolver problemas en los que se asigna el mismo peso a todos los datos, 1 para los positivos ³ y 0 para los negativos, ya que no se puede imponer la restricción de regresión, la técnica más adecuada para resolver este tipo de problemas es la regresión logística, que consiste en asignar un valor arbitrariamente grande o pequeño a los datos en función de la distancia a la frontera de decisión, permitiendo una mayor precisión en la obtención del resultado (14).

3

Con este contexto, el resultado de la regresión logística es la probabilidad de que un punto de entrada determinado pertenezca a una clase específica. La salida de la regresión logística siempre se encuentra en [0,1] (13).

3

Este tipo de problema puede verse como un caso particular de regresión no lineal, en el que el valor objetivo (probabilidad de pertenecer o no a la clase de cada dato) se transforma a través de la función logística de la siguiente manera (14):

$$\log \frac{\pi_i(\vec{x})}{1 - \pi_i(\vec{x})}; \quad \pi_i(\vec{x}) = \begin{cases} 1: \vec{x} \in C_i \\ 0: \vec{x} \notin C_i \end{cases}$$

3

De este modo, podemos generar valores arbitrariamente grandes en función de su separación de la frontera de decisión.

Para realizar las predicciones de probabilidad, se utiliza la función sigmoide (la inversa de la función logística) (14).

$$\phi(z) = \frac{1}{1 + e^{-z}} \quad z = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_m x_m = \sum_{i=0}^m \theta_i x_i = \theta^t x$$

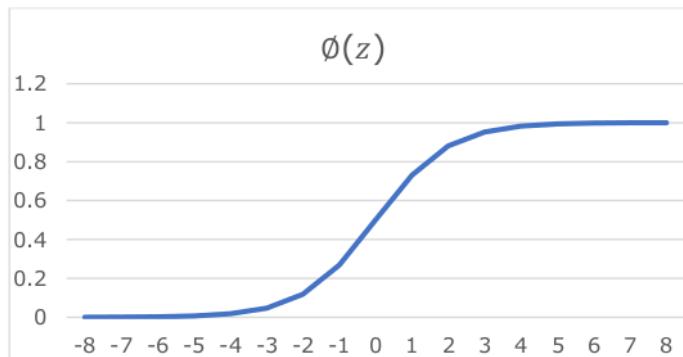


Diagrama 4. Función Sigmóide.

Para trabajar con el modelo de regresión logística, debemos considerar lo siguiente: m es el número de conjuntos de entrenamiento, X_i son las variables de entrada, Y_i son las variables por predecir y Z representa el hiperplano límite de decisión. Basándonos en lo anterior, podemos formular nuestra hipótesis utilizando la función sigmoide de la siguiente manera (14):

| Y | $h_0(X)$ |
|-----|-------------------|
| 1 | $h_0(X) \geq 0.5$ |
| 0 | $h_0(X) < 0.5$ |

Tabla 5. Formulación de la hipótesis, Regresión Logística.

$$h_0(x) = \phi(\theta^t x)$$

2.3.1. Razón de probabilidades (Odds Ratio).

Estas se definen como la relación entre la probabilidad de éxito con la probabilidad de fracaso y podemos definirla de la siguiente manera (13):

$$\frac{\text{Posibilidad de que suceda}}{\text{Posibilidad de que no suceda}} = \frac{P}{1 - P}$$

Uno de los casos más sencillos para exemplificar la razón de probabilidad es el de la moneda. La razón de probabilidad de que al lanzar una moneda obtengas cara es de 1. Esto se debe a que hay una probabilidad del 0.5 de obtener cara y una probabilidad de 0.5 de obtener cruz. En otras palabras, existe un 50% de probabilidad de obtener cara y un 50% de obtener cruz. Sin embargo, si la moneda estuviera manipulada de tal forma de que la probabilidad de obtener cara fuera de 0.8 y de 0.2 de obtener cruz, entonces la razón de probabilidad de que salga cara es de $0.8/0.2 = 4$. Es decir, es cuatro veces más probable obtener cara en lugar de cruz (13).

2.3.2. Función Logit.

Al aplicar la función de logaritmo natural a la razón de probabilidad se obtiene la función logit. Esta transfiere una variable en el rango de $[1,0]$ a una nueva en el rango de $[-\infty, \infty]$. Para ver esta relación podemos utilizar la siguiente expresión matemática:

$$L = \ln \left(\frac{P}{1 - P} \right)$$

19

3. MINERÍA DE DATOS Y APRENDIZAJE AUTOMÁTICO EN EL CUIDADO DE LA SALUD.²¹²

En el campo de la medicina se generan diariamente grandes cantidades de datos. Antes de la llegada de los ordenadores, estos datos eran registrados principalmente en papel. Hoy en día, con la rápida tendencia a la digitalización y con el avance y aparición de nuevas tecnologías, se generan cantidades masivas de datos (Big Data). La cantidad del conjunto de datos digitales ¹⁸⁹ salud es tan grande y compleja que resulta difícil su procesamiento mediante herramientas y métodos de gestión de datos comunes. La cantidad de datos sobre salud hoy en día es abrumadora, no solo por su volumen, sino también por la diversidad en los datos y la rapidez con la que deben ser gestionados.

A través del análisis y la minería de grandes cantidades datos, las organizaciones sanitarias pueden obtener diversos beneficios como la gestión eficaz de la salud de una población, la detección de fraudes en la atención médica y la gestión de los recursos hospitalarios, entre otros. Uno de los principales es la detección temprana de enfermedades, lo que se traduce en un tratamiento más eficaz y a la vez en la prevención de complicaciones que podrían desarrollarse si la enfermedad no se detecta a tiempo (5).

3.1. Ventajas del aprendizaje automático en el cuidado de la salud.

Derivado de la digitalización de la información, hoy en día contamos con grandes cantidades de datos, con los cuales, mediante una adecuada clasificación y con el apoyo de diferentes métodos y herramientas de aprendizaje automático, podemos obtener una variedad de ventajas en diferentes áreas de atención a la salud. Una de las principales áreas de oportunidad es la detección temprana de enfermedades y sus complicaciones. Gracias al análisis de bases de datos históricas con grandes cantidades de información, podemos predecir eventos que podrían conducir a una mejor atención de los pacientes con distintas enfermedades y con ello ofrecerles determinadas herramientas que les ayuden a llevar un control adecuado de su enfermedad evitando así complicaciones (5).

Aunque la detección temprana de enfermedades es uno de los principales objetivos abordados en la minería de datos a gran escala, este representa solo una pequeña parte de las grandes ventajas que podemos obtener en la atención de la salud. A continuación, se muestran algunas de las ¹⁷³ aplicaciones de la minería de datos a gran escala y el aprendizaje automático en el ámbito de la atención a la salud (5):

8

- **Operaciones clínicas.** Mediante la clasificación y análisis de datos a gran escala, se pueden efectuar investigaciones para realizar una comparación entre los tratamientos médicos ¹³⁷ existentes con el fin de determinar procedimientos ²¹³s relevantes tanto desde el punto de vista clínico como de rentabilidad en el diagnóstico y tratamiento de los pacientes (5).

- **Investigación y desarrollo.** Mediante distintos modelos de predicción y clasificación, se pueden mejorar los ensayos clínicos para catalogar a los pacientes con la finalidad de ajustar los tratamientos de forma personalizada, reduciendo con ello las fallas en los ensayos mediante el desarrollo de nuevos y mejores tratamientos médicos. Además, se puede realizar un análisis de los ensayos e historiales clínicos para identificar indicadores de efectos adversos en los medicamentos antes de su aplicación o salida al mercado (5).
- **Salud pública.** Con el análisis de datos a gran escala sobre enfermedades, se pueden identificar patrones para llevar un seguimiento de los brotes y su transmisión mejorando la vigilancia de la salud pública logrando una respuesta mucho más rápida. También contribuye al rápido y eficaz desarrollo de vacunas. Además, con la posibilidad de procesar grandes cantidades de datos se logra obtener información importante que puede utilizarse para determinar las necesidades, prestar servicios, predecir situaciones y prevenir crisis en materia de salud pública (5).
- **Medicina basada en evidencia.** Al combinar y analizar datos financieros, operativos, clínicos y genómicos, estructurados y no estructurados, se puede lograr la estandarización de los tratamientos, predecir los pacientes con riesgo de enfermedad o complicaciones y ofrecer una atención médica más eficiente (5).
- **Análisis genómico.** La secuenciación de genes puede ser más eficiente y rentable. El análisis genómico pasa a formar parte del proceso normal en la toma de decisiones para la atención del paciente y el manejo del historial clínico (5).
- **Análisis para la prevención de fraudes.** Gracias a estas tecnologías es posible analizar un gran número de incidencias que aportan información importante para la prevención de fraudes, la reducción de gastos innecesarios y la disminución de abusos (5).
- **Monitoreo remoto.** Pueden recogerse grandes cantidades de datos para su análisis en tiempo real y a distancia. Los sistemas de monitorización también pueden ser diseñados para la seguridad y la predicción de eventos adversos (5).
- **Análisis del perfil del paciente.** Se pueden realizar modelos avanzados de análisis en los perfiles de pacientes para identificar indicadores de enfermedades o sus complicaciones. Con el análisis de grandes cantidades de datos, se podemos realizar modelos predictivos o de clasificación para la detección temprana de enfermedades o complicaciones y así disponer de mejores tratamientos preventivos (5).

A continuación, se muestra un diagrama sobre la influencia de las distintas técnicas de aprendizaje automático aplicado a las ciencias para el cuidado de la salud.



Diagrama 5. Aprendizaje automático en el cuidado de la salud.

3.2. Las 4 “V” del análisis de datos y el cuidado de la salud.

Al igual que en otros ámbitos, el análisis relacionado a grandes cantidades de datos (Big Data) para la atención de la salud se describe a través de cuatro características principales: volumen, velocidad, variedad y veracidad (5).

- **Volumen.** Con la llegada de las nuevas tecnologías, los datos relacionados con la salud han crecido y se han acumulado de forma exponencial. Estos, siguen generando día a día una cantidad incalculable de información, en las que se incluyen historias clínicas, imágenes médicas, datos de ensayos clínicos, de secuencias genéticas, lecturas de sensores biométricos y datos personales de pacientes, entre otros. Además, conforme surgen nuevas tecnologías, estas proporcionan cada vez más y más información (5).
- **Velocidad.** Los avances en la gestión ¹³⁵ de datos, como la virtualización y la computación en la nube, facilitan el desarrollo de plataformas para capturar, almacenar y manipular datos en tiempo real y a gran escala con un flujo constante y a un ritmo sin precedentes (5).
- **Variedad.** Con la capacidad de realizar análisis en tiempo real de volúmenes tan grandes de datos que están en constante movimiento, hoy en día podemos generar muchas aplicaciones que proporcionan una gran cantidad de soluciones en diferentes áreas para la atención de la salud (5).
- **Veracidad.** La precisión en las arquitecturas, herramientas, plataformas, algoritmos y metodologías para el análisis y la extracción de datos a ¹³⁹ escala en la atención de la salud es de vital importancia. La seguridad y la vida de los pacientes dependen de que se disponga de información precisa, ya que la calidad de los datos, especialmente los no estructurados, es muy variable y a menudo puede ser incorrecta (5).

55 4. DIABETES MELLITUS.

La diabetes mellitus es una enfermedad crónica en la que el páncreas, el órgano que produce la insulina, hormona que regula los niveles de azúcar en la sangre no produce suficiente o no la utiliza eficazmente, lo que provoca un aumento significativo de los niveles de azúcar en la sangre (hiperglucemia), que con el tiempo puede dañar gravemente muchos órganos y sistemas de las personas que padecen la enfermedad (1).

Existen dos tipos principales de diabetes, la diabetes tipo 1 y la diabetes tipo 2. La primera caracteriza porque el páncreas produce poco o nada de insulina por sí mismo. La diabetes tipo 2 es la más común y representa del 85 al 90 por ciento de los casos, presentándose principalmente en adultos y se vincula con factores de riesgo relacionados con los hábitos del paciente, por lo que puede ser prevenida. También podemos clasificar otro tipo de diabetes, la diabetes gestacional que puede llegar a presentarse en algunas mujeres durante el embarazo (2).

72 4.1. Tipos de diabetes.

La diabetes mellitus es un síndrome de hiperglucemia que llega a presentarse en las personas por diferentes causas. En sus clasificaciones, la diabetes tipo 1 representa entre el 5 y el 10 por ciento de los nuevos casos de pacientes diagnosticados con diabetes, mientras que la diabetes tipo 2 representa entre el 90 y el 95 por ciento. Una de las principales diferencias entre estas dos, es que en la diabetes tipo 1 suele haber una eliminación completa o casi total de las reservas de insulina, mediada únicamente por la respuesta inmunogénica de los portadores de cierto genotipo, mientras que la diabetes tipo 2 es de origen poligenético y puede aparecer derivado de factores ambientales como una mala alimentación y un estilo de vida sedentario entre otros hábitos de las personas (15).

114 4.1.1. Diabetes Tipo 1.

Se caracteriza por una producción deficiente de insulina. Por ello, los pacientes con este tipo de diabetes necesitan mantener un control diario de los niveles de glucosa en la sangre y suministrarse insulina por vía intramuscular en el organismo. En la actualidad, derivado de la naturaleza por la cual surge este tipo de diabetes en las personas, no es posible prevenirla mediante la tecnología y conocimientos sobre la enfermedad, con los que contamos hasta ahora (1).

218 Los principales síntomas de esta enfermedad incluyen una excreción urinaria excesiva, sed y hambre constante, pérdida de peso, alteraciones visuales y cansancio y pueden aparecer de manera repentina (1).

4.1.2. Diabetes tipo 2.

Se caracteriza porque, aunque el páncreas tiene la capacidad de producir insulina, 165 organismo es incapaz de utilizar esta hormona de forma eficaz. En gran medida, la causa de la diabetes de tipo 2 es la inactividad y el sobrepeso de las personas, por lo que se puede prevenir (2).

7

Los síntomas pueden ser similares a los de la diabetes de tipo 1, pero menos intensos, por lo que se detecta regularmente una vez que han aparecido las complicaciones de la enfermedad (1).

156

En la siguiente tabla se hace una comparación general entre los tipos 1 y 2 de la diabetes mellitus (16):

| Característica | Diabetes Tipo 1 | Diabetes Tipo 2 |
|----------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Naturaleza de la enfermedad. | Trastorno autoinmune caracterizado por la destrucción de las células beta productoras de insulina. | Trastorno de deficiencia de insulina por insuficiencia del páncreas. |
| Inicio. | Inicio rápido. Los síntomas comienzan de manera repentina. (En días o semanas) | Inicio leve a moderado. Los síntomas comienzan más lentos. (Durante meses o años) |
| Síntomas. | Aumentan los niveles de glucosa en la sangre en niveles muy altos. Polifagia, polidipsia, poliuria, cetoacidosis. | Aumentan los niveles de glucosa en la sangre en niveles que van de moderados a altos. Polidipsia y poliuria leves, fatiga, dolor de cabeza. |
| Factores de riesgo. | Antecedentes familiares de enfermedades autoinmunes, principalmente diabetes tipo 1. (Riesgo 10 veces superior al de la población general). | Se relaciona más con factores ambientales como los hábitos. Sobrepeso, obesidad, mala alimentación, sedentarismo. Antecedentes familiares de diabetes tipo 2 o gestacional. Étnico, (Mayor riesgo en africanos e hispanos). |
| Edad en que suele desarrollarse. | Típicamente desde la vida temprana hasta la adolescencia aun que puede presentarse a cualquier edad. | Por lo general aparece en la adultez, sin embargo, cada vez tiende a iniciar en edades más tempranas. |
| Plan tratamiento. | Dependencia 25 psoluta de insulina. Modificaciones en el estilo de vida. (Llevar un plan de alimentación y ejercicio, modificar hábitos nocivos, etc.) | Llevar un adecuado plan alimenticio y de ejercicio en conjunto con un tratamiento con medicamentos orales. Modificar hábitos nocivos. (Un porcentaje cada vez mayor de pacientes llega a requerir el suministro de insulina con el tiempo). |
| Forma de prevenir. | No existe en la actualidad. Los casos futuros pueden predecirse mediante diferentes métodos de análisis genético. (Objeto de grandes esfuerzos de investigación). | Para la mayoría de los casos potenciales, mediante modificaciones en los hábitos nocivos. Llevar a cabo un adecuado plan de alimentación y ejercicio. |
| Forma revertirlo. | No existe en la actualidad. (Objeto de grandes esfuerzos de investigación). | No existe en la actualidad. 72 Sin embargo, los pacientes pueden controlar la enfermedad y reducir el riesgo de complicaciones a través 1 modificaciones en los hábitos y mediante un plan de alimentación y ejercicio adecuado. |
| Complicaciones. | Emergencias agudas de hipoglucemia y cetoacidosis que conducen a la inconsciencia hipoglucémica. Los efectos crónicos de la | Emergencias agudas de hipoglucemia y cetoacidosis que conducen a la inconsciencia hipoglucémica. Los efectos crónicos de la |

| | | |
|--|------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| | hiperglucemia pueden conducir a retinopatía, nefropatía, neuropatía, enfermedad cardiovascular, etc. | hiperglucemia pueden conducir a retinopatía, nefropatía, neuropatía, enfermedad cardiovascular, etc. |
|--|------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|

42

Tabla 6. Comparativa entre características de diabetes tipo 1 y diabetes tipo 2.

42

También podemos comparar mediante la siguiente tabla, otras características entre la diabetes tipo 1 y la diabetes tipo 2 (17):

| Característica | Diabetes tipo 1 | Diabetes tipo 2 |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Frecuencia. | 5% – 10% | 90% - 95% |
| Edades. | Comúnmente se presenta en niños y jóvenes adolescentes aun que puede presentarse a cualquier edad. | Generalmente se presenta a edades avanzadas, aunque también puede presentarse en niños y jóvenes adolescentes. |
| Patogénesis. | 133 generalmente autoinmune. Se caracteriza por la destrucción de las células beta pancreáticas. | No hay 146 inmunidad. Existe una resistencia a la insulina que genera deficiencia progresiva de la misma. |
| Niveles de péptido C. | Muy bajos o indetectables. | Detectables. |
| Diagnóstico primero. | Los autoanticuerpos (GAD65, IA-2, IAA, ZnT8) pueden estar presentes. | Ausencia de autoanticuerpos. |
| Terapia medicación. | Necesidad absoluta de suministro de insulina. Pueden ser múltiples inyecciones diarias o mediante una bomba de insulina. | Agentes orales y/o fármacos hipoglucemiantes inyectables no insulínicos. En ocasiones puede ser necesario el suministro de insulina. |
| Modo prevención. | Ninguna conocida Ensayos clínicos en curso. | Estilo de vida (adecuada alimentación y actividad física constante. Control del peso corporal) Los medicamentos orales (metformina, acarbosa) pueden ser útiles cuando se cuenta con un diagnóstico de prediabetes. |

31

Tabla 7. Comparativa entre características de diabetes tipo 1 y diabetes tipo 2.

4.2. Métodos de diagnóstico.

31 la actualidad pueden ser utilizadas las siguientes pruebas para la detección y diagnóstico de la diabetes mellitus (15):

- **HbA1c:** La prueba de hemoglobina glucosilada se utiliza más comúnmente debido a la fiabilidad de los resultados y la facilidad con que puede ser realizada. Se considera un valor de HbA1c $\geq 6.5\%$ para realizar un diagnóstico positivo de diabetes (15).
- **FPG:** Para el análisis de glucosa en plasma sanguíneo es 34 necesario un ayuno de al menos 8 horas antes de ser realizada. Niveles $\geq 126 \text{ mg/dl}$ de glucosa [plasmática en ayunas pueden indicar un diagnóstico positivo de diabetes. Esta prueba debe ser realizada de nuevo al día siguiente para confirmar el diagnóstico (15).
- **PTGO:** La prueba de tolerancia a la glucosa se realiza mediante el suministro de una carga de glucosa de 75g generalmente. 2 horas después, se analizan 118

los niveles de glucosa plasmática, niveles ≥ 200 mg/dl de glucosa plasmática son un indicador para diagnosticar la diabetes mellitus. (En mujeres embarazadas, el suministro de la carga de glucosa generalmente es de 100g) (15).

168

- Síntomas de hiperglucemia como poliuria, polidipsia y pérdida de peso de forma repentina también son indicativos de diabetes mellitus. Si el paciente presente síntomas de hiperglucemia y se le realiza un análisis casual o de forma aleatoria de glucosa plasmática y presenta niveles ≥ 200 mg/dl puede ser diagnosticado con diabetes mellitus (15).

90

En ocasiones y según el criterio de cada médico, si una persona presenta niveles de glucosa en la sangre superiores a 195 mg/dl normales, pero no lo suficientemente altos como para cumplir con los criterios para el diagnóstico de diabetes mellitus, se le puede diagnosticar con “prediabetes”, para lo cual se le suele recomendar tomar medidas preventivas para controlar sus niveles de glucosa plasmática antes de que estos puedan resultar nocivos para su salud. Sin embargo, cabe señalar que el detectar niveles altos de glucosa puede ser un indicador de que el metabolismo del paciente comienza a presentar alteraciones en su funcionamiento por lo que se le deberá dar un seguimiento más estrecho al paciente para prevenir complicaciones.

4.3. Complicaciones agudas de la diabetes.

La diabetes está estrechamente asociada con complicaciones agudas que pueden dar lugar a alteración en gran parte de los órganos y sistemas de las personas, tales como accidentes cardiovasculares o cerebrovasculares y otras lesiones que pueden poner en riesgo la vida del paciente, si no se tratan de manera oportuna (18).

125

A continuación, se describen algunas de las principales complicaciones que se derivan de la diabetes, así como su tratamiento y forma de prevención (18):

4.3.1. Hipoglucemias.

7

Puede definirse como una concentración de glucosa en sangre venosa inferior a los 60 mg/dl. Está estrechamente relacionada al tratamiento farmacológico de la diabetes mediante medicamentos orales que ayudan al control de los niveles de glucosa en la sangre o el suministro de insulina. Cualquier paciente que se encuentra en tratamiento puede llegar a sufrirla, aunque es más común en paciente que presentan una larga evolución de la diabetes mellitus y que se encuentran en tratamiento intensivo con insulina (18).

80

La hipoglucemia puede presentarse de manera leve con síntomas como ansiedad, inquietud, taquicardia, palpitaciones, temblores, disminución en la capacidad de concentración, mareo, hambre o visión borrosa entre otros, pero que no intervienen en el desempeño normal de sus actividades. Sin embargo, también puede llegar a presentarse de manera moderada con síntomas como el deterioro evidente de la

función motora del paciente que le impide tratar los síntomas de manera autónoma, confusión o conducta inadecuada y graves como crisis convulsivas y coma (18).

Las principales complicaciones de la hipoglucemias³³ van desde accidentes cardiovasculares o cerebrovasculares agudos, hasta encefalopatía hipoglucémica o daño permanente en la corteza cerebral derivado de constantes episodios de hipoglucemias graves (18).

El principal tratamiento para la hipoglucemia consta de la ingesta de una o dos cucharadas de hidratos de carbono de absorción rápida (100 ml de bebida azucarada, 2 cucharadas de azúcar o miel). En pacientes tratados con inhibidores de las alfaglucosidasas (acarbosa/miglitol) asociados a sulfonilureas o insulina, se recomienda la administración de glucosa pura. La hipoglucemia grave secundaria a sulfonilureas requiere de observación estrecha del paciente siendo motivo de ingreso hospitalario y tratamiento posterior (18).

Con la finalidad de prevenir episodios de hipoglucemia grave, se le recomienda al paciente poner atención a cualquier síntoma de hipoglucemia leve para que lleve una oportuna atención. El autoanálisis de glucosa en sangre capilar y un adecuado control en los hábitos del paciente resulta la mejor protección frente a las complicaciones que se derivan de esta alteración en el metabolismo de los pacientes (18).

4.3.2. Hiperglucemia.

55

El déficit total o parcial de insulina puede ocasionar un incremento en los niveles de glucosa en la sangre a lo que se le denomina hiperglucemia y que puede derivar en diferentes complicaciones metabólicas para el paciente (18).

Una de las principales complicaciones de la hiperglucemia es la ketoacidosis diabética. Esta complicación puede desarrollarse principalmente en la diabetes tipo 1 aunque también se han presentado casos en pacientes con diabetes tipo 2, comúnmente cuando estos se encuentran en situaciones de estrés. Esta complicación se presenta principalmente cuando existe un déficit absoluto o relativo de insulina que genera concentraciones de glucosa en la sangre superiores a los 300 mg/dl. Entre los factores que desencadenan en una ketoacidosis más frecuentemente, destacan los procesos infecciosos y errores en la administración de insulina durante el tratamiento siendo más comúnmente derivado de la omisión de alguna dosis por parte del paciente o una prescripción inadecuada de una pauta en el tratamiento (18).

Mediante el trastorno metabólico subyacente, las manifestaciones clínicas asociadas a esta complicación resultan fáciles de reconocer para su diagnóstico. En una fase inicial el paciente puede presentar poliuria, polidipsia, perdida ponderal, astenia y anorexia. A medida que avanza el cuadro se pueden presentar náuseas,

vómitos, dolor abdominal, alteraciones de la conciencia y en un pequeño porcentaje de paciente se puede llegar a presentar un coma (18).

El diagnóstico de cetoacidosis diabética se trata con ingreso hospitalario. Mediante el tratamiento se busca corregir las anomalías del metabolismo graso e hidrocarbonado mediante la administración de insulina. También es necesario tratar el trastorno hidroeléctrico a través de la reposición de líquidos y iones con líquido isotónico para revertir los factores precipitantes. La reposición de fluidos es lo más urgente ya que la insulina puede no actuar de manera adecuada cuando la perfusión periférica es deficiente (18).

Durante el tratamiento por cetoacidosis diabética resulta fundamental la reposición de potasio en el organismo. A menor concentración de potasio plasmático inicial, mayor deberá ser la cantidad y rapidez de administración de este. Se trata de mantener la cantidad de este elemento en la sangre en cifras superiores a 3.5 mEq/l en todo momento (18).

La auto monitorización por parte del paciente le permitirá llevar un control en los niveles de glucemia y cetonuria permitiéndole así un diagnóstico precoz de la cetoacidosis para un mejor tratamiento. Por otro lado, un adecuado control en el tratamiento y los hábitos del paciente son la mejor herramienta para prevenir este tipo de complicaciones (18).

4.3.3. Coma hiperglucémico hiperosmolar no cetósico.

Esta complicación se presenta principalmente en pacientes con diabetes tipo 2 en edades avanzadas lo que genera una tasa de mortalidad más elevada que la generada por la cetoacidosis diabética. El cuadro clínico comienza con un deterioro agudo de la función en el sistema nervioso central, deshidratación severa, y puede presentarse en pacientes comúnmente no diagnosticados de manera previa con diabetes (en el 35% de los casos, esta es la primera manifestación de diabetes en el paciente) (18).

Además de la depresión sensorial y los signos neurológicos, esta se caracteriza por una glucemia plasmática mayor a los 600 mg/dl con osmolaridad superior a los 320¹⁰⁴ mOsmol/l y ausencia de cuerpos cetónicos. Los síntomas propios de la hiperglucemias como poliuria y polidipsia, acompañados de deshidratación progresiva, náuseas, vómitos, convulsiones y disminución de la conciencia suelen aparecer de forma constante en el transcurso de varios días y pueden conducir a un coma profundo (18).

Su tratamiento se realiza mediante ingreso hospitalario. Similar al tratamiento para la cetoacidosis diabética, la hiperglucemias hiperosmolar no cetósica se controla mediante la reposición de líquidos con suero salino isotónico y la administración de insulina subcutánea. Una vez que la glucemia desciende por debajo de los 250

² mg/dl se administra suero glucosado al 5% con un aporte mínimo de glucosa de entre 100 y 150 mg por día (18).

La forma de prevenir esta complicación es la misma que la descrita para la prevención de la cetoacidosis diabética, solo que, en esta se deberá extremar el control metabólico derivado de la existencia de algún factor de riesgo (18).

4.3.4. Acidosis láctica.

Se trata de una complicación metabólica poco frecuente que se identifica por una descompensación aguda en la glucemia del paciente. Se caracteriza por un aumento de la concentración hemática de lactato superior a 5mEq/l y un PH inferior a 7.35 (18).

² El cuadro clínico suele presentarse de forma brusca y repentina con taquipnea, deshidratación, dolor abdominal y grado variable de coma (18).

² Su relación con la diabetes puede deberse a una reducción del aporte de oxígeno, una hipoxia hística, una disfunción en el miocardio, una infección o al uso de biguanidas. Está estrechamente asociada a la cetoacidosis diabética o a una descomposición hiperglucémica hiperosmolar no cetósica en combinación de hipoxia tisular (18).

El tratamiento consta de medicamentos antidiabéticos orales (biguanidas), fundamentalmente metformina, sobre todo cuando se trata de pacientes con alguna otra enfermedad como insuficiencia renal. Además de la reposición de líquidos y electrolitos, se administra al paciente bicarbonato en grandes cantidades para elevar el PH a 7.2 y el bicarbonato sérico a 12 mEq/l. Por otro lado, es de vital importancia mantener un estrecho control del estado respiratorio y circulatorio del paciente debido a la alta probabilidad de shock e insuficiencia cardiaca (18).

⁷ 4.4. Complicaciones crónicas de la diabetes.

Aunado a los desórdenes metabólicos derivados de la diabetes, los pacientes suelen presentar con frecuencia otros factores de riesgo como obesidad, hipertensión arterial o dislipidemia. Así también, en algunos casos el paciente mantiene hábitos nocivos como el tabaquismo o el consumo de alcohol en exceso lo que puede desencadenar complicaciones a largo plazo. En este contexto, se les denomina complicaciones crónicas y se pueden clasificar como complicaciones macrovasculares. En este tipo de complicaciones se ven afectadas las arterias (arteroesclerosis) dando como resultado distintas afecciones tales como la enfermedad cardiaca coronaria y la enfermedad cerebrovascular periférica (18)

106 4.4.1. Macroangiopatía.

Es la afección arterioesclerótica de los vasos de mediano y gran calibre. A diferencia de en pacientes no diabéticos, la arteriosclerosis puede ocurrir comúnmente en pacientes más jóvenes³³ con una gravedad y extensión mayores y con peor pronóstico, siendo las enfermedades cardiovasculares la principal causa de morbilidad y mortalidad entre las personas con diabetes mellitus. Aproximadamente entre el 70 y 80 por ciento de los⁵¹ pacientes diabéticos mueren a consecuencia de enfermedades cardiovasculares. La presencia de microalbuminuria o proteinuria es un importante factor precursor o determinante de enfermedad cardiovascular (18).

Para prevenir complicaciones graves derivadas de enfermedad vascular, tales como accidentes cardiovasculares² cerebrovasculares o la muerte, resulta de vital importancia llevar a cabo un adecuado control de los factores de riesgo, poniendo especial énfasis en las alteraciones lipídicas, la hipertensión arterial, la obesidad y el tabaquismo mediante cambios en el estilo de vida del paciente motivándole a mejorar sus hábitos, llevando un control adecuado del peso corporal a través de un plan de alimentación y ejercicio adecuados (18).

²

La asociación Americana de Diabetes recomienda² la utilización de fármacos antiagregantes plaquetarios (ácido acetilsalicílico) como medida de prevención primaria en pacientes diabéticos con perfil cardiovascular³⁸ de alto riesgo. Esto derivado de estudios que han demostrado una significativa eficacia en la reducción en el riesgo de infarto al miocardio (18).

4.4.2. Cardiopatía isquémica.

En pacientes diabéticos la cardiopatía isquémica puede estar presente de forma previa a su detección debido a que, en estos, el riesgo de desarrollar dicho padecimiento es de 2 a 5 veces superior que en pacientes no diabéticos (18).

La forma en que este padecimiento suele presentarse muestra ciertas peculiaridades en los³⁸ pacientes diabéticos además de los síntomas comunes como la angina de pecho, el infarto agudo de miocardio, la insuficiencia cardiaca o la muerte súbita. Por ejemplo, en el caso de Ángor e infarto agudo² de miocardio, el paciente puede cursar con, además de los síntomas clásicos, sudación, astenia, náuseas, vomito, disnea o sincope, mostrando una incidencia tres veces mayor en estos pacientes que en la población en general además de un riesgo de shock cardiógenico e insuficiencia cardiaca marcadamente superior (18).

4.4.3. Cardiopatía isquémica silente.

Esta alteración se encuentra mayormente en pacientes diabéticos. Debido a que generalmente el paciente suele no presentar síntomas y que solo es posible su detección mediante estudios como el electrocardiograma, Holter o prueba de

esfuerzo, resulta necesario practicar a los pacientes un electrocardiograma de manera periódica al menos una vez por año (18).

4.4.4. Insuficiencia cardiaca.

23

Los pacientes diabéticos tienen un riesgo 5 veces superior al de la población en general de presentar insuficiencia cardiaca como complicación de la enfermedad. Esta muestra una tendencia mayor a presentarse en mujeres diabéticas (18).

Para prevenir enfermedades cardiovasculares en pacientes diabéticos, es necesario establecer un control estricto de la glucemia y la dislipemia. Aunado a esto, el consumo de ácido acetilsalicílico puede ayudar a disminuir en hasta un 32% la posibilidad de muerte. Por otra parte, en pacientes fumadores el riesgo aumenta considerablemente por lo que resulta de vital importancia abandonar el tabaquismo por completo (18).

4.4.5. Arteriopatía periférica.

Otra de las complicaciones cardiovasculares en pacientes diabéticos es la arteriopatía periférica que suele presentarse 4 veces más en hombres y hasta 8 veces más en mujeres en comparación con la prevalencia en la población general. La lesión aparece principalmente en los miembros inferiores, aunque en algunas ocasiones puede presentarse en miembros superiores. Esta llega a provocar dolor en piernas que puede imposibilitar al paciente para caminar, agravándose conforme este recorre cierta distancia y empeorando con el tiempo, hasta el punto en que el dolor intenso puede permanecer incluso estando en reposo, hasta el punto en que si la enfermedad sigue progresando puede producir lesiones como ulceras o gangrena (18).

El tratamiento para este tipo de complicaciones consta de técnicas endovasculares como la angioplastia con balón, la ateroectomía, la angioplastia con láser y la revascularización (18).

La mejor manera de prevenir este tipo de complicaciones es mediante un control adecuado de los factores de riesgo como mantener un peso adecuado, controlar niveles lipídicos en la sangre, vigilar constantemente la tensión arterial, llevar una dieta y un programa de ejercicio adecuados, dejar de fumar y llevar un control adecuado de la glucemia (18).

4.4.6. Accidente cerebrovascular.

Otra forma de complicación vascular en los pacientes diabéticos es el accidente cerebrovascular que se presenta 2 veces más ¹⁸⁷ que en los no diabéticos. Uno de los principales factores de riesgo de accidente cerebrovascular en los pacientes diabéticos es la hipertensión. El 50% de los casos de muerte súbita en pacientes diabéticos e hipertensos se debe a este tipo de complicaciones (18).

174

El buen control de los factores de riesgo, especialmente la hipertensión arterial, aunado a la auscultación de carótidas, son las medidas que se tomaran para intentar prevenir y controlar la enfermedad cerebrovascular. Además, estos pacientes pueden ser tratados con fármacos anticoagulantes y ácido acetilsalicílico. Así también, dependiendo de lo agudo y el diagnóstico de la lesión, esta podría ser tratada mediante intervención quirúrgica (18).

62

4.4.7. Estenosis de la arteria renal y aneurisma de la aorta abdominal.

Otras de las complicaciones vasculares comunes en pacientes diabéticos es la estenosis de la arteria renal y aneurisma de la aorta abdominal. La auscultación de soplos abdominales representa la principal sospecha de estenosis en la arteria renal, que se comprueba mediante ecosonograma, mostrando alteraciones y asimetría en el tamaño de los riñones (18).

4.5. Otras complicaciones.

Como se menciona anteriormente, derivado de un mal control de la diabetes y los malos hábitos en los pacientes²⁵, pueden presentarse un gran número de complicaciones. Además de las complicaciones crónicas y agudas de la enfermedad, los pacientes diabéticos pueden llegar a presentar complicaciones en otras partes del organismo (18).

4.5.1. Pie diabético.

29

Derivado de una hiperglucemia prolongada, el pie diabético es una alteración clínica de base etiopatogénica neuropática que provoca lesiones o ulceraciones en los pies del paciente con diabetes en lo que puede presentar o no isquemia en vasos y arterias de la extremidad. Esta complicación⁶⁷ representa una de las principales causas de amputación no traumática de las extremidades inferiores de los pacientes diabéticos (18).

Con la intención de establecer un tratamiento adecuado, a este tipo de complicaciones se les clasifica por grados de riesgo que van del 0 al 5 de forma ascendente por el nivel de gravedad de la lesión. El grado 0 representa un pie en riesgo que no muestra lesiones, el grado 1 representa una lesión pequeña como una ulceración superficial, el grado 2 representa una lesión más delicada como una ulceración profunda que penetra en el tejido celular subcutáneo y que afecta a tendones y ligamentos pero que no presenta infección o afecciones en el hueso, el grado 3 representa una ulceración profunda acompañada de inflamación, absceso u osteítis, el grado 4 representa una lesión localizada que exterioriza gangrena generalmente en talón, dedos o zonas distales del pie y el grado 5 que representa lesión grave gangrenada en gran parte de la extremidad (18).

⁸⁵ Para prevenir este tipo de complicaciones es necesario identificar a los pacientes con mayor probabilidad de padecer neuropatía o arteriopatía como los fumadores, con más de 10 años de evolución y con un deficiente control ⁴⁴ glucémico. Además, la detección oportuna de pie diabético en grado 0 a través de la inspección periódica por parte del paciente y del médico responsable, uso de calzado adecuado y las visitas regulares con el podólogo son la mejor herramienta para prevenir complicaciones más graves y evitar la amputación de la extremidad (18).

4.5.2. Alteraciones en la piel.

Así también, existe un gran número de alteraciones cutáneas asociadas a la diabetes mellitus ⁹³ con mayor predisposición a las infecciones y alteraciones neuropáticas como dermopatía diabética, necrobiosis lipoídica, bullosis diabeticorum, granuloma anular, xantomas eruptivos, lipoatrofia y lipohipertrofia (18).

4.5.3. Alteraciones bucales.

De igual forma, estos pacientes ²³ pueden presentar frecuentemente complicaciones en la cavidad bucal, como caries dental, candidiasis oral, mucomicosis, glositis romboidal media, xerostomía, síndrome de ardor bucal, agrandamiento en la glándulas salivales y alteraciones del gusto, entre otras (18).

4.6. Datos y cifras de la diabetes en México.

Según la Federación Internacional de Diabetes para 2019 en México existían 12.8 millones de pacientes diabéticos con una tendencia de aumento que prevé que para el 2045 la cantidad de personas con esta enfermedad en nuestro país aumentará a 22.9 millones. Esto sitúa a México en el sexto lugar en prevalencia de diabetes a nivel mundial después de India, China, India, Estados Unidos, Pakistán y Brasil (19).

Para el año 2020 en México 151,000 personas fallecieron a causa de la diabetes. ⁸¹ Esto representa un 14% del total de las defunciones ocurridas dicho año en el país. La tasa de mortalidad por diabetes para el 2020 fue de 11.95 personas por cada 10 mil habitantes. Estas cifras fueron las más altas reportadas en los últimos 10 años. En la siguiente gráfica se muestra la tasa de mortalidad según las cifras anteriores, por género en México (20):

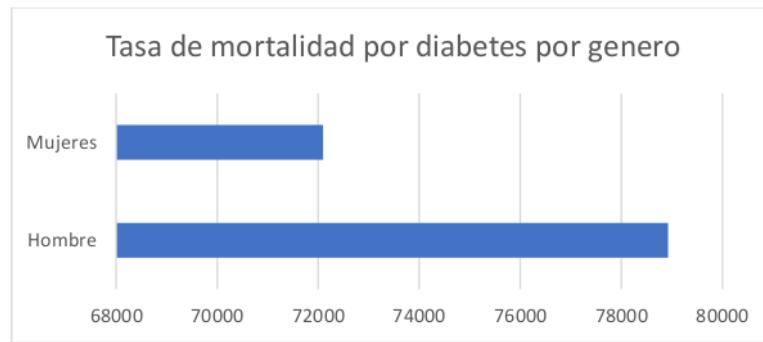


Diagrama 6. Tasa de mortalidad por diabetes en México según el género.

52

Se ha observado una tendencia que indica que conforme aumenta la edad de las personas⁵⁹ en nuestro país incrementa el diagnóstico de la enfermedad. De acuerdo con la encuesta Nacional de Salud y Nutrición (ENSANUT) del 2018, realizada por el Instituto Nacional de Geografía y Estadística (INEGI), a nivel nacional, el 28.5% de la población entre 60 y 69 años declaró haber sido diagnosticada con diabetes, ³² lo que representa a 2.3 millones de personas. En la siguiente gráfica se muestra la prevalencia de diabetes en la población de 20 años y más, por grupo de edad y por género en 2018 (20):

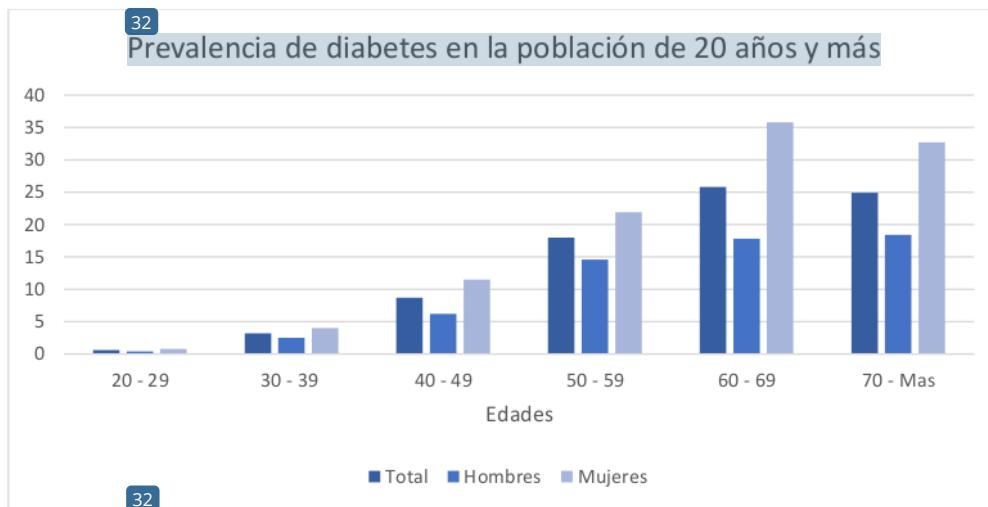


Diagrama 7. Prevalencia de diabetes en la población de 20 años y más. México (ENSANUT 2018).

Dado que esta enfermedad afecta diferentes órganos y sistemas de los pacientes, de ella se deriva una cantidad importante de complicaciones que pueden afectar significativamente la salud de los pacientes. Gracias a los datos recabados en la encuesta ENSANUT del 2018 del INEGI podemos conocer una aproximación de la tasa de prevalencia de las distintas afecciones desarrolladas por los pacientes como consecuencia de la enfermedad. A continuación, se muestra la prevalencia de las distintas complicaciones en los pacientes diabéticos (21):

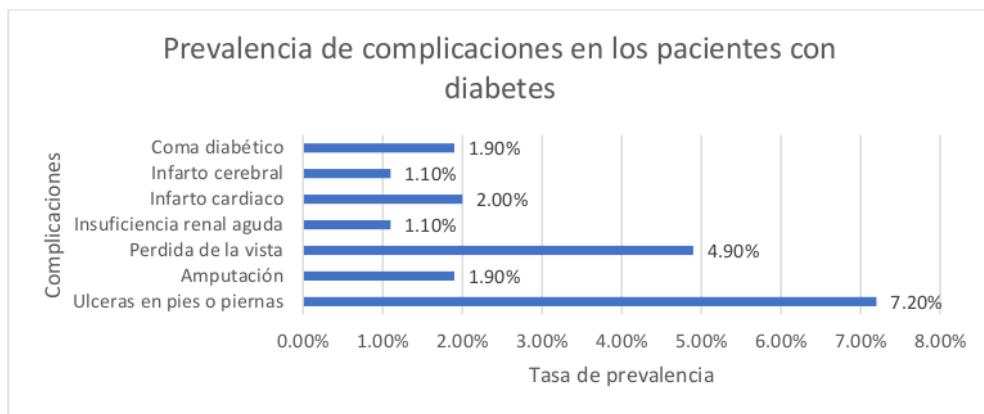


Diagrama 8. Tasa de prevalencia de complicaciones por diabetes. México (INSANUT 2018).

29

5. MODELO DE REGRESIÓN LOGÍSTICA PARA EL ¹⁶⁹ANÁLISIS DE HÁBITOS Y COMPLICACIONES EN PACIENTES CON DIABETES.

La diabetes es una enfermedad que puede llegar a afectar todos los órganos y sistemas de las personas. Resulta por tanto de vital importancia dotar al paciente de una herramienta que le permita llevar un control y tratamiento adecuado de su padecimiento.

Estudios han demostrado que mediante el uso de herramientas de aprendizaje automático y el análisis de datos podemos predecir y detectar de manera oportuna distintas enfermedades. Sin embargo, estos estudios se limitan a la predicción de la enfermedad, dejando de lado la parte de la prevención de un sin número de complicaciones derivadas de los malos hábitos del paciente.

En la actualidad contamos también con una gran variedad de dispositivos electrónicos inteligentes de uso personal que monitorean distintas señales y aspectos de la salud del usuario. No obstante, estos están enfocados para un uso generalizado de las personas, resultando necesario el desarrollo de una herramienta que aproveche dichas funciones y se enfoque en el monitoreo y control de pacientes diabéticos.

56

5.1. Encuesta Nacional de Salud y Nutrición 2018.

Para conocer la prevalencia y distribución de las enfermedades agudas y crónicas degenerativas ⁸⁸ y las condiciones generales de salud y nutrición de la población en México, el Instituto Nacional de Estadística y Geografía (INEGI) en coordinación con el ¹⁴¹ Instituto Nacional de Salud Pública (ISNP) y la Secretaría de Salud Pública de México, realizó la Encuesta Nacional de Salud y Nutrición (ENSANUT) en 2018. Para ello, se aplicaron 10 cuestionarios de salud y 8 de nutrición a una muestra de 50,000 hogares para los componentes de salud y 32,000 para los de nutrición, sumando así 126.5 millones de habitantes encuestados en total (21)..

58

5.2. Descripción del conjunto de datos de la muestra.

De las variables disponibles en nuestro conjunto de datos, una determina si la persona encuestada tiene diabetes por lo que utilizamos dicha variable para realizar un filtrado y seleccionar los registros de pacientes con la enfermedad obteniendo 4555 registros en total. Previo al análisis de los datos, realizamos algunos ajustes para que los valores en los registros categóricos fueran binarios con la finalidad de que el formato precise las respuestas positivas (Sí) a través del número “1” y el “0” para las respuestas negativas (No).

En la siguiente tabla se muestra una descripción acerca del conjunto de datos utilizados para el análisis.

| ID | Nemónico | Dato | Tipo | Longitud | Código | Clases | |
|--------------------------------------------------------------------------------------------------------------|----------|------------------------------------------------------------------|----------|----------|--------|--------|----------|
| | | | | | | Valor | Etiqueta |
| 1 | P3_1 | ¿El paciente tiene Diabetes? | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| Derivado de la Diabetes. ¿El paciente ha presentado alguna de las siguientes complicaciones? | | | | | | | |
| 2 | P3_18_1 | Ulceras en pies o piernas. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 3 | P3_18_2 | Amputación | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 4 | P3_18_3 | Disminución de la vista | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 5 | P3_18_4 | Perdida de la vista | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 6 | P3_18_5 | Insuficiencia renal aguda | Booleano | 1 | 0,1 | 0 | No |
| 86 | | | | | | 1 | Si |
| 7 | P3_18_6 | Infarto cardiaco | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 8 | P3_18_7 | Infarto cerebral | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 9 | P3_18_8 | Coma diabético | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 10 | P3_18_9 | Perdida del conocimiento, confusión o desmayo | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| Para el cuidado y control de la Diabetes. ¿El paciente realiza las siguientes actividades de forma habitual? | | | | | | | |
| 11 | P3_3 | Acude de manera frecuente al médico especialista. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 12 | P3_13_1 | Lleva a cabo un plan alimenticio. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 13 | P3_13_2 | Lleva a cabo un plan de ejercicio. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 14 | P3_16_1 | Acude con frecuencia para revisión de la vista. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 15 | P3_16_2 | Toma una aspirina diaria. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 16 | P3_16_3 | Revisa diariamente sus pies. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 17 | P3_16_6 | Revisa sus niveles de tensión arterial. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 18 | P3_16_7 | Se aplica la vacuna contra la influenza. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 19 | P3_16_8 | Se aplica la vacuna contra el neumococo. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 20 | P3_16_10 | Toma medicamento para control del colesterol. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 21 | P3_16_11 | Toma medicamento para el control de la presión arterial. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 22 | P3_16_12 | Dejó de fumar. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 23 | P3_16_13 | Acude con frecuencia para revisión dental. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 24 | P3_16_14 | Recibe orientación sobre su enfermedad. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 25 | P3_16_16 | Evita utilizar calzado incomodo. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 26 | P3_16_19 | Realiza medidas preventivas. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 27 | P3_8 | Toma pastillas o insulina para controlar sus niveles de glucosa. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |
| 28 | P3_17 | Revisa sus pies con frecuencia en búsqueda de lesiones. | Booleano | 1 | 0,1 | 0 | No |
| | | | | | | 1 | Si |

| | | | | | | | |
|----|--------|---------------------------------|----------|---|-----|---|----|
| 29 | P13_2 | No fuma. | Booleano | 1 | 0,1 | 0 | No |
| 30 | P13_11 | No consume bebidas alcohólicas. | Booleano | 1 | 0,1 | 0 | No |

Tabla 8. Descripción del conjunto de datos de la muestra.

5.3. Estado del arte.

155

Uno de los principales retos a la hora de aplicar técnicas de aprendizaje automático para la detección temprana de enfermedades se centra en identificar y modelar los datos para obtener información precisa que proporcione los marcadores particulares para la predicción de una enfermedad en concreto.

205

A continuación, realizamos un análisis de los estudios más recientes sobre minería de datos y aprendizaje automático para la detección temprana de enfermedades, centrando nuestra investigación en la diabetes. A continuación, mostramos una síntesis de los resultados de los estudios analizados en la presente revisión del estado del arte:

130

- Presentaron el sistema 5G-Smart Diabetes en el que, a través de la aplicación de tecnologías recientes como Wearable 2.0 que permite la integración del humano en la nube para los sistemas sanitarios de nueva generación, y la aplicación de técnicas de aprendizaje automático y minería de datos a gran escala, realizaron un estudio sobre una gran cantidad de datos de pacientes que padecen diabetes. Con esto lograron construir un gran banco de datos para su aplicación 5G-Smart Diabetes que interactúa con dispositivos wearables, teléfonos inteligentes y grandes nubes de datos para ofrecer de manera efectiva un diagnóstico personalizado que sugiere un tratamiento específico para cada paciente en particular (22).
- Realizaron un estudio en el que aplicaron algoritmos de clasificación en un entorno distribuido para predecir la diabetes utilizando Apache Spark. Abordan las diferencias entre Spark y Hadoop con respecto a la tolerancia a fallos, la ejecución y el soporte de operaciones paralelas (23).

En el estudio, los autores concluyen que Spark es mejor que Hadoop en muchos aspectos ya que Spark desarrolla soluciones más rápidas y eficientes para analizar grandes cantidades de datos en mucho menor tiempo (23).

10

Para el desarrollo de la investigación, se basaron en datos con información sobre 43 parámetros importantes como la presión arterial, el nivel de insulina en suero y el índice de masa corporal (23).

- Utilizaron el análisis predictivo en HUE para predecir las enfermedades persistentes relacionadas con la diabetes mellitus mediante la recopilación de un conjunto de datos de los indios Pima, logrando así, a través de un

102

modelo de clasificación mediante la técnica de máquina de vector soporte o SVM⁴¹ por sus siglas en inglés, un modelo eficaz para llevar un seguimiento del número de personas que sufren de diabetes en esa población (24).

- Realizaron un análisis de investigación mediante la aplicación del algoritmo de árboles de decisión en el entorno Hadoop/Map Reduce para predecir y clasificar por enfermedad la diabetes mellitus tipo 1 y 2 proporcionando un marco de referencia para mejorar a través del análisis de datos a gran escala la disponibilidad y asequibilidad de la atención sanitaria, principalmente en barrios rurales y urbanos (25).
- Utilizando el algoritmo escalable de clasificación de bosque aleatorio, proponen un modelo predictivo que identifica con precisión la tasa de clasificación del riesgo de diabetes (26).

Para realizar el análisis predictivo, tomaron en cuenta varios factores con datos demográficos, parámetros de pacientes de un hospital y varios indicadores específicos de la diabetes, obteniendo un resultado del 87.5% de precisión en su análisis (26).

- Basándose²⁰¹ en algunos factores característicos de la diabetes como la glucosa, el índice de masa corporal, la insulina y la edad entre otros, desarrollaron varios modelos para predecir la diabetes (27).

Para lograr una mayor precisión en la clasificación de los datos, realizaron una comparación de la base de datos del estudio con una base de datos histórica (27).

Los modelos implementados se muestran a continuación junto con el porcentaje de precisión obtenido con cada una de las técnicas: clasificador de vector soporte, 60%, clasificador de bosque aleatorio, 91%, clasificador de árbol de decisión, 86%, clasificador de árbol extra, 91%, algoritmo Ada Boost, 93%, perceptrón, 76%, algoritmo de análisis discriminante lineal, 94%, regresión logística, 96%, K-Nearest Neighbour (KNN), 90%, Gaussian Naive Bayes, 93%, algoritmo Bagging, 90% y clasificador Gradient Boost, 93% (27).

- muestran un trabajo de investigación en el que utilizan varias técnicas de aprendizaje automático aplicadas a bases de datos sobre diabetes para determinar cómo dichos modelos pueden ayudar para la predicción de la enfermedad (3).

5

Para la clasificación de los datos utilizan los algoritmos de máquina de vector soporte (SVM), bosque aleatorio y K-Nearest Neighbour (KNN). Los investigadores observaron que los algoritmos de bosque aleatorio y SVM presentan mejores resultados que el algoritmo KNN.

En cuanto a la precisión de las pruebas conjuntas, el algoritmo bosque aleatorio presentó un 85% siendo el mejor resultado obtenido de los cuatro experimentos. SVM presentó un 84% y KNN un 81% (3).

La siguiente tabla se muestran tanto los métodos y herramientas utilizados como los resultados obtenidos en las pruebas presentadas en esta revisión del estado del arte.

| Algoritmos | Herramientas | Precisión | 1 Análisis | Referencia |
|-----------------------------------------|---------------------------|-----------|---------------------------------------|------------|
| Máquina de vector soporte (SVM) | 5G-smart Diabetes | 93.0% | Diagnóstico de diabetes | (22) |
| Arboles de decisiones | 5G-smart Diabetes | 91.0% | Diagnóstico de diabetes | (22) |
| Redes Neuronales Artificiales (ANN) | 5G-smart Diabetes | 82.0% | Diagnóstico de diabetes | (22) |
| Clasificador Redes Bayesianas | Apache Spark | 72.8% | Predicción de diabetes | (23) |
| Máquina de vector soporte (SVM) | Apache Spark | 65.3% | Predicción de diabetes | (23) |
| Redes Neuronales Artificiales (ANN) | Apache Spark | 66.6% | Predicción de diabetes | (23) |
| Máquina de vector soporte (SVM) | Hadoop/Apache Spark | No aplica | Clasificación de pacientes (Diabetes) | (24) |
| Arboles de decisiones | Hadoop/Mapreduce | No aplica | Predicción de diabetes | (25) |
| Clasificador de Bosque Aleatorio | Mapreduce | 87.5% | Predicción de diabetes | (26) |
| Arboles de Decisiones | Información no disponible | 86.0% | Predicción de diabetes | (27) |
| Clasificador Gaussian Redes Bayesianas | Información no disponible | 93.0% | Predicción de diabetes | (27) |
| Análisis de discriminación lineal (LDA) | Información no disponible | 94.0% | Predicción de diabetes | (27) |
| Máquina de vector soporte (SVM) | Información no disponible | 60.0% | Predicción de diabetes | (27) |
| Bosque Aleatorio | Información no disponible | 91.0% | Predicción de diabetes | (27) |
| Arboles Extra | Información no disponible | 91.0% | Predicción de diabetes | (27) |
| AdaBoost | Información no disponible | 93.0% | Predicción de diabetes | (27) |
| Perceptrón | Información no disponible | 76.0% | Predicción de diabetes | (27) |
| Regresión Logística | Información no disponible | 96.0% | Predicción de diabetes | (27) |
| Clasificador por Impulso del Gradiente | Información no disponible | 93.0% | Predicción de diabetes | (27) |
| Embolsamiento | Información no disponible | 90.0% | Predicción de diabetes | (27) |
| K-Vicino Cercano (KNN) | Información no disponible | 90.0% | Predicción de diabetes | (27) |
| Bosque Aleatorio | Programación en Python | 85.0% | Predicción de diabetes | (3) |
| Máquina de Vector Soporte (SVM) | Programación en Python | 84.0% | Predicción de diabetes | (3) |
| K-Vicino Cercano (KNN) | programación en Python | 81.0% | Predicción de diabetes | (3) |

Tabla 9. Comparativa entre métodos, herramienta y resultados en el estado del arte.

27

El aprendizaje automático es una rama de la inteligencia artificial (IA) que consiste en la aplicación de programas informáticos que aprenden con base en la experiencia para realizar tareas de forma más eficiente a medida que evoluciona. Su principal objetivo es el desarrollo de teorías, técnicas y algoritmos que permitan a un sistema modificar su comportamiento a través de la inferencia inductiva, que se basa en la observación de datos que representan información sobre un proceso o fenómeno estadístico. En la actualidad, el aprendizaje automático supone una gran ventaja para la aplicación de sistemas que dan soporte a diferentes áreas dedicadas a la atención de la salud (3).

51

A continuación, se presenta un breve resumen de las herramientas utilizadas por los investigadores referenciados en este análisis:

8

- **Hadoop.** en día, se genera una gran cantidad de datos no estructurados a un ritmo acelerado (Big Data), por ello, resulta necesario contar con herramientas para su almacenamiento y análisis. Hadoop es una herramienta de código abierto que trabaja con un sistema de archivos distribuido en un clúster de hardware básico. En este contexto, para el almacenamiento, procesamiento, manejo y análisis de datos en los que se incluyen audios, vídeos e imágenes entre otros, se requieren herramientas con una enorme potencia de cálculo y un gran tamaño de almacenamiento para lo que Hadoop resulta una excelente herramienta (25) Hoy.

El sistema de archivos distribuidos Hadoop (HDFS) tiene la capacidad de manejar un enorme volumen de datos porque se divide y almacena en varios nodos del clúster. HDFS almacena en diagonal las diferentes tecnologías para recuperar la pérdida de información en caso de fallo y es adecuado para el almacenamiento y procesamiento distribuido proporcionando varias funciones como la distribución, el almacenamiento, el permiso de procesamiento de archivos y la autenticación. (25).

- **Mapreduce.** Es un framework de programación que soporta enormes cantidades de datos, en paralelo con una gran cantidad de grupos de hardware (miles de nodos) de forma fiable y tolerante a fallos. Por lo general, divide el conjunto de datos de entrada en fragmentos independientes que son procesados por las tareas de mapas completamente en paralelo. El framework ordena los resultados que entregan los mapas para posteriormente realizar un trabajo de reducción, de ahí su nombre. Una vez realizado el trabajo anterior, Mapreduce se encarga de programar las tareas, monitorizarlas y volver a ejecutar las tareas fallidas (28).

Generalmente, esta herramienta utiliza los mismos nodos para el cálculo y el almacenamiento de datos, es decir, el framework Mapreduce y el sistema de archivos distribuido Hadoop se ejecutan en el mismo conjunto de nodos, lo que permite a Mapreduce programar las tareas de forma eficaz en los nodos de datos ya presentes, generando un ancho de banda muy elevado para todo el clúster (28).

- **5G-Smart Diabetes.** Es una herramienta enfocada a la prevención de la diabetes y su tratamiento tras la hospitalización. Su objetivo es conseguir un monitoreo continuo de los estados fisiológicos del paciente con el fin de lograr proporcionarle un tratamiento adecuado y personalizado. Para conseguir un control fisiológico adecuado de la enfermedad, es necesario analizar diferentes marcadores particulares, tales como los hábitos y características generales del paciente (29).

La arquitectura del sistema 5G-Smart Diabetes incluye tres capas: La primera es la capa de detección donde se recoge la información fisiológica y los hábitos del paciente a través de artículos inteligentes que recogen las señales corporales¹⁶⁴ en tiempo real como la temperatura, un electrocardiograma y la saturación de oxígeno en sangre, en la segunda capa se realiza un diagnóstico particular mediante el análisis de datos a gran escala utilizando métodos de aprendizaje automático que generan modelos personalizados para predecir la enfermedad y por último la capa de intercambio de datos donde se incluyen los datos del usuario y las redes sociales. En esta última, se comparte información que servirá para estudios en el futuro (29).

176

- **Python.** Es un lenguaje de programación para la integración de sistemas con una forma de trabajo rápida y eficaz. Python¹²⁷ está desarrollado bajo una licencia de código abierto aprobada por el modelo de interconexión de sistemas abierto, OSI por sus siglas en inglés, lo que hace que su uso y distribución sean gratuitos (30).

El índice de paquetes de Python consta de miles de módulos de terceros. Tanto su librería estándar como los módulos aportados por la comunidad permiten disponer de un amplio abanico de posibilidades como el desarrollo web, el acceso a bases de datos, el análisis científico y numérico, la educación, la programación en red y el desarrollo de software entre otros (30).

- **Apache Spark.** trata de un motor de análisis unificado para el procesamiento de datos a gran escala que consigue un alto rendimiento tanto para los datos por lotes como para la transmisión de datos, utilizando un programador DAG de última generación, un optimizador de consultas y un motor de ejecución física. Spark ofrece más de 80 operadores de alto nivel que facilitan la creación de aplicaciones paralelas y puede utilizarse de forma interactiva desde los intérpretes de comandos (shells), Scala, Python, R y SQL. Se puede ejecutar en Hadoop, Apache Mesos y Kubernetes entre otros, de forma independiente o en la nube y puede acceder a diversas fuentes de datos. También puede acceder a datos en HDFS, Alluxio, Apache Cassandra, Apache Hbase, Apache Hive y cientos de otras fuentes (31) Se.

¹⁵⁸as librerías pueden combinarse perfectamente en una sola aplicación, incluyendo SQL y DataFrames, Mlib para machine learning, Graphx y Spark Streaming (31) Se.

107

- **Apache Cassandra.** Es un sistema de ¹⁰⁹gestión de bases de datos de código abierto escrito en Java, que permite escalabilidad y alta disponibilidad sin comprometer el rendimiento. Es una plataforma para datos de misión crítica con escalabilidad lineal y tolerancia a fallos, probada en hardware básico o infraestructura en la nube. El soporte de Cassandra para la replicación a través de múltiples centros de datos proporciona una menor latencia para los usuarios (31) Se.

45

Cassandra ofrece un sólido soporte para múltiples centros de datos con replicación asíncrona sin necesidad de un servidor maestro que permita operaciones de baja latencia para todos los usuarios. Los datos de Cassandra se replican automáticamente en múltiples nodos, lo que permite la tolerancia a fallos, además los nodos que fallan pueden ser reemplazados sin tiempo de inactividad (31) Se.

5

Los investigadores han utilizado diversas técnicas de aprendizaje automático aplicadas a distintas bases de datos sobre la diabetes, con el fin de determinar ya sea una solución para la detección temprana, el tratamiento o la prevención de esta enfermedad. A continuación, presentamos un breve resumen de las técnicas y algoritmos utilizados por los investigadores en este análisis:

- **Red Neuronal Artificial (ANN):** Es un algoritmo de aprendizaje profundo inspirado en los patrones de conectividad ²¹de las neuronas de la corteza cerebral humana, en el que se presenta una capa de entrada, una capa oculta y una capa de salida. En estas, se asigna un grado de importancia a varios aspectos del modelo siendo capaz de diferenciar unos de otros. En otras palabras, están compuestas por neuronas con pesos y sesgos que tienen la capacidad de aprender. En este algoritmo el preprocesamiento requerido es menor en comparación con otros algoritmos de aprendizaje automático (32).
- **Máquinas de Vector Soporte (SVM):** Son algoritmos de aprendizaje supervisado que proporcionan una solución para la clasificación de datos y el analizar patrones. Consisten principalmente en la construcción de un hiperplano con diferentes dimensiones en el que los entrenamientos lineales se dividen fácilmente por clases. Para los entrenamientos no lineales, en los que la separación en clases no es posible, las SVM introducen las Funciones Kernel que transforman los datos en un espacio de alta dimensionalidad para así lograr separarlos de forma lineal minimizando el error de la prueba y mejorando la clasificación de los datos (33).

- **Arboles de decisión:** Son modelos de clasificación de datos utilizados en inteligencia artificial que se caracterizan por su contribución visual a la toma de decisiones. Para realizar una prueba con este método²² se utilizan dos bases de datos. El objetivo principal de estos modelos es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Estos algoritmos están representados por un conjunto de nodos, hojas y ramas: El nodo principal o raíz consiste en el atributo a partir del cual se realizará el proceso de clasificación, los nodos internos corresponden a las preguntas sobre el atributo particular del problema, las ramas que salen, a partir de estos nodos, están etiquetadas con posibles valores del atributo y los nodos finales u hojas corresponden a una decisión (34)..
- **K Vecinos Cercanos (KNN):** Consiste en un algoritmo supervisado que busca entre las observaciones más cercanas a las que intenta predecir y clasificar un punto de interés basándose en la mayoría de los datos circundantes. La clasificación KNN consta de dos etapas, la primera determina los vecinos más cercanos y la segunda lleva a cabo la determinación de la clase utilizando¹⁹³ los vecinos. Este algoritmo calcula la distancia entre los datos a clasificar y el resto de los datos de entrenamiento, selecciona los elementos más cercanos y realiza una votación por mayoría para que la clase que domine sea la que determine la clasificación final (35)
- **AdaBoost:** Es un algoritmo que incorpora varios clasificadores que podrían considerarse débiles utilizando un enfoque iterativo para aprender de los errores y formar un clasificador más fuerte. Funciona dando más peso a las instancias difíciles de clasificar y quitando peso a las que ya están bien catalogadas. En otras palabras, podemos decir que el tipo de aprendizaje de AdaBoost combina varios algoritmos básicos para formar un algoritmo optimizado (36).
- **Análisis de Discriminación Lineal:** Es un método de clasificación de datos que maximiza la relación entre la²⁰⁴ varianza de las clases asegurando la mayor separabilidad de los datos. Consiste en la aplicación de la técnica de regresión lineal para encontrar una combinación de características que diferencien dos o más tipos de datos para el reconocimiento de patrones en el aprendizaje automático (37).

¹⁴⁷ La regresión lineal es un método⁷³ de análisis estadístico que busca determinar la relación cuantitativa entre dos o más variables. Cuando la variable dependiente y la variable independiente muestran una relación lineal, la función de mínimos cuadrados de la ecuación de regresión lineal puede utilizarse para establecer un modelo matemático de la relación entre la variable dependiente y la variable independiente. El objetivo de este método es encontrar los parámetros más adecuados y utilizar una línea recta para ajustar los puntos de datos precipitados (38).

- 3
- **Regresión Logística:** Se utiliza principalmente para la clasificación de datos. A diferencia de la Regresión Lineal, en la Regresión Logística los puntos de datos no están dispuestos en filas de líneas. La Regresión Logística busca encontrar la línea límite de la clasificación, que está representada por la fórmula de regresión. El clasificador utiliza el algoritmo de optimización para encontrar el mejor coeficiente en la fórmula de regresión. La clasificación recibe un conjunto arbitrario de entradas para obtener mediante una función la representación de salida de la clasificación de los datos de entrada (38).
 - **Redes Bayesianas:** Consiste en una colección de algoritmos de clasificación basados en el teorema de Bayes. Este clasificador proporciona excelentes resultados cuando se utiliza para el análisis de datos de texto. Este algoritmo se utiliza como clasificador probabilístico. El modelo Redes Bayesianas utiliza conceptos de modelos de mezcla que son capaces de establecer la máxima probabilidad del componente que consiste en el teorema de Bayes (39).
 - **K-Means:** Es uno de los algoritmos de clústeres deterministas más sencillos. El objetivo de este algoritmo es dividir N número de observaciones en k grupos determinados por el usuario, en los que cada observación pertenece a un cluster con un centroide cercano a la observación. Este algoritmo se utiliza ampliamente en la segmentación de imágenes, la clasificación de patrones, el reconocimiento de objetos y la minería de datos (40).

5.4. Métodos y herramientas implementados en el modelo.

Dado que el formato del conjunto de datos con el que trabajamos es categórico, decidimos utilizar el algoritmo de Regresión Logística para realizar un modelo de clasificación. Para el desarrollo de este y el análisis de los datos utilizamos Jupyter Notebook y las librerías de Matplotlib, Pandas, Numpy, Math, SKLearn y Statsmodel como herramientas para análisis de datos y estadísticas con programación en Python.

```
In [1]: # Importamos las librerías necesarias

# Tratamiento de datos
import pandas as pd
import numpy as np
import math
import statsmodels.api as sm
%matplotlib inline

# Gráficos
import matplotlib.pyplot as plt
from matplotlib import style
import seaborn as sns

# Preprocesado y modelado
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import plot_confusion_matrix
import statsmodels.api as sm
import statsmodels.formula.api as smf

# Configuración matplotlib
plt.rcParams['image.cmap'] = "bwr"
plt.rcParams['figure.dpi'] = "100"
plt.rcParams['savefig.bbox'] = "tight"
style.use('ggplot') or plt.style.use('ggplot')

# Configuración warnings
import warnings
warnings.filterwarnings('ignore')
```

Activar Window
Ve a Configuración p

Imagen 1. Librerías utilizadas en Python.

Para el **83** diseño del modelo utilizamos la base de datos "CS_Adultos.csv" que contiene información relacionada con la salud y los hábitos de los encuestados.

```
In [2]: # Importamos el archivo CS_ADULTOS.csv obtenido de la pagina del 'INEGI' y lo convertimos en 'DataFrame'
data = pd.read_csv('/Users/Equipo/DataMining/salud/CS_ADULTOS.csv')
data = pd.DataFrame(data)
data = data.fillna(0)
```

Imagen 2. Importación y conversión en DataFrame del archivo CS_Adultos.csv del INEGI

Seleccionamos solo las variables que de alguna manera se relacionan con la diabetes para establecer cómo variables dependientes todas aquellas que proporcionan información de alguna complicación de la diabetes en los pacientes y como variables independientes, aquellas que se relacionan con los hábitos del paciente en el tratamiento de la enfermedad.

```
In [3]: # Filtramos y eliminamos las columnas del dataframe que no guardan relación con la columna 'P3_1' (Diabetes)
data = data.drop(['UVM', 'IV_SEL', 'HOGAR', 'NUMIREM', 'P1_2', 'P1_3', 'P1_1', 'P1_6', 'P1_8', 'P1_9', 'P1_10_1',
'P1_10_2', 'P1_10_3', 'P1_10_4', 'P1_10_5', 'P1_10_6', 'P1_10_7', 'P1_10_8', 'P1_10_9', 'P1_10_10', 'P2_1_1',
'P2_1_2', 'P2_1_3', 'P2_1_4', 'P2_1_5', 'P2_1_7', 'P2_2', 'P2_1_6', 'P3_2', 'P3_3V', 'P3_4', 'P3_5D', 'P3_5M',
'P3_5A', 'P3_6', 'P3_7_1', 'P3_7_2', 'P3_7_3', 'P3_7_4', 'P3_7_5', 'P3_7_6', 'P3_7_7', 'P3_7_8', 'P3_7_9',
'P3_7_10', 'P3_7_11', 'P3_7_12', 'P3_9A', 'P3_10M', 'P3_10A', 'P3_11', 'P3_12', 'P3_13_3', 'P3_13_4',
'P3_13_5', 'P3_15_1', 'P3_15_2', 'P3_15_3', 'P3_15_4', 'P3_15_5', 'P3_15_6', 'P3_15_7', 'P3_14_8', 'P3_16_4',
'P3_16_5', 'P3_16_17', 'P3_16_18', 'P4_2M', 'P4_2A', 'P4_3', 'P4_4', 'P4_5M', 'P4_5A', 'P4_6', 'P4_7', 'P4_8_1',
'P4_8_2', 'P4_8_3', 'P4_8_4', 'P4_8_5', 'P4_9_1', 'P4_10_1', 'P4_10_2', 'P4_10_3', 'P4_10_4',
'P4_10_5', 'P4_10_5V', 'P4_10_6', 'P4_10_6V', 'P4_10_6C', 'P5_1', 'P5_2_2', 'P5_2_3', 'P5_3', 'P5_4', 'P5_5',
'P5_6', 'P5_7', 'P6_1_1', 'P6_1_2', 'P6_2_1', 'P6_2_2', 'P6_2_3', 'P6_2_4', 'P6_2_5', 'P6_2_6', 'P6_2_7',
'P6_3', 'P6_4', 'P6_5_1', 'P6_5_2', 'P6_5_3', 'P6_5_4', 'P6_6', 'P6_7_1', 'P6_7_2', 'P6_7_3', 'P6_7_4',
'P6_8_1', 'P6_8_2', 'P6_8_3', 'P6_8_4', 'P6_8_5', 'P6_8_6', 'P6_9', 'P7_1', 'P7_2', 'P7_3', 'P7_4_1', 'P7_5_1',
'P7_2_1', 'P7_3_1', 'P7_2_2', 'P7_3_2', 'P7_1_3', 'P7_2_3', 'P7_3_3', 'P7_4_2', 'P7_5_2', 'P7_4_3', 'P7_5_3', 'P8_1',
'P8_2_1', 'P8_2_2', 'P8_2_3', 'P8_2_4', 'P8_2_5', 'P8_2_6', 'P8_2_7', 'P8_2_8', 'P8_2_9', 'P8_2_10', 'P8_2_11',
'P8_2_12', 'P8_2_13', 'P8_2_14', 'P8_2_15', 'P8_3_1', 'P8_3_2', 'P8_3_3', 'P8_3_4', 'P8_3_5', 'P8_3_6',
'P8_3_7', 'P8_3_8', 'P8_3_9', 'P8_3_10', 'P8_3_11', 'P8_3_12', 'P8_3_13', 'P8_3_14', 'P8_3_15', 'P8_3_16',
'P8_3_17', 'P8_4', 'P8_5', 'P8_6M', 'P8_6A', 'P8_7', 'P8_8', 'P8_9', 'P8_10', 'P8_11_1', 'P8_11_2', 'P8_11_3',
'P8_11_4', 'P8_11_5', 'P8_13', 'P8_14_1', 'P8_14_2', 'P8_14_3', 'P8_14_4', 'P8_14_5', 'P8_14_6', 'P8_14_7', 'P8_14_8',
'P8_15', 'P8_16', 'P8_17_1', 'P8_17_2', 'P8_17_3', 'P8_17_4', 'P8_17_5', 'P8_17_6', 'P8_17_7', 'P8_17_8',
'P8_17_9', 'P8_17_10', 'P8_17_11', 'P8_17_12', 'P8_17_13', 'P8_17_14', 'P8_17_15', 'P8_17_16', 'P8_18']
```

Imagen 3. Selección de variables relacionadas con la variable "Diabetes" (P3_1).

Una vez habiendo seleccionado las variables relacionadas con las diabetes, utilizamos la variable P3_1 que indica si los encuestados tienen o no diabetes para seleccionar una muestra con registros de personas con diabetes.

```
In [6]: # Utilizando la variable de Diabetes (P3_1) se realizó un filtrado para seleccionar a los pacientes con diabetes.
PDiabeticos = data[data['P3_1']>0]
PDiabeticos = PDiabeticos.sample(n=4555, random_state=1)
PDiabeticos
print(PDiabeticos['P3_1'].value_counts())

1    4555
Name: P3_1, dtype: int64
```

Imagen 4. Selección de la muestra con pacientes diabéticos.

Seleccionamos como variables independientes, las que indican información sobre complicaciones relacionadas con la diabetes. Decidimos realizar un análisis por separado para cada variable dependiente con la intención de incluir las demás variables referentes a complicaciones de la diabetes como variables independientes y con ello observar también la relación entre estas. De nuestra muestra, definimos un 80% como datos de entrenamiento y un 20% de prueba.

```
In [8]: Y = PDiabeticos['P3_18_1']
X = PDiabeticos.drop('P3_18_1',axis =1)

In [9]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=1234, shuffle      = True)
```

Imagen 5. Selección de variables y datos de prueba y entrenamiento

Aplicamos el modelo de regresión logística mediante la librería de statsmodel y definimos como parámetros para el modelo un nivel de significancia para el análisis de asociación de $(\alpha) = 0.5$. Utilizamos los coeficientes de las variables independientes para conocer el tipo de asociación con las variables dependientes y utilizamos los odds ratio para medir la razón de probabilidades entre los hábitos de los pacientes y las complicaciones de la diabetes. Por último, obtuvimos los parámetros resultantes y la precisión del modelo para realizar el análisis de los resultados.

```
In [11]: from sklearn.linear_model import LogisticRegression
modelo = LogisticRegression(random_state = 0)
modelo.fit(X_train, Y_train)
Y_predict_py = modelo.predict(X_test)

In [12]: print(f"La precisión del modelo es de: {100*modelo.score(X_test,Y_test)}%")
La precisión del modelo es de: 92.86498353457738%
```



```
In [13]: print("-----Modelo de Regresión Logística-----")
modelo_odds = pd.DataFrame(np.exp(modelo2.params), columns=['Odds Ratio'])
modelo_odds['P>|Z|']=modelo2.pvalues
modelo_odds[['2.5%', '97.5%']] = np.exp(modelo2.conf_int())
modelo_odds['Coeficiente']=modelo2.params
print(modelo_odds)
print("-----")
print(f"La precisión del modelo es de: {100*modelo.score(X_test,Y_test)}%")
print("-----")
```

Imagen 6. Aplicación del modelo de Regresión Logística mediante Statsmodel en Python.

5.5. Descripción del modelo.

Para el desarrollo del presente modelo utilizamos información de la Encuesta Nacional de Salud y Nutrición 2018 del INEGI. De esta investigación consideramos para el análisis el conjunto de datos "CS_Adultos.csv" que contiene información relacionada con la salud y los hábitos de los encuestados, para con ello realizar un modelo que nos permita determinar cómo es que infieren los hábitos de los pacientes con diabetes en la prevención de complicaciones propias de la enfermedad. En la base de datos encontramos en su mayoría registros con variables categóricas (21).

- Herramientas: Python Jupyter Notebook.
- Librerías: Matplotlib, Pandas, Numpy, Math, SKLearn y Statsmodel
- Método: Algoritmo de Regresión Logística (Statsmodels).
- Nombre de la base de datos: CS_Adultos.csv. (INEGI), (ENSANUT).
- Total, de registros: 43,070.
- Muestra: 4,555.
- Clases: 1 (Diabéticos).
- Datos de entrenamiento: (80%) = 3,644.
- Datos de prueba: (20%) = 911.
- Variables dependientes: 7 (Complicaciones).
- Variables independientes: 20 (Hábitos) + 8 (Complicaciones).
- Nivel de significancia (α) para el análisis de asociación: 0.05

5.6. Resultados para cada una de las variables.

A continuación, se muestran los resultados obtenidos de modelo propuesto para cada una de las variables sobre complicaciones de la diabetes. En estos se muestra tanto los resultados en Python con la precisión del modelo, la matriz de confusión y las tablas de relación entre las variables con asociación significativa como las observaciones sobre cada una de ellas.

5.6.1. Resultados para la variable “Ulceras en pies o piernas”.

De los 4555 registros de pacientes diabéticos 328 mencionaron presentar ulceras en pies o piernas como consecuencia de la enfermedad. Se realizó el análisis de la regresión logística mediante el modelo propuesto. A continuación, se presentan los resultados obtenidos:

Resultados del modelo en Python (Statsmodel), “Ulceras en pies o piernas”.

| -----Modelo de Regresión Logística----- | | | | | |
|-----------------------------------------|------------|--------------|----------|-----------|--------------|
| | Odds Ratio | P> Z | 2.5% | 97.5% | Coefficiente |
| P3_1 | 0.014630 | 3.787781e-30 | 0.007080 | 0.030230 | -4.224699 |
| P3_3 | 1.487800 | 1.115102e-01 | 0.912105 | 2.426859 | 0.397299 |
| P3_8 | 0.826658 | 4.326657e-01 | 0.513821 | 1.329963 | -0.190364 |
| P3_13_1 | 1.029816 | 8.479888e-01 | 0.762594 | 1.390677 | 0.029380 |
| P3_13_2 | 0.832856 | 3.997874e-01 | 0.544098 | 1.274861 | -0.182894 |
| P3_16_1 | 1.171059 | 4.224153e-01 | 0.796219 | 1.722362 | 0.157908 |
| P3_16_2 | 1.306448 | 3.206139e-01 | 0.770899 | 2.214048 | 0.267312 |
| P3_16_3 | 1.052499 | 7.733431e-01 | 0.743011 | 1.490001 | 0.051168 |
| P3_16_7 | 0.872266 | 6.009971e-01 | 0.522655 | 1.455738 | -0.136661 |
| P3_16_8 | 1.109897 | 7.658881e-01 | 0.558743 | 2.204720 | 0.104268 |
| P3_16_10 | 0.950178 | 8.328778e-01 | 0.591085 | 1.527427 | -0.051106 |
| P3_16_11 | 1.297398 | 1.885056e-01 | 0.880122 | 1.912507 | 0.260360 |
| P3_16_12 | 3.023680 | 4.127232e-04 | 1.636306 | 5.587366 | 1.106475 |
| P3_16_13 | 0.349898 | 8.725368e-03 | 0.159631 | 0.766949 | -1.050113 |
| P3_16_14 | 0.710876 | 3.950145e-01 | 0.323801 | 1.560064 | -0.341258 |
| P3_16_16 | 0.993152 | 9.750209e-01 | 0.645980 | 1.526907 | -0.006871 |
| P3_16_19 | 0.865622 | 4.293851e-01 | 0.605187 | 1.238131 | -0.144307 |
| P3_17 | 2.250689 | 4.686431e-04 | 1.428609 | 3.545826 | 0.811236 |
| P3_18_2 | 11.111817 | 1.294572e-16 | 6.281442 | 19.656709 | 2.408009 |
| P3_18_3 | 3.192499 | 5.102432e-14 | 2.359931 | 4.318793 | 1.160804 |
| P3_18_4 | 1.793692 | 1.894779e-02 | 1.101044 | 2.922074 | 0.584276 |
| P3_18_5 | 1.850972 | 1.738312e-01 | 0.762141 | 4.495359 | 0.615711 |
| P3_18_6 | 1.002522 | 9.952626e-01 | 0.436481 | 2.302624 | 0.002519 |
| P3_18_7 | 2.278181 | 5.551111e-02 | 0.980790 | 5.291765 | 0.823377 |
| P3_18_8 | 1.761090 | 1.206000e-01 | 0.861877 | 3.598469 | 0.565933 |
| P3_18_9 | 1.432957 | 8.742065e-02 | 0.948581 | 2.164670 | 0.359740 |
| P13_2 | 0.990594 | 9.651593e-01 | 0.648244 | 1.513747 | -0.009450 |
| P13_11 | 0.885990 | 4.586460e-01 | 0.643266 | 1.220300 | -0.121050 |

La precisión del modelo es de: 92.86498353457738%

Imagen 7. Resultados RL. Python (Statsmodel). Variable “Ulceras en pies o piernas”.

Matriz de confusión, “Ulceras en pies o piernas”.

| | | |
|---------------------|---------------------|----------------------|
| VN 838 | FN 63 | TFN 6.91% |
| FP 2 | VP 8 | TVN 91.98% |
| TFP 0.21% | TVP 0.87% | TP 92.89% |

Tabla 10. Matriz de confusión de resultados para la variable “Ulceras en pies o piernas”.

- **Nomenclatura.**

Total, de registros (TR) = 911

Varaderos positivos (VP) = 8

Verdaderos negativos (VN) = 838

Falsos positivos (FP) = 2
 Falsos negativos (FN) = 63
 Tasa de error (TE) = 7.14%
 Tasa de precisión (TP) = **92.86%**

- **Tasas de resultados.**

$$TVN = \frac{VN}{TR} = 0.9198 = \mathbf{91.98 \%}$$

$$TFN = \frac{FN}{TR} = 0.0691 = \mathbf{6.91 \%}$$

$$TVP = \frac{VP}{TR} = 0.0087 = \mathbf{0.87 \%}$$

$$TP = \frac{VN + VP}{TR} = 0.9286 = \mathbf{92.86 \%}$$

$$TFP = \frac{FP}{TR} = 0.0021 = \mathbf{0.21 \%}$$

$$TE = \frac{FN + FP}{TR} = 0.0714 = \mathbf{7.14 \%}$$

Tabla de relación entre variables con asociación significativa, “Ulceras en pies o piernas”.

| Variable Dependiente P3_18_1 Ulceras en pies o piernas | | | | |
|--------------------------------------------------------------|----------------------------|------------------------------------------------------------------|------------------------------------|--------------------------------|
| Característica | Hábitos de prevención | | Complicaciones | |
| Variables Independientes | P3_16_12 Dejó de fumar. | P3_17 Revisa sus pies con frecuencia en búsqueda de lesiones. | P3_18_3 Disminución de la vista | P3_18_4 Perdida de la vista |
| Asociación | Positiva | Positiva | Positiva | Positiva |
| Odds Ratio | 3.023680 | 2.250689 | 3.192499 | 1.793692 |
| Probabilidad | 3 - 1 | 2.2 - 1 | 3.1 - 1 | 1.7 - 1 |
| Tasa de probabilidad | 75.00% | 68.75% | 75.60% | 62.96% |

Tabla 11. Relación entre variables con asociación significativa (Ulceras en pies o piernas)

Observaciones, “Ulceras en pies o piernas”.

Con un nivel de significancia de $\alpha = 0.05$, para la variable de complicación de “Ulceras en pies o piernas” encontramos una asociación significativa tanto con las variables relacionadas con los hábitos del paciente: “Dejó de fumar” y “Revisa sus pies en búsqueda de lesiones” como con las variables “Disminución de la vista” y “perdida de la vista” referentes a otras complicaciones de la enfermedad. Con base en los resultados de asociación podemos observar que para el paciente diabético el hábito de revisar sus pies en búsqueda de lesiones puede reducir en un 68.57% la probabilidad de que desarrolle ulceras. Además, el que haya dejado de fumar como medida preventiva para el control de la enfermedad puede prevenir en un 75.00% la posibilidad de que desarrolle esta complicación. Por otro lado, en relación con la asociación con las variables referentes a otras complicaciones, observamos que la complicación de ulceras en pies y piernas se relaciona en un 75.60% con la disminución de la vista en el paciente y un 62.96% con la perdida de la vista.

5.6.2. Resultados para la variable “Amputación”.

De los 4555 registros de pacientes diabéticos 87 mencionaron haber sufrido la amputación de alguna extremidad como consecuencia de la enfermedad. Se realizó el análisis de la regresión logística mediante el modelo propuesto. A continuación, se presentan los resultados obtenidos:

Resultados del modelo en Python (Statsmodel), “Amputación”.

| -----Modelo de Regresión Logística----- | | | | | |
|-----------------------------------------|------------|--------------|----------|-----------|-------------|
| | Odds Ratio | P> Z | 2.5% | 97.5% | Coeficiente |
| P3_1 | 0.002259 | 9.548500e-14 | 0.000455 | 0.011230 | -6.092728 |
| P3_3 | 0.887380 | 8.086499e-01 | 0.337396 | 2.333883 | -0.119482 |
| P3_8 | 2.838236 | 1.001473e-01 | 0.818415 | 9.842908 | 1.043183 |
| P3_13_1 | 1.425088 | 2.529576e-01 | 0.776401 | 2.615759 | 0.354234 |
| P3_13_2 | 0.947709 | 8.945283e-01 | 0.428400 | 2.096529 | -0.053707 |
| P3_16_1 | 1.493686 | 2.888838e-01 | 0.711581 | 3.135409 | 0.401247 |
| P3_16_2 | 1.144291 | 7.892092e-01 | 0.425972 | 3.073919 | 0.134785 |
| P3_16_3 | 1.092927 | 8.101598e-01 | 0.529326 | 2.256627 | 0.088860 |
| P3_16_7 | 1.548812 | 3.473538e-01 | 0.621908 | 3.857192 | 0.437488 |
| P3_16_8 | 1.017118 | 9.773047e-01 | 0.315868 | 3.275200 | 0.016973 |
| P3_16_10 | 1.411698 | 4.407813e-01 | 0.587517 | 3.392056 | 0.344793 |
| P3_16_11 | 0.625222 | 2.898113e-01 | 0.262044 | 1.491741 | -0.469649 |
| P3_16_12 | 0.756949 | 6.948883e-01 | 0.188264 | 3.043456 | -0.278459 |
| P3_16_13 | 0.635413 | 4.917715e-01 | 0.174418 | 2.314836 | -0.453481 |
| P3_16_14 | 0.929783 | 9.179174e-01 | 0.232839 | 3.712848 | -0.072804 |
| P3_16_16 | 1.832000 | 1.337334e-01 | 0.830368 | 4.041850 | 0.605408 |
| P3_16_19 | 0.533198 | 9.706884e-02 | 0.253674 | 1.120729 | -0.628863 |
| P3_17 | 0.955569 | 9.125339e-01 | 0.424689 | 2.150070 | -0.045448 |
| P3_18_1 | 11.290320 | 2.441432e-16 | 6.324573 | 20.154931 | 2.423946 |
| P3_18_3 | 1.839463 | 5.675830e-02 | 0.982623 | 3.443463 | 0.609474 |
| P3_18_4 | 4.671135 | 2.549086e-05 | 2.279310 | 9.572856 | 1.541402 |
| P3_18_5 | 1.535404 | 5.850442e-01 | 0.329444 | 7.155883 | 0.428794 |
| P3_18_6 | 2.288913 | 1.863795e-01 | 0.670187 | 7.817484 | 0.828077 |
| P3_18_7 | 2.936847 | 1.168494e-01 | 0.763996 | 11.289416 | 1.077337 |
| P3_18_8 | 0.881612 | 8.626771e-01 | 0.211442 | 3.675896 | -0.126004 |
| P3_18_9 | 1.077621 | 8.576316e-01 | 0.476159 | 2.438821 | 0.074756 |
| P13_2 | 1.063795 | 8.945145e-01 | 0.426435 | 2.653768 | 0.061843 |
| P13_11 | 0.996440 | 9.915880e-01 | 0.513515 | 1.933522 | -0.003566 |

La precisión del modelo es de: 97.25576289791438%

Imagen 8. Resultados RL. Python (Statsmodel). Variable “Amputación”.

- **Matriz de confusión:**

| | | |
|---------------------|---------------------|----------------------|
| VN 885 | FN 25 | TFN 2.74% |
| FP 0 | VP 1 | TVN 97.14% |
| TFP 0.00% | TVP 0.10% | TP 97.25% |

Tabla 12. Matriz de confusión de resultados para la variable “Amputación”.

- **Nomenclatura:**

Total, de registros (TR) = 911

Varaderos positivos (VP) = 1

Verdaderos negativos (VN) = 885

Falsos positivos (FP) = 0

Falsos negativos (FN) = 25

Tasa de error (TE) = 2.75%

Tasa de precisión (TP) = **97.25%**

- **Tasas de resultados:**

$$TVN = \frac{VN}{TR} = 0.9714 = \mathbf{97.14 \%} \quad TFN = \frac{FN}{TR} = 0.0274 = \mathbf{2.74 \%}$$

$$TVP = \frac{VP}{TR} = 0.0010 = \mathbf{0.10 \%} \quad TP = \frac{VN + VP}{TR} = 0.9725 = \mathbf{97.25 \%}$$

$$TFP = \frac{FP}{TR} = 0.0000 = \mathbf{0.00 \%} \quad TE = \frac{FN + FP}{TR} = 0.0275 = \mathbf{2.75 \%}$$

Tabla de relación entre variables con asociación significativa, “Amputación”.

| Variable Dependiente P3_18_2 Amputación | | | | |
|-----------------------------------------------|---------------------------------------------|----------------------------------------------|-------------------------------------|---------------------------------|
| Característica | Hábitos de prevención | Complicaciones | | |
| Variables Independientes | P3_16_19 No realiza medidas preventivas. | 149 P3_18_1 Ulceras en pies o piernas. | P3_18_3 Disminución de la vista. | P3_18_4 Perdida de la vista. |
| Asociación | Negativa | Positiva | Positiva | Positiva |
| Odds Ratio | 0.53498 | 11.290320 | 1.839463 | 4.671135 |
| Probabilidad | 1.8 - 1 | 11.2 - 1 | 1.8 - 1 | 4.6 - 1 |
| Tasa de probabilidad | 64.28% | 91.80% | 68.28% | 82.14% |

Tabla 13. Relación entre variables con asociación significativa (Amputación).

Observaciones, “Amputación” .

Con un nivel de significancia de $\alpha = 0.05$, para la variable de complicación “Amputación”, encontramos una asociación significativa tanto con la variable “No realiza medidas preventivas, como con las variables, “Ulceras en pies o piernas”, “Disminución de la vista” y “Perdida de la vista”, referentes a otras complicaciones de la enfermedad.

Con base en los resultados de asociación logramos observar que en los pacientes diabéticos la presencia de ulceras en pies o piernas puede desencadenar la amputación de alguna de las extremidades inferiores con una probabilidad del 91.80%, además esta se relaciona con la disminución de la capacidad visual en un 68.28% y un 82.14% con la pérdida total de la vista.

Por otro lado, al no realizar medidas para prevenir complicaciones de la diabetes, el paciente tiene una probabilidad del 64.28% de terminar con la amputación de alguna de sus extremidades inferiores.

5.6.3. Resultados para la variable “Perdida de la vista”:

De los 4555 registros de pacientes diabéticos 226 mencionaron haber sufrido perdida de la vista como consecuencia de la enfermedad. Se realizó el análisis de la regresión logística mediante el modelo propuesto. A continuación, se presentan los resultados obtenidos:

Resultados del modelo en Python (Statsmodel), “Perdida de la vista”.

| -----Modelo de Regresión Logística----- | | | | | |
|-----------------------------------------|------------|--------------|----------|----------|-------------|
| | Odds Ratio | P> Z | 2.5% | 97.5% | Coeficiente |
| P3_1 | 0.030398 | 5.020624e-21 | 0.014683 | 0.062935 | -3.493363 |
| P3_3 | 0.911435 | 7.143892e-01 | 0.554670 | 1.497673 | -0.092735 |
| P3_8 | 0.851451 | 5.291645e-01 | 0.515982 | 1.405027 | -0.160814 |
| P3_13_1 | 1.469007 | 2.729011e-02 | 1.044043 | 2.066947 | 0.384587 |
| P3_13_2 | 0.906075 | 6.766586e-01 | 0.569959 | 1.440407 | -0.098633 |
| P3_16_1 | 1.072231 | 7.581350e-01 | 0.687866 | 1.671371 | 0.069742 |
| P3_16_2 | 1.252668 | 4.590902e-01 | 0.689967 | 2.274277 | 0.225276 |
| P3_16_3 | 1.068435 | 7.527752e-01 | 0.707724 | 1.612992 | 0.066195 |
| P3_16_7 | 1.174806 | 5.697398e-01 | 0.674097 | 2.047436 | 0.161103 |
| P3_16_8 | 0.674183 | 3.462852e-01 | 0.296795 | 1.531434 | -0.394254 |
| P3_16_10 | 1.099820 | 7.338302e-01 | 0.635542 | 1.903263 | 0.095146 |
| P3_16_11 | 0.732137 | 2.225877e-01 | 0.443607 | 1.208332 | -0.311788 |
| P3_16_12 | 1.342765 | 4.745856e-01 | 0.598610 | 3.012012 | 0.294731 |
| P3_16_13 | 1.632527 | 1.460227e-01 | 0.843091 | 3.161160 | 0.490129 |
| P3_16_14 | 0.908296 | 8.211947e-01 | 0.394431 | 2.091625 | -0.096185 |
| P3_16_16 | 0.675429 | 1.527921e-01 | 0.394420 | 1.156648 | -0.392407 |
| P3_16_19 | 0.965921 | 8.685349e-01 | 0.640664 | 1.456306 | -0.034673 |
| P3_17 | 1.026786 | 9.023760e-01 | 0.673044 | 1.566451 | 0.026434 |
| P3_18_1 | 1.697248 | 3.659940e-02 | 1.033512 | 2.787246 | 0.529008 |
| P3_18_2 | 4.515220 | 2.405042e-05 | 2.243209 | 9.088412 | 1.507454 |
| P3_18_3 | 0.998970 | 9.951127e-01 | 0.718293 | 1.389321 | -0.001031 |
| P3_18_5 | 3.519520 | 5.684804e-03 | 1.442680 | 8.586120 | 1.258325 |
| P3_18_6 | 1.246400 | 6.288115e-01 | 0.510283 | 3.044411 | 0.220259 |
| P3_18_7 | 1.389236 | 5.634391e-01 | 0.455422 | 4.237778 | 0.328754 |
| P3_18_8 | 2.063327 | 7.158889e-02 | 0.938366 | 4.536949 | 0.724320 |
| P3_18_9 | 2.429154 | 1.194686e-04 | 1.545557 | 3.817906 | 0.887543 |
| P13_2 | 1.414595 | 2.060306e-01 | 0.826350 | 2.421588 | 0.346843 |
| P13_11 | 0.987413 | 9.466558e-01 | 0.681322 | 1.431020 | -0.012667 |

La precisión del modelo es de: 94.7310647639956%

Imagen 9. Resultados RL. Python (Statsmodel). Variable “Perdida de la vista”.

- **Matriz de confusión:**

| | | |
|---------------------|---------------------|----------------------|
| VN 863 | FN 47 | TFN 5.15% |
| FP 1 | VP 0 | TVN 94.73% |
| TFP 0.10% | TVP 0.00% | TP 94.73% |

Tabla 14. Matriz de confusión de resultados para la variable “Perdida de la vista”.

- **Nomenclatura:**

Total, de registros (TR) = 911

Varaderos positivos (VP) = 0

Verdaderos negativos (VN) = 863

Falsos positivos (FP) = 1

Falsos negativos (FN) = 47

Tasa de error (TE) = 5.27%

Tasa de precisión (TP) = **94.73%**

- **Tasas de resultados:**

$$TVN = \frac{VN}{TR} = 0.9473 = \mathbf{94.73\%}$$

$$TFN = \frac{FN}{TR} = 0.0515 = \mathbf{5.15\%}$$

$$TVP = \frac{VP}{TR} = 0.0000 = \mathbf{0.00\%}$$

$$TP = \frac{VN + VP}{TR} = 0.9473 = \mathbf{94.73\%}$$

$$TFP = \frac{FP}{TR} = 0.0010 = \mathbf{0.10\%}$$

$$TE = \frac{FN + FP}{TR} = 0.0527 = \mathbf{5.27\%}$$

Tabla de relación entre variables con asociación significativa, “Perdida de la vista”:

| Variable Dependiente P3_18_4 Perdida de la vista | | | | |
|--------------------------------------------------------|----------------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------------------------|
| Característica | Hábitos de prevención | Complicaciones | | |
| Variables Independientes | P3_13_1 Lleva a cabo un plan alimenticio. | P3_18_1 Ulceras en pies o piernas. | P3_18_5 Insuficiencia renal aguda. | P3_18_9 Perdida del conocimiento confusión o desmayo |
| Asociación | Positiva | Positiva | Positiva | Positiva |
| Odds Ratio | 1.469007 | 1.697248 | 3.519520 | 2.429154 |
| Probabilidad | 1.4 - 1 | 1.7 - 1 | 3.5 - 1 | 2.4 - 1 |
| Tasa de probabilidad | 59.34% | 62.96% | 77.77% | 70.58% |

Tabla 15. Relación entre variables con asociación significativa (Perdida de la vista).

Observaciones, “Perdida de la vista”.

Con un nivel de significancia de $\alpha = 0.05$, para la variable de complicación de “Perdida de la vista” encontramos una asociación significativa con la variable “Lleva a cabo un plan alimenticio”. Así también observamos una asociación significativa con las variables “Ulceras en pies o piernas”, “Insuficiencia renal aguda” y “Perdida del conocimiento, confusión o desmayo”, referentes otras complicaciones de la enfermedad.

Con base en los resultados obtenidos, podemos deducir que el llevar a cabo un plan alimenticio adecuado podría disminuir la posibilidad de que el paciente termine con la pérdida de la vista en un 59.34%.

Así también logramos distinguir una relación entre la presencia de ulceras en pies o piernas en el paciente diabético con la complicación de perdida de la vista con un 62.96%, un 77.77% con problemas de insuficiencia renal aguda y un 70.58% con eventos de pérdida del conocimiento, confusión o desmayo.

5.6.4. Resultados para la variable “Insuficiencia renal aguda”.

De los 4555 registros de pacientes diabéticos 53 mencionaron haber sufrido insuficiencia renal aguda como consecuencia de la enfermedad. Se realizó el análisis de la regresión logística mediante el modelo propuesto. A continuación, se presentan los resultados obtenidos:

Resultados del modelo en Python (Statsmodel), “Insuficiencia renal aguda”.

| -----Modelo de Regresión Logística----- | | | | | |
|-----------------------------------------|---------------|--------------|----------|-----------|--------------|
| | Odds Ratio | P> Z | 2.5% | 97.5% | Coefficiente |
| P3_1 | 8.184011e-04 | 4.017132e-10 | 0.000088 | 0.007595 | -7.108158 |
| P3_3 | 1.252474e+00 | 7.291727e-01 | 0.350192 | 4.479522 | 0.225121 |
| P3_8 | 1.610310e-01 | 2.525002e-04 | 0.060558 | 0.428204 | -1.826158 |
| P3_13_1 | 9.980320e-01 | 9.961033e-01 | 0.452702 | 2.200274 | -0.001970 |
| P3_13_2 | 2.109890e-01 | 1.172902e-01 | 0.030106 | 1.478643 | -1.555949 |
| P3_16_1 | 7.610584e-01 | 6.415073e-01 | 0.241120 | 2.402164 | -0.273045 |
| P3_16_2 | 3.456313e-39 | 1.000000e+00 | 0.000000 | inf | -88.560616 |
| P3_16_3 | 9.499479e-01 | 9.185931e-01 | 0.354860 | 2.542974 | -0.051348 |
| P3_16_7 | 1.388108e+00 | 6.295648e-01 | 0.366179 | 5.262033 | 0.327942 |
| P3_16_8 | 1.395224e+00 | 6.999492e-01 | 0.256467 | 7.590238 | 0.333055 |
| P3_16_10 | 3.054779e+00 | 3.436706e-02 | 1.085694 | 8.595125 | 1.116707 |
| P3_16_11 | 1.130942e+00 | 8.195913e-01 | 0.392818 | 3.256043 | 0.123051 |
| P3_16_12 | 3.666830e+00 | 1.004037e-01 | 0.778211 | 17.277642 | 1.299328 |
| P3_16_13 | 8.946500e-02 | 1.786084e-01 | 0.002655 | 3.014867 | -2.413908 |
| P3_16_14 | 3.154555e-01 | 4.818295e-01 | 0.012683 | 7.851116 | -1.153421 |
| P3_16_16 | 4.124279e-01 | 1.918950e-01 | 0.109054 | 1.559750 | -0.885694 |
| P3_16_19 | 1.783937e+00 | 2.658528e-01 | 0.643536 | 4.945227 | 0.578823 |
| P3_17 | 1.527066e+00 | 4.730558e-01 | 0.480433 | 4.853808 | 0.423348 |
| P3_18_1 | 1.653065e+00 | 3.067178e-01 | 0.630542 | 4.333770 | 0.502631 |
| P3_18_2 | 2.302783e+00 | 3.088809e-01 | 0.461861 | 11.481392 | 0.834119 |
| P3_18_3 | 3.372876e+00 | 8.985334e-03 | 1.354871 | 8.396588 | 1.215766 |
| P3_18_4 | 3.8660790e+00 | 5.050015e-03 | 1.501857 | 9.926168 | 1.350872 |
| P3_18_6 | 7.017880e+00 | 8.916488e-04 | 2.233551 | 22.149544 | 1.948461 |
| P3_18_7 | 4.425814e+00 | 7.629873e-02 | 0.854495 | 22.923281 | 1.487454 |
| P3_18_8 | 3.055417e+00 | 6.398193e-02 | 0.937146 | 9.961706 | 1.116916 |
| P3_18_9 | 3.975383e+00 | 1.445336e-03 | 1.700545 | 9.293294 | 1.380121 |
| P13_2 | 1.117155e+00 | 8.629610e-01 | 0.317525 | 3.930513 | 0.110786 |
| P13_11 | 5.027689e+00 | 3.454692e-02 | 1.124582 | 22.477386 | 1.614961 |

La precisión del modelo es de: 98.35345773874863%

Imagen 10. Resultados del modelo. Python (Statsmodel). Insuficiencia renal”.

- **Matriz de confusión:**

| | | |
|---------------------|---------------------|----------------------|
| VN 895 | FN 14 | TFN 1.53% |
| FP 1 | VP 1 | TVN 98.24% |
| TFP 0.10% | TVP 0.10% | TP 98.35% |

Tabla 16. Matriz de confusión de resultados para la variable “Insuficiencia renal aguda”.

- **Nomenclatura:**

Total, de registros (TR) = 911

Varaderos positivos (VP) = 1

Verdaderos negativos (VN) = 895

Falsos positivos (FP) = 1

Falsos negativos (FN) = 14

Tasa de error (TE) = 1.65%

Tasa de precisión (TP) = **98.35%**

- **Tasas de resultados:**

$$TVN = \frac{VN}{TR} = 0.9824 = \mathbf{98.24\%} \quad TFN = \frac{FN}{TR} = 0.0153 = \mathbf{1.53\%}$$

$$TVP = \frac{VP}{TR} = 0.0010 = \mathbf{0.10\%} \quad TP = \frac{VN + VP}{TR} = 0.9835 = \mathbf{98.35\%}$$

$$TFP = \frac{FP}{TR} = 0.0010 = \mathbf{0.10\%} \quad TE = \frac{FN + FP}{TR} = 0.0165 = \mathbf{1.65\%}$$

Tabla de relación entre variables con asociación significativa, “Insuficiencia renal aguda”.

| Variable Dependiente P3_18_5 Insuficiencia renal aguda | | | | | | | |
|--------------------------------------------------------------|--------------------------------------------------------|-------------------------------------------------------|------------------------------------|---------------------------------|-----------------------------|--------------------------|------------------------------------------------------|
| Caract. | Hábitos de prevención | | | Complicaciones | | | |
| Variables Independientes | P3_8 Toma pastillas/insulina para controlar la glucosa | P3_16_10 Toma medicamento para control del colesterol | P13_11 Consumo bebidas alcohólicas | P3_18_3 Disminución de la vista | P3_18_4 Pérdida de la vista | P3_18_6 Infarto cardiaco | P3_18_9 Pérdida del conocimiento confusión o desmayo |
| Asociación | Negativa | Positiva | Positiva | Positiva | Positiva | Positiva | Positiva |
| Odds Ratio | 0.161031 | 3.054779 | 5.027689 | 3.372876 | 3.86079 | 7.01788 | 3.975383 |
| Probabilidad | 6.2 - 1 | 3 - 1 | 5 - 1 | 3.3 - 1 | 3.8 - 1 | 7 - 1 | 3.9 - 1 |
| Tasa de probabilidad | 86.11% | 75.00% | 83.33% | 76.74% | 79.16% | 87.50% | 79.59% |

Tabla 17. Relación entre variables con asociación significativa (Insuficiencia renal aguda).

Observaciones, “Insuficiencia renal aguda”.

Con un nivel de significancia de $\alpha = 0.05$, para la variable de complicación de “Insuficiencia renal aguda” encontramos una asociación significativa tanto con las variables relacionadas con los hábitos del paciente, “Toma pastillas o insulina para el control de los niveles de glucosa”, “Toma medicamento para el control del colesterol” y “Consumo bebidas alcohólicas” como con las variables “Disminución de la vista”, “Pérdida de la vista”, “Infarto cardiaco” y “Pérdida del conocimiento, confusión o desmayo” referentes a otras complicaciones de la enfermedad.

Con base en los resultados de asociación podemos observar que para el paciente diabético el llevar acabo un adecuado control de los niveles de glucosa en la sangre, con el apoyo de medicamentos o la aplicación de insulina reduce en un 86.11% la probabilidad de que desarrolle Insuficiencia renal aguda.

Además, el que requiera del consumo de medicamento para el control del colesterol puede aumentar la probabilidad de que desarrolle este problema en un 75.00%. Asimismo, el consumo de bebidas alcohólicas representa una probabilidad del 83.33% de desencadenar esta enfermedad en el paciente.

Por otro lado, la insuficiencia renal aguda en el paciente diabético se relaciona en un 76.74% con la disminución de la vista, un 79.16% con la pérdida total de la vista, un 87.5% con infarto cardiaco y un 79.59% con incidentes de pérdida del conocimiento, confusión o desmayo.

5.6.5. Resultados para la variable “Infarto cardiaco”.

De los 4555 registros de pacientes diabéticos 94 mencionaron haber sufrido un infarto cardiaco como consecuencia de la enfermedad. Se realizó el análisis de la regresión logística mediante el modelo propuesto. A continuación, se presentan los resultados obtenidos:

- **Matriz de confusión:**

| | | |
|---------------------|---------------------|----------------------|
| VN 888 | FN 22 | TFN 2.41% |
| FP 0 | VP 1 | TVN 97.47% |
| TFP 0.00% | TVP 0.10% | TP 97.69% |

Tabla 18. Matriz de confusión de resultados para la variable “Infarto cardiaco”.

- **Nomenclatura:**

Total, de registros (TR) = 911

Varaderos positivos (VP) = 1

Verdaderos negativos (VN) = 888

Falsos positivos (FP) = 0

Falsos negativos (FN) = 22

Tasa de error (TE) = 2.31%

Tasa de precisión (TP) = **97.69%**

- **Tasas de resultados:**

$$TVN = \frac{VN}{TR} = \frac{888}{911} = 0.9747 = \mathbf{97.47 \%}$$

$$TFN = \frac{FN}{TR} = \frac{22}{911} = 0.0241 = \mathbf{2.41 \%}$$

$$TVP = \frac{VP}{TR} = \frac{1}{911} = 0.0010 = \mathbf{0.10 \%}$$

$$TP = \frac{VN + VP}{TR} = \frac{888 + 1}{911} = 0.9769 = \mathbf{97.69 \%}$$

$$TFP = \frac{FP}{TR} = \frac{0}{911} = 0.0000 = \mathbf{0.00 \%}$$

$$TE = \frac{FN + FP}{TR} = \frac{22 + 0}{911} = 0.0231 = \mathbf{2.31 \%}$$

Resultados del modelo en Python (Statsmodel), “Infarto cardiaco”.

| -----Modelo de Regresión Logística----- | | | | | | |
|-----------------------------------------|------------|--------------|----------|-----------|--------------|--|
| | Odds Ratio | P> Z | 2.5% | 97.5% | Coefficiente | |
| P3_1 | 0.003974 | 2.293851e-13 | 0.000906 | 0.017422 | -5.528074 | |
| P3_3 | 0.857142 | 7.291188e-01 | 0.358218 | 2.050964 | -0.154152 | |
| P3_8 | 3.827462 | 4.176738e-02 | 1.051298 | 13.934751 | 1.342202 | |
| P3_13_1 | 1.846366 | 2.423281e-02 | 1.083132 | 3.147415 | 0.613219 | |
| P3_13_2 | 0.684152 | 3.189280e-01 | 0.324321 | 1.443210 | -0.379576 | |
| P3_16_1 | 1.129151 | 7.212755e-01 | 0.579347 | 2.200721 | 0.121466 | |
| P3_16_2 | 1.943673 | 9.232387e-02 | 0.896510 | 4.213968 | 0.664580 | |
| P3_16_3 | 0.904289 | 7.576561e-01 | 0.477277 | 1.713338 | -0.100607 | |
| P3_16_7 | 1.188477 | 6.749186e-01 | 0.530329 | 2.663402 | 0.172673 | |
| P3_16_8 | 0.915395 | 8.718943e-01 | 0.312591 | 2.680651 | -0.088399 | |
| P3_16_10 | 1.348477 | 4.207846e-01 | 0.651227 | 2.792251 | 0.298976 | |
| P3_16_11 | 1.250433 | 5.047095e-01 | 0.648474 | 2.411174 | 0.223490 | |
| P3_16_12 | 0.669009 | 6.001564e-01 | 0.148833 | 3.007223 | -0.401958 | |
| P3_16_13 | 0.726266 | 5.818321e-01 | 0.232669 | 2.267007 | -0.319840 | |
| P3_16_14 | 1.128223 | 8.424278e-01 | 0.343406 | 3.706655 | 0.120643 | |
| P3_16_16 | 0.902977 | 7.906287e-01 | 0.425079 | 1.918157 | -0.102058 | |
| P3_16_19 | 0.958594 | 9.039032e-01 | 0.482496 | 1.904478 | -0.042288 | |
| P3_17 | 0.936081 | 8.508919e-01 | 0.470124 | 1.863861 | -0.066054 | |
| P3_18_1 | 1.249875 | 5.685111e-01 | 0.580672 | 2.690309 | 0.223044 | |
| P3_18_2 | 2.062636 | 2.267211e-01 | 0.637726 | 6.671307 | 0.723985 | |
| P3_18_3 | 1.478672 | 1.478336e-01 | 0.870595 | 2.511470 | 0.391145 | |
| P3_18_4 | 1.339452 | 5.024186e-01 | 0.570173 | 3.146644 | 0.292260 | |
| P3_18_5 | 4.508113 | 6.682435e-03 | 1.518446 | 13.384136 | 1.505879 | |
| P3_18_7 | 2.780060 | 9.788676e-02 | 0.828388 | 9.329840 | 1.022472 | |
| P3_18_8 | 4.523171 | 4.552016e-04 | 1.945410 | 10.516591 | 1.509213 | |
| P3_18_9 | 2.842038 | 1.150719e-03 | 1.514016 | 5.334939 | 1.044522 | |
| P13_2 | 0.514438 | 4.992585e-02 | 0.264704 | 0.999785 | -0.664680 | |
| P13_11 | 1.240155 | 4.988769e-01 | 0.664599 | 2.314155 | 0.215236 | |

La precisión del modelo es de: 97.6905311778291%

Imagen 11. Resultados RL. Python (Statsmodel). Infarto cardiaco”.

Tabla de relación entre variables con asociación significativa:

| | | Variable Dependiente P3_18_6 Infarto cardiaco | | | | |
|--------------------------|--|-----------------------------------------------------------|---------------------------------------------|-------------------------------------|------------------------|---------------------------|
| Característica | | Hábitos de prevención | | | Complicaciones | |
| Variables Independientes | | P3_8 Toma pastillas/insulina para controlar la glucosa | P3_13_1 Lleva a cabo un plan alimenticio | P3_16_2 Toma una aspirina diaria | P13_2 Dejó de fumar | P3_18_8 Coma diabético |
| Asociación | | Positiva | Positiva | Positiva | Negativa | Positiva |
| Odds Ratio | | 3.827462 | 1.846366 | 1.943673 | 0.514438 | 4.523171 |
| Probabilidad | | 3.8 - 1 | 1.8 - 1 | 1.9 - 1 | 1.9 - 1 | 4.5 - 1 |
| Tasa de probabilidad | | 79.16% | 64.28% | 65.51% | 65.51% | 81.81% |
| | | | | | | 73.68% |

Tabla 19. Relación entre variables con asociación significativa (Infarto cardiaco).

Observaciones, “Infarto cardiaco”.

Con un nivel de significancia de $\alpha = 0.05$, para la variable de complicación “Infarto cardiaco” encontramos una asociación significativa tanto con las variables relacionadas con los hábitos del paciente, “Toma pastillas o insulina para el control de los niveles de glucosa”, “Lleva a cabo un plan alimenticio”, “Dejó de fumar” y “Toma una aspirina diaria” como con las variables “Coma diabético” y “Perdida del conocimiento, confusión o desmayo”, referentes a otras complicaciones de la enfermedad.

Con base en los resultados de asociación podemos observar que para el paciente diabético el llevar acabo un adecuado control de los niveles de glucosa en la sangre, con el apoyo de medicamentos o la aplicación de insulina reduce en un 79.16% la probabilidad de presentar un infarto cardiaco. Además, el llevar a cabo un plan alimenticio reduce en un 64.28% la probabilidad de presentar esta afección. Asimismo, el consumir una aspirina diaria como medida preventiva puede reducir 65.51% la probabilidad de que presente dicha complicación, al igual que dejar el hábito del consumo de tabaco.

Por otro lado, los eventos de infarto cardiaco en el paciente diabético se relacionan en un 81.81% con eventos de coma diabético, y un 73.68% con incidentes de pérdida del conocimiento, confusión o desmayo.

5.6.6. Resultados para la variable “Infarto cerebral”.

De los 4555 registros de pacientes diabéticos 53 mencionaron haber sufrido un infarto cerebral como consecuencia de la enfermedad. Se realizó el análisis de la regresión logística mediante el modelo propuesto. A continuación, se presentan los resultados obtenidos:

- **Matriz de confusión:**

| | | |
|---------------------|---------------------|----------------------|
| VN 898 | FN 12 | TFN 1.31% |
| FP 0 | VP 1 | TVN 98.57% |
| TFP 0.00% | TVP 0.10% | TP 98.68% |

Tabla 20. Matriz de confusión de resultados para la variable “Infarto cerebral”.

- **Nomenclatura:**

Total, de registros (TR) = 911
 Varaderos positivos (VP) = 1
 Verdaderos negativos (VN) = 898
 Falsos positivos (FP) = 0
 Falsos negativos (FN) = 12
 Tasa de error (TE) = 1.32%
 Tasa de precisión (TP) = **98.68%**

- **Tasas de resultados:**

$$TVN = \frac{VN}{TR} = 0.9857 = \mathbf{98.57\%} \quad TFN = \frac{FN}{TR} = 0.0131 = \mathbf{1.31\%}$$

$$TVP = \frac{VP}{TR} = 0.0010 = \mathbf{0.10\%} \quad TP = \frac{VN + VP}{TR} = 0.9868 = \mathbf{98.68\%}$$

$$TFP = \frac{FP}{TR} = 0.0000 = \mathbf{0.00\%} \quad TE = \frac{FN + FP}{TR} = 0.0132 = \mathbf{1.32\%}$$

Resultados del modelo en Python (Statsmodel), “Infarto cerebral”.

| -----Modelo de Regresión Logística----- | | | | | | |
|-----------------------------------------|------------|--------------|----------|-----------|--------------|--|
| | Odds Ratio | P> Z | 2.5% | 97.5% | Coefficiente | |
| P3_1 | 0.002079 | 5.313233e-11 | 0.000329 | 0.013152 | -6.175821 | |
| P3_3 | 1.997510 | 3.446063e-01 | 0.475699 | 0.387756 | 0.691901 | |
| P3_8 | 0.858492 | 8.056672e-01 | 0.254585 | 2.894939 | -0.152578 | |
| P3_13_1 | 0.843484 | 6.618071e-01 | 0.393412 | 1.808450 | -0.170214 | |
| P3_13_2 | 1.593680 | 3.208430e-01 | 0.635933 | 3.999503 | 0.466046 | |
| P3_16_1 | 1.579553 | 3.356134e-01 | 0.622886 | 4.005526 | 0.457142 | |
| P3_16_2 | 0.748009 | 6.719755e-01 | 0.195102 | 2.867820 | -0.290340 | |
| P3_16_3 | 0.671988 | 4.073245e-01 | 0.262427 | 1.720734 | -0.397515 | |
| P3_16_7 | 0.709550 | 5.754285e-01 | 0.213551 | 2.357565 | -0.343125 | |
| P3_16_8 | 1.178958 | 8.299750e-01 | 0.262363 | 5.297777 | 0.164631 | |
| P3_16_10 | 0.808161 | 7.069149e-01 | 0.266267 | 2.452892 | -0.212995 | |
| P3_16_11 | 2.375997 | 4.978485e-02 | 0.000814 | 5.640770 | 0.865417 | |
| P3_16_12 | 2.736505 | 1.123471e-01 | 0.789773 | 9.481789 | 1.006682 | |
| P3_16_13 | 1.042694 | 9.509900e-01 | 0.274895 | 3.955008 | 0.041808 | |
| P3_16_14 | 1.896289 | 3.640906e-01 | 0.476176 | 7.551643 | 0.639899 | |
| P3_16_16 | 0.853839 | 7.690877e-01 | 0.297316 | 2.452076 | -0.158013 | |
| P3_16_19 | 0.686604 | 4.193135e-01 | 0.275686 | 1.710005 | -0.375997 | |
| P3_17 | 1.406523 | 5.053132e-01 | 0.515546 | 3.837301 | 0.341121 | |
| P3_18_1 | 1.981290 | 1.360187e-01 | 0.806403 | 4.867926 | 0.683748 | |
| P3_18_2 | 1.885563 | 3.991080e-01 | 0.431726 | 8.235198 | 0.634226 | |
| P3_18_3 | 2.443039 | 2.582437e-02 | 1.113777 | 5.358735 | 0.893243 | |
| P3_18_4 | 1.462597 | 5.045905e-01 | 0.478758 | 4.468207 | 0.380214 | |
| P3_18_5 | 2.613456 | 2.117381e-01 | 0.578636 | 11.803881 | 0.960673 | |
| P3_18_6 | 2.599378 | 1.352596e-01 | 0.742146 | 9.104366 | 0.955272 | |
| P3_18_8 | 4.934589 | 4.148749e-03 | 1.656771 | 14.697364 | 1.596269 | |
| P3_18_9 | 2.123892 | 9.118455e-02 | 0.886257 | 5.089853 | 0.753250 | |
| P13_2 | 0.979473 | 9.690150e-01 | 0.343956 | 2.789220 | -0.020740 | |
| P13_11 | 0.888417 | 7.707681e-01 | 0.400847 | 1.969046 | -0.118314 | |

La precisión del modelo es de: 98.6827661909989%

Imagen 12. Resultados RL. Python (Statsmodel). Infarto cerebral”.

Tabla de relación entre variables con asociación significativa, “Infarto cerebral”.:

| Variable Dependiente P3_18_7 Infarto cerebral | | | | |
|-----------------------------------------------------|---------------------------------------------------------------------|-----------------------------------------------------|---------------------------|---------------------------------------------------------|
| Característica | Hábitos de prevención | Complicaciones | | |
| Variables Independientes | P3_16_11 Toma medicamento para el control de la tensión arterial | P3_18_3 ⁸⁶ Disminución de la vista | P3_18_8 Coma diabético | P3_18_9 Perdida del conocimiento confusión o desmayo |
| Asociación | Positiva | Positiva | Positiva | Positiva |
| Odds Ratio | 2.375997 | 2.443039 | 4.934589 | 2.123892 |
| Probabilidad | 2.3 - 1 | 2.4 - 1 | 4.9 - 1 | 2 - 1 |
| Tasa de probabilidad | 69.69% | 70.58% | 83.05% | 66.66% |

Tabla 21. Relación entre variables con asociación significativa (Infarto cerebral).

Observaciones, “Infarto cerebral”.

Con un nivel de significancia de $\alpha = 0.05$, para la variable de complicación “Infarto cerebral” encontramos una asociación significativa tanto con la variable relacionada con el hábito del paciente, “Toma medicamento para el control de la tensión arterial” como con las variables “Disminución de la vista”, “Coma diabético” y “Perdida del conocimiento, confusión o desmayo”, referentes a otras complicaciones de la enfermedad.

Con base en los resultados de asociación podemos observar que para el paciente diabético que además es hipertenso, el llevar acabo un adecuado control de los niveles de tensión arterial con el apoyo de medicamentos reduce en un 69.69% la probabilidad de presentar un infarto cerebral.

Por otro lado, los eventos de infarto cerebral en el paciente diabético se relacionan en un 70.58% con disminución de la capacidad visual, un 83.05% con eventos de coma diabético y un 66.66% con incidentes de pérdida del conocimiento, confusión o desmayo.

5.6.7. Resultados para la variable “Coma diabético”.

De los 4555 registros de pacientes diabéticos 90 mencionaron haber sufrido un coma diabético como consecuencia de la enfermedad. Se realizó el análisis de la regresión logística mediante el modelo propuesto. A continuación, se presentan los resultados obtenidos:

Resultados del modelo en Python (Statsmodel), “Coma diabético”.

| | Odds Ratio | P> Z | 2.5% | 97.5% | Coefficiente |
|----------|------------|--------------|----------|-----------|--------------|
| P3_1 | 0.003434 | 1.595467e-13 | 0.000761 | 0.015501 | -5.673913 |
| P3_3 | 2.013553 | 2.209243e-01 | 0.656543 | 6.175372 | 0.699901 |
| P3_8 | 1.016633 | 9.766366e-01 | 0.337056 | 3.066384 | 0.016496 |
| P3_13_1 | 5.396428 | 6.921982e-03 | 0.202549 | 0.775890 | 0.925260 |
| P3_13_2 | 1.652533 | 2.329533e-01 | 0.723919 | 3.772339 | 0.502309 |
| P3_16_1 | 1.105079 | 7.998925e-01 | 0.510358 | 2.392831 | 0.099917 |
| P3_16_2 | 1.714200 | 2.613865e-01 | 0.669267 | 4.390596 | 0.538946 |
| P3_16_3 | 0.569932 | 1.448975e-01 | 0.267629 | 1.213703 | -0.562239 |
| P3_16_7 | 1.978675 | 1.247257e-01 | 0.827956 | 4.728699 | 0.682427 |
| P3_16_8 | 1.645871 | 3.728915e-01 | 0.550072 | 4.924608 | 0.498270 |
| P3_16_10 | 0.767190 | 5.727284e-01 | 0.305461 | 1.926864 | -0.265020 |
| P3_16_11 | 1.295765 | 4.899592e-01 | 0.620962 | 2.703880 | 0.259101 |
| P3_16_12 | 1.613366 | 4.825895e-01 | 0.424490 | 6.131955 | 0.478323 |
| P3_16_13 | 0.417463 | 2.238812e-01 | 0.102156 | 1.705979 | -0.873559 |
| P3_16_14 | 0.348977 | 2.218257e-01 | 0.064462 | 1.889260 | -1.052749 |
| P3_16_16 | 0.902372 | 8.088946e-01 | 0.393887 | 2.067286 | -0.102728 |
| P3_16_19 | 1.158484 | 6.894448e-01 | 0.563025 | 2.383707 | 0.147113 |
| P3_17 | 0.761818 | 4.779506e-01 | 0.359349 | 1.615047 | -0.272048 |
| P3_18_1 | 1.852091 | 1.081175e-01 | 0.873276 | 3.928013 | 0.616315 |
| P3_18_2 | 0.457651 | 3.705562e-01 | 0.082698 | 2.532635 | -0.781648 |
| P3_18_3 | 1.846405 | 4.994386e-02 | 1.000150 | 3.408698 | 0.613240 |
| P3_18_4 | 2.078557 | 7.574496e-02 | 0.926981 | 4.660722 | 0.731674 |
| P3_18_5 | 2.172017 | 2.145629e-01 | 0.638103 | 7.393255 | 0.775656 |
| P3_18_6 | 5.637617 | 1.079277e-04 | 2.349110 | 13.529690 | 1.729461 |
| P3_18_7 | 5.329349 | 3.548446e-03 | 1.730626 | 16.411385 | 1.673229 |
| P3_18_9 | 15.668369 | 3.428238e-21 | 8.853997 | 27.727339 | 2.751644 |
| P13_2 | 0.858138 | 7.188950e-01 | 0.373041 | 1.974050 | -0.152990 |
| P13_11 | 0.982048 | 9.584506e-01 | 0.496772 | 1.941370 | -0.018115 |

La precisión del modelo es de: 97.69484083424807%

Imagen 12. Resultados RL. Python (Statsmodel). Variable “Coma diabético”.

- **Matriz de confusión:**

| | | |
|---------------------|---------------------|----------------------|
| VN 889 | FN 20 | TFN 2.19% |
| FP 1 | VP 1 | TVN 97.58% |
| TFP 0.10% | TVP 0.10% | TP 97.69% |

Tabla 22. Matriz de confusión de resultados para la variable “Coma diabético”.

- **Nomenclatura:**

Total, de registros (TR) = 911

Varaderos positivos (VP) = 1

Verdaderos negativos (VN) = 889

Falsos positivos (FP) = 1

Falsos negativos (FN) = 20

Tasa de error (TE) = 2.31%

Tasa de precisión (TP) = **97.69%**

- **Tasas de resultados:**

$$TVN = \frac{VN}{TR} = 0.9758 = \mathbf{97.58\%}$$

$$TFN = \frac{FN}{TR} = 0.0219 = \mathbf{2.19\%}$$

$$TVP = \frac{VP}{TR} = 0.0010 = \mathbf{0.10\%}$$

$$TP = \frac{VN + VP}{TR} = 0.9769 = \mathbf{97.69\%}$$

$$TFP = \frac{FP}{TR} = 0.0010 = \mathbf{0.10\%}$$

$$TE = \frac{FN + FP}{TR} = 0.0231 = \mathbf{2.31\%}$$

Tabla de relación entre variables con asociación significativa, “Coma diabético”.

| Variable Dependiente P3_18_8 Coma diabético | | | | | |
|---------------------------------------------------|---------------------------------------------|------------------------------------|-----------------------------|-----------------------------|---------------------------------------------------------|
| Caract. | Hábitos de prevención | Complicaciones | | | |
| Variables Independientes | P3_13_1 Lleva a cabo un plan alimenticio | P3_18_3 Disminución de la vista | P3_18_6 Infarto cardiaco | P3_18_7 Infarto cerebral | P3_18_9 Perdida del conocimiento confusión o desmayo |
| Asociación | Positiva | Positiva | Positiva | Positiva | Positiva |
| Odds Ratio | 5.396428 | 1.846405 | 5.637617 | 5.329349 | 15.668369 |
| Probabilidad | 5.3 - 1 | 1.8 - 1 | 5.6 - 1 | 5.3 - 1 | 15.6 - 1 |
| Tasa de probabilidad | 84.12% | 64.28% | 84.84% | 84.12% | 93.97% |

Tabla 23. Relación entre variables con asociación significativa (Coma diabético).

Observaciones, “Coma diabético”.

Con un nivel de significancia de $\alpha = 0.05$, para la variable de complicación “Coma diabético” encontramos una asociación significativa tanto con la variable relacionada con los hábitos del paciente, “Lleva a cabo un plan alimenticio” como con las variables “Disminución de la vista”, “Infarto cardiaco”, “Infarto cerebral” y “Perdida del conocimiento, confusión o desmayo” referentes a otras complicaciones de la enfermedad.

Con base en los resultados de asociación podemos observar que para el paciente diabético el llevar acabo un adecuado plan alimenticio puede reducir en un 84.12% la probabilidad de que presente un coma diabético.

Por otro lado, los eventos de coma diabético en el paciente diabético se relacionan en un 64.28% con la disminución de la vista, un 84.84% con eventos de infarto al corazón, un 84.12% con infarto cerebral y un 93.97% con incidentes de pérdida del conocimiento, confusión o desmayo.

5.7. Diabetes - Intelligent Notifys System (Diabet-INS).

194

Con base en los resultados obtenidos mediante el modelo desarrollado, el cual muestra el impacto que ejercen los hábitos en la salud del paciente y para proporcionar a este y al personal a cargo de la gestión de su salud, una herramienta que les permita llevar un control adecuado de la enfermedad, se desarrolló el sistema “Diebetes - Intelligent Notifys System”, el cual por un lado, motiva al paciente a realizar diferentes actividades que le ayudaran en la prevención de complicaciones y por otro funciona como instrumento para la gestión del expediente clínico completo del paciente, almacenando dicha información para formar una base de datos histórica que servirá para análisis y desarrollo de futuras investigaciones sobre la enfermedad.

35

5.7.1. Herramientas para el desarrollo del sistema.

- **JavaScript:** Es un lenguaje de programación, subconjunto de TypeScript que proporciona contenido dinámico para el desarrollo de aplicaciones y sistemas web. En la actualidad, es el lenguaje de programación más utilizado para el desarrollo de tecnologías nuevas y avanzadas en el mundo de la informática (41).
- **React:** Es una biblioteca de JavaScript que permite construir interfaces de usuario interactivas. Con esta se pueden diseñar vistas simples para cada estado de una aplicación. Con React se puede actualizar y renderizar de forma eficiente componentes cuando los datos cambian. Así también, puede renderizar desde el servidor usando Node y potencializar aplicaciones móviles usando React Native. Además, puede crear componentes

encapsulados que manejan su propio estado convirtiéndolos en interfaces de usuarios complejas (42).

13

- **Next.js:** Es un marco de desarrollo de código abierto elaborado por Node.js que permite funcionalidades de sistemas web basadas en React como renderización estática híbrida y de servidor, compatibilidad con TypeScript, agrupación inteligente y precarga de rutas entre otras (43).
- **Mongoose.js:** Es una biblioteca de programación de JavaScript orientada a objetos que proporciona una conexión entre MongoDB (Sistema de base de datos) y el marco de trabajo de la aplicación web Express. Esta proporciona una solución sencilla, basada en esquemas para el modelado de datos. (44).
- **Express:** Es una infraestructura flexible de aplicaciones web del entorno en tiempo ejecución multiplataforma de código abierto para la capa de servicio Node.js, que proporciona un conjunto sólido de características para las aplicaciones web y móviles (45).
108
- **MongoDB:** Es un sistema de base de datos NoSQL de código abierto orientado a documentos. Este almacena una gran variedad de datos en forma de documentos JSON y BSON (formatos de texto sencillos para el intercambio de datos) lo que permite consultar y analizar datos de forma eficiente. Además, con la herramienta MongoDB Atlas, puede gestionar varias nubes por lo que se pueden distribuir datos entre regiones geográficas y proveedores en la nube (46).
- **Heroku:** Es una plataforma en la nube que se encarga de los procesos de despliegue, configuración, escalado, ajuste y gestión de aplicaciones, basada en contenedores. Esta adopta los lenguajes de programación de las aplicaciones más comunes como Node, Ruby, Java, Scala, PHP entre otros. Cuenta con más de 200 complementos que proporcionan servicios que van desde bases de datos, alertas de tiempo de actividad, servicios de mensajería, copias de seguridad automáticas hasta métricas o entrega de correo electrónico (47).

5.7.2. Descripción del sistema.

El sistema está destinado para gestión tanto de las características y expediente clínico del paciente diabético como de las actividades para la prevención de complicaciones y el control óptimo de la enfermedad. Este cuenta con tres enfoques. El primero va dirigido al paciente pues mediante esta herramienta, puede llevar a cabo el control tanto de las actividades, recomendaciones e indicaciones del médico especialista como del monitoreo de los parámetros fisiológicos, el control en el consumo de medicamentos y el registro de los estudios de laboratorio realizados.

El segundo enfoque va dirigido al personal médico y encargado de la salud de los pacientes pues les permite llevar a cabo la gestión de la salud de diferentes pacientes. Con este, el medico puede agregar al sistema diferentes pacientes para llevar un control del expediente clínico y el progreso en las actividades de prevención de todos sus pacientes. Así también, este puede asignar nuevas actividades de manera personalizada para el control¹⁸⁵ cada paciente en particular. Además, puede acceder al historial de expedientes en la base de datos. Por último, un administrador que estará a cargo de gestionar las credenciales de todos los usuarios y el acceso a la base de datos del sistema.

A continuación, se describen cada una de las vistas del sistema y sus funciones:

- **Registro/Acceso:** Permite el registro para usuario nuevo. Para ello se requiere proporcionar datos de identificación como nombre, apellidos y profesión, un correo electrónico y la creación de una contraseña de acceso.

Registro

Email
victor.moreno3447@alumnos.udg.mx

Nombre
Victor Ernesto

Apellido
Moreno Gonzalez

Ocupacion
Ingeniero

Password

Repite el Password

Imagen 13. Vista “Registro” del sistema (D-INS).

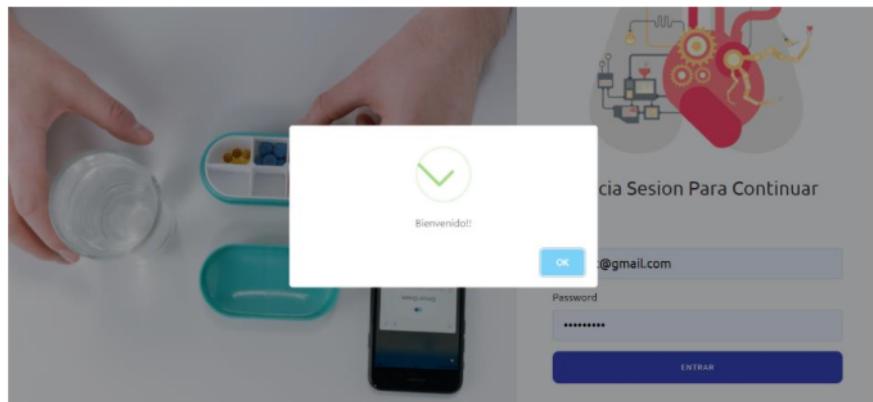


Imagen 14. Vista “Acceso” del sistema (D-INS).

- **Vista Hogar:** Página principal en la cual se muestran las características principales del sistema y el menú de navegación. En esta se encuentra también, al igual que en las demás vistas, el acceso a las notificaciones sobre el progreso en las actividades asignadas a cada paciente.

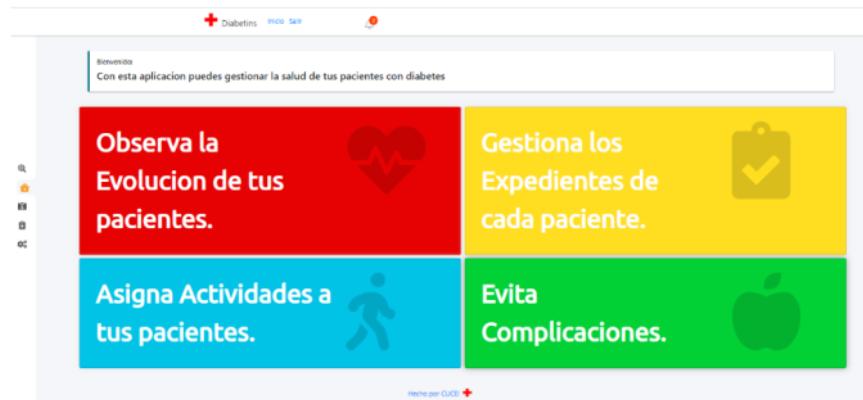


Imagen 15. Vista “Hogar” del sistema (D-INS).

- **Pacientes:** Aquí se pueden registrar los datos de nuevos pacientes. Para el registro de nuevo paciente se deberán llenar datos como, nombre, peso, estatura, edad, sexo, tipo de diabetes, datos de contacto, fecha en que fue diagnosticado y una imagen de identificación.

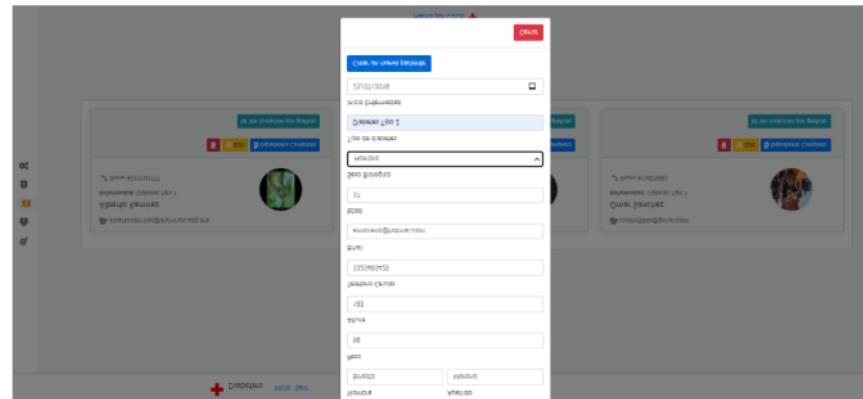


Imagen 15. Vista “Paciente” del sistema (D-INS) (Registro de nuevo paciente).

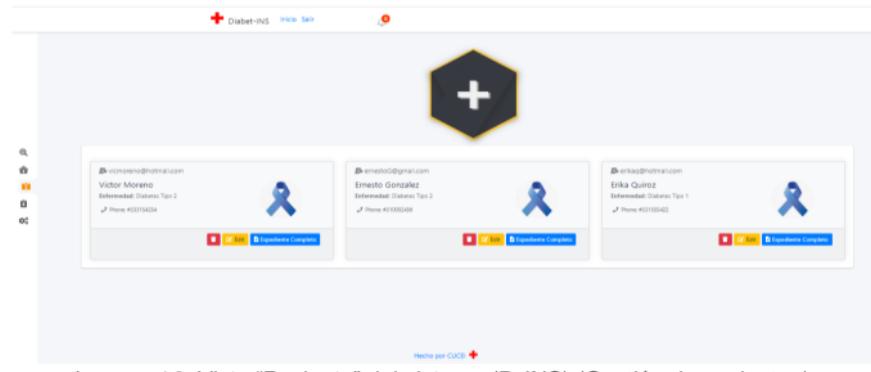


Imagen 16. Vista “Paciente” del sistema (D-INS) (Gestión de pacientes).

Una vez registrado un nuevo paciente se podrá acceder al expediente clínico donde el usuario podrá registrar y gestionar datos completos sobre análisis de laboratorio, citas médicas, régimen alimenticio, plan de ejercicios, actividades de prevención e indicaciones medicas entre otros. Además, desde esta área se puede gestionar el historial de expedientes.

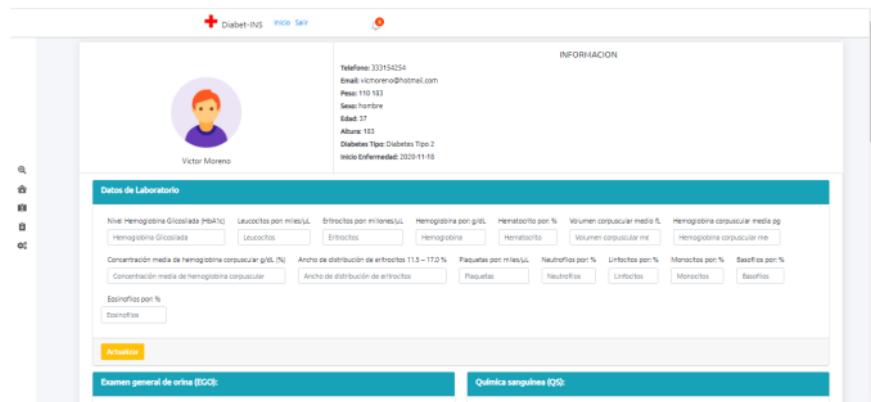


Imagen 17. Vista “Paciente – Expediente” del sistema (D-INS) (Gestión de expediente).



Imagen 18. Vista “Paciente – Historial” del sistema (D-INS) (Gestión de expedientes).

- **Estatus:** En esta página se presenta el estatus del progreso en las actividades e indicaciones médicas. El progreso se muestra mediante una barra que indica el avance de cada paciente. Aquí también se pueden gestionar las actividades destinadas al paciente con la posibilidad de generar nuevas indicaciones y actividades acordes al perfil del paciente en particular.

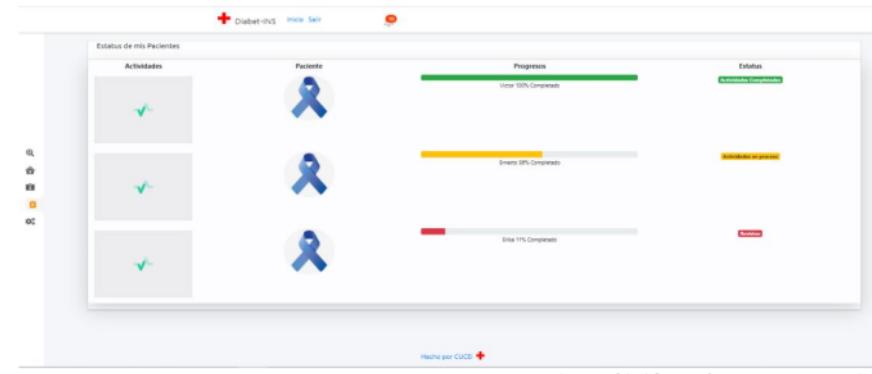


Imagen 19. Vista “Estatus – Progreso” del sistema (D-INS) (Gestión de progreso).

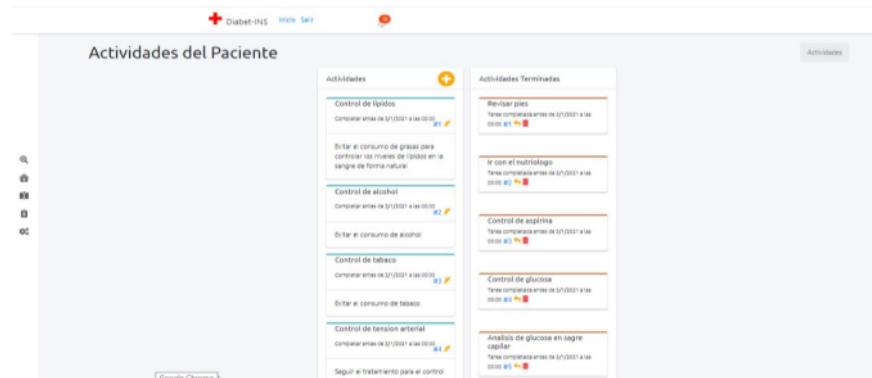


Imagen 20. Vista “Estatus – Actividades” del sistema (D-INS) (Gestión de actividades).

This screenshot shows the 'Actualizar actividad' (Update Activity) dialog box from the D-INS system. It allows users to edit activity details. The 'Nombre de la Actividad' (Activity Name) is set to 'Visitar al nutrólogo'. The 'Descripción de la Actividad' (Activity Description) is 'Visitar al nutrólogo para realizar ajustes en...'. The 'Fecha' (Date) is set to '28/12/2021'. The 'Hora' (Time) is set to '12:00:00 p.m.'. There are buttons for 'Actualizar' (Update), 'Cancelar' (Cancel), and 'Eliminar' (Delete).

Imagen 21. Vista “Estatus – Nueva actividad” del sistema (D-INS) (Gestión de actividades).

- **Perfil:** Muestra los datos y las credenciales del usuario como nombre, correo electrónico y en el caso de personal médico la cantidad de pacientes a cargo. Dependiendo de las credenciales de cada usuario, estos pueden ser: usuario general (Paciente o familiar a cargo), personal de salud (Médico a cargo) o Administrador (Personal a cargo de gestionar las credenciales de los demás usuarios).

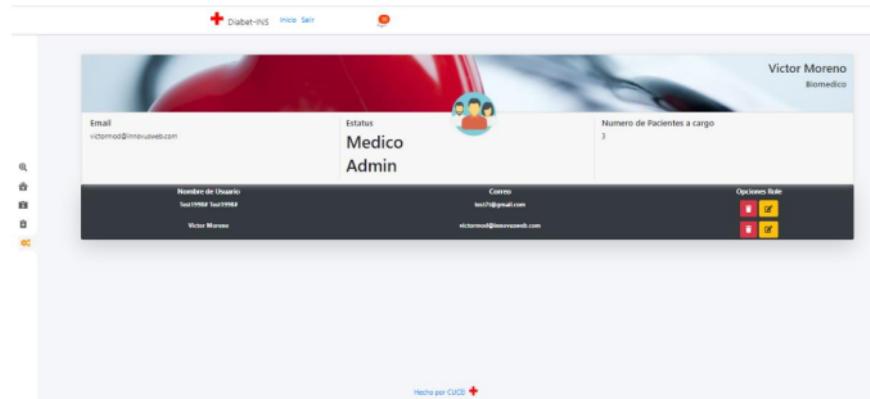


Imagen 22. Vista “Perfil” del sistema (D-INS) (Gestión de usuarios).

- **Buscar:** Para el caso de los usuarios “Administrador” y “Personal médico”, permite buscar en la base de datos a pacientes registrados en el sistema para poder acceder a su expediente e historial clínico completo, con ello, el personal médico puede seleccionar a los pacientes para agregarlos a su control de pacientes a cargo.

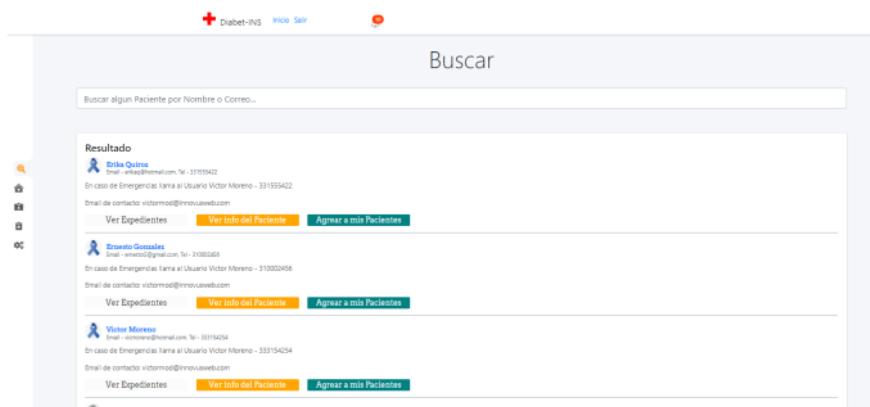


Imagen 23. Vista “Buscar” del sistema (D-INS) (Gestión de pacientes a cargo).

- 198
- **Base de datos del sistema:** En la base de datos el administrador del sistema puede acceder a la información de todos los registros. Entre estos se

encuentra la base de datos de usuarios, pacientes, expedientes, actividades y notificaciones.

```

{
  "_id": "05ec1d1e41394839f818c7e7327c721842",
  "role": "user",
  "notificaciones": [],
  "pacientes": [],
  "name": "TestUser",
  "email": "testuser@gmail.com",
  "password": "TestUser.S0m3g0t7p0w3h4g0n3r3y020_q7n3yngos0ma",
  "lastName": "TestUser",
  "occupation": "TestUser"
}
  
```

Imagen 24. Base de datos del sistema (D-INS) (Gestión de datos).

5.7.3. Actividades y notificaciones de prevención.

Además de la posibilidad de programar actividades acordes a las características de cada paciente, con base en los resultados del modelo desarrollado se diseñaron actividades que estarán preestablecidas en el sistema para todos los pacientes con la finalidad de motivarlos a generar los distintos hábitos que garantizarán la prevención de las complicaciones abordadas en el análisis de investigación.

Estas están diseñadas para indicar al paciente de manera periódica el desarrollo de las diferentes actividades mediante notificaciones que le presentan la razón de probabilidad de evitar cierta complicación al realizar de manera adecuada dicha actividad y le notifica la razón de probabilidad de desarrollar dicha complicación en caso de no realizarla a tiempo.

| Actividad | Respuestas en el sistema | Notificación |
|-------------------------------------------------------------------------------------|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Revisa tus pies y piernas en búsqueda de lesiones o alteraciones en la piel. | Actividad realizada | Al revisar tus pies o piernas reduces un 68.57% la probabilidad de desarrollar ulceras. |
| | Actividad no realizada | Es importante que revises tus pies y piernas. El no realizarlo de manera habitual aumenta 68.57% tus probabilidades de desarrollar ulceras que pueden convertirse en lesiones graves. |
| Seguir el plan alimenticio indicado por el especialista | Actividad realizada | Siguiendo el plan alimenticio indicado disminuyes 59.34% tus probabilidades de pérdida de la vista, un 84.12% tus probabilidades de un evento de coma diabético y 64.28% la probabilidad de un infarto cardíaco. Además, previenes otras complicaciones como insuficiencia renal aguda e infarto cerebral. |
| | Actividad no realizada | Al no seguir el plan alimenticio indicado aumentas 59.34% tus |

| | | |
|----------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| | | probabilidades de perder la vista, 84.12% tus probabilidades de un evento de coma diabético y 64.28% tus posibilidades de sufrir un infarto cardiaco. Además, podrías desarrollar otras complicaciones como insuficiencia renal aguda e infarto cerebral. ¹¹⁹ |
| Llevar a cabo un control adecuado de los niveles de glucosa en la sangre. (Tratamiento farmacológico indicado por el especialista) | Actividad realizada | llevar acabo un adecuado control de los niveles de glucosa en la sangre, con el apoyo del tratamiento indicado, reduce en un 86.11% ³⁴ probabilidades de desarrollar Insuficiencia renal aguda y 79.16 la probabilidad de sufrir un infarto cardiaco. |
| | Actividad no realizada | Al no cumplir con el tratamiento indicado por tu médico de forma adecuada aumentas tus probabilidades de desarrollar insuficiencia renal aguda en un 86.11% y de presentar un infarto cardiaco en un 79.16% |
| Llevar a cabo un control adecuado de los niveles de glucosa en la sangre. (Control mediante análisis habitual de glucosa en sangre capilar) | Actividad realizada | Al revisar ³¹ frecuencia tus niveles de azúcar en la sangre puedes prevenir diferentes complicaciones de la diabetes como: disminución o pérdida total de la vista, insuficiencia renal aguda, ulceraciones en pies o piernas e infarto cardiaco. |
| | Actividad no realizada | Es importante que revises de forma habitual tus niveles de glucosa en sangre capilar pues un control inadecuado puede desencadenar complicaciones como disminución o perdida de la vista, aparición de ulceras o amputación en pies o piernas, insuficiencia renal aguda e infarto cardiaco. |
| Reducir el consumo de grasas para controlar los niveles de lípidos en la sangre | Actividad realizada | Mantener tus niveles de lípidos en la sangre evita que requieras para ello el apoyo de medicamentos, disminuyendo así un 75% tus probabilidades de desarrollar insuficiencia renal aguda. |
| | Actividad no realizada | El que tus niveles de lípidos en la sangre no sean los adecuados puede provocar la necesidad de medicamentos para su control, lo que aumenta 75% tus probabilidades de desarrollar insuficiencia renal aguda. |
| Evitar el consumo de alcohol | Actividad realizada | Al evitar el consumo de bebidas alcohólicas disminuyes 83.33% tus probabilidades de desarrollar insuficiencia renal aguda. Además, previenes otras complicaciones como disminución o perdida de la vista e infarto cardiaco. |
| | Actividad no realizada | El consumir bebidas alcohólicas aumenta 83.33% tus probabilidades de desarrollar insuficiencia renal aguda y contribuye al desarrollo de otras complicaciones como disminución o perdida de la vista e infarto cardiaco. |

| | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Evitar el consumo de tabaco | Actividad realizada | Al evitar el consumo de tabaco o dejar de fumar disminuyes 75% tus posibilidades de desarrollar ulceras en pies o piernas. |
| | Actividad no realizada | El fumar aumenta 75% tus probabilidades de desarrollar ulceras en pies o piernas, lo que podría generarte problemas más graves como es la amputación de la extremidad. |
| Tomar una aspirina 100mg de liberación prolongada. | Actividad realizada | Al tomar una aspirina diaria como hábito de prevención reduce las probabilidades de sufrir un infarto al corazón en un 65.51% |
| | Actividad no realizada | Al no tomar una aspirina diaria podrías aumentar tus probabilidades de sufrir un infarto cardiaco en un 65.51% |
| Realizar un monitoreo de los niveles de tensión arterial / Seguir el tratamiento para el control de los niveles de tensión arterial (Para pacientes hipertensos) 37 | Actividad realizada | El cumplir de manera adecuada el tratamiento para controlar tus niveles de tensión arterial te ayuda a prevenir un 69.69% la probabilidad de sufrir un infarto cerebral. |
| | Actividad no realizada | El no seguir el tratamiento para el control de tus niveles de tensión arterial de forma adecuada aumenta 69.69% tus probabilidades de sufrir un infarto cerebral. |

Tabla 24. Actividades y notificaciones del sistema para la prevención de complicaciones.

Estas notificaciones además de figurar en el sistema se envían al correo electrónico tanto del personal médico encargado de la gestión de la salud del paciente como del usuario o familiar a cargo. Con ello, tanto el médico como la persona a cargo pueden permanecer al tanto del comportamiento de su paciente, permitiéndoles así actuar a tiempo para prevenir complicaciones derivadas de malos hábitos.

CONCLUSIONES.

De acuerdo con la evidencia obtenida mediante el análisis de los datos sobre hábitos y características de la población con diabetes en México y gracias al modelo propuesto logramos⁸⁵ determinar cómo es que influyen los hábitos del paciente y el impacto de estos en la prevención de complicaciones derivadas de la enfermedad. Pudimos observar una relación bastante significativa entre estos dos aspectos con lo que logramos concluir de manera específica cuál es la importancia de cada una de las rutinas de cuidado y control de la enfermedad y la probabilidad de prevención al seguir las indicaciones de forma adecuada.

Con base en los resultados¹¹⁶ asociación entre estas particularidades en el paciente diabético, concluimos que uno de los hábitos más importantes para el control de la enfermedad es llevar a cabo un plan alimenticio adecuado, pues con ello puede reducir sus probabilidades de sufrir un infarto cardiaco en un 64.28%, prevenir un coma diabético en un 84.12% y prevenir el deterioro de la capacidad visual en un 59.34%.

Por otro lado, sabemos que el pie diabético es uno de los principales problemas que se puede presentar en esta clase de pacientes. Mediante los resultados de nuestro modelo logramos determinar que el hábito de revisar sus pies o piernas puede prevenirla en un 68.57% la probabilidad de desarrollar ulceras, las cuales se relacionan en un 91.80% con la amputación de la extremidad. Además, el consumo de tabaco en estos pacientes aumenta las probabilidades de desarrollar estas complicaciones en un 75.00%.

Así también, logramos determinar cómo influyen en estos pacientes otros hábitos nocivos como el consumo de alcohol, azúcares o grasas. Por una parte, el consumo de alcohol aumenta 83.33% las probabilidades de que el paciente presente insuficiencia renal aguda. De la misma forma, el paciente previene¹⁴⁴ en un 86.11% la probabilidad de presentar este problema mediante un control adecuado de sus niveles de glucosa en la sangre, siguiendo de forma apropiada el tratamiento farmacológico indicado por el médico lo cual también previene¹⁹² en un 79.16% la probabilidad de sufrir un infarto cardiaco. Así también, el que los niveles de lípidos en la sangre del paciente no sean los adecuados puede provocar que requiera de tratamiento mediante medicamentos que pueden aumentar en un 75.00% las probabilidades de que el paciente desarrolle insuficiencia renal aguda.

Además, es común que el paciente diabético padezca de hipertensión por lo que para prevenir otra¹⁵³ complicaciones resulta necesario llevar un tratamiento farmacológico para el control de los niveles de tensión arterial. En nuestro estudio logramos determinar para estos pacientes que el llevar a cabo de manera adecuada su tratamiento puede disminuir un 69.69% la probabilidad de sufrir un infarto cerebral.

De igual forma, la diabetes aumenta el riesgo de anomalías a nivel arterial. A través de nuestro estudio logramos determinar que el consumo diario de ácido

acetilsalicílico (aspirina) como terapia en el paciente diabético, puede prevenir en un 65.51% la probabilidad de que llegue a sufrir un infarto cardiaco.

Derivado de lo anterior, podemos determinar que nuestro sistema “Diabet-INS” ofrece, por un lado, para el paciente, una herramienta para llevar un control adecuado de la enfermedad que le incentiva a corregir los hábitos nocivos y a realizar actividades que le ayudan a prevenir complicaciones. Con este, el paciente tiene la posibilidad de sobrellevar su padecimiento de forma saludable. Por otro lado, funciona como una herramienta con la que los médicos o personal responsable del cuidado de la salud del paciente pueden llevar una mejor gest¹⁶⁷ tanto del expediente clínico como de los factores físicos y fisiológicos en los que se desarrolla la enfermedad de su paciente. Finalmente, debido a que todos los datos recogidos en el sistema se irán almacenando, formando una base de datos históricos, nuestro sistema puede proporcionar información importante que podría ser utilizada para futuras investigaciones sobre la diabetes.

Así pues, las diferentes técnicas de aprendizaje automático en conjunto con el análisis y la minería de datos forman una herramienta poderosa tanto para la prevención de complicaciones y detección temprana de enfermedades como para un sinfín de soluciones para diferentes áreas dedicadas al cuidado de la salud.

REFERENCIAS

1. **OMS.** Organizacion Mundial de la Salud (OMS). [En línea] 17 de 11 de 2020. www.who.int/es/news-room/fact-sheets/detail/diabetes.
2. **OPS.** Organización Panamericana de la Salud. [En línea] 29 de 09 de 2021. <https://www.paho.org/es/temas/diabetes>.
3. **Denis Cedeno Moreno, Miguel Vargas Lombardo.** *Applications of machine learning with supervised classification algorithms: In the context of health.* Panama : International Ingineering Sciences and Technology Conference (IESTEC), 30 de December de 2019, IEEE Access, Vol. 7.
4. **Microsoft.** Microsoft Docs. *Documentacion de Analysis Services.* [En línea] 09 de 01 de 2019. <https://docs.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions>.
5. **Wullianallur Raghupathi; Viju Raghupathi.** *Big Data Analytics in Healthcare: Promise and Potential. Health Information Science and Systems.* [En línea] 3 de February de 2014.
6. **Larose, Daniel T y Larose, Chantal D.** *Discovering Knowledge in Data : An Introduction to Data Mining.* s.l. : John Wiley & Sons, Incorporated, 2014.
7. **Márquez, María Pérez.** *Minería de datos a través de ejemplos.* Madrid : Alfaomega, 2014.
8. **Myatt Glenn J., Wayne P. Johnson.** *Making Sense of Data I : A Practical Guide to Exploratory Data Analysis and Data Mining.* s.l. : John Wiley & Sons, Incorporated. ProQuest Ebook Central, 2014.
9. **Charu C. Aggarwal, and Chandan K. Reddy.** *Data Clustering : Algorithms and Applications.* s.l. : CRC Press. ProQuest Ebook Central, 2013.
10. **Sánchez, Juan Jesús Cañas Escamilla. José R. Galo.** RED Descartes. *Estadística y probabilidad de 3º de ESO.* [En línea] 30 de 04 de 2021. https://proyectodescartes.org/iCartesLibri/materiales_didacticos/IntroduccionEstadisticaProbabilidad/3ESO/6MedidasDispersion.html.
11. **Stephens, Rod.** *Beginning Database Design Solutions.* s.l. : John Wiley & Sons, Incorporated. ProQuest Ebook Central, 2008.
12. **Murphy, Kevin P.** *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series).* London England. Cambridge Massachusetts : The MIT Press, 24 Agosto 2012.
13. **Lee, Wei-Meng.** *Python Machine Learning.* s.l. : John Wiley & Sons, Incorporated, 2019.
14. **Jesus Garcia, Jose M Molina, Antonio Berlanga, Miguel A Patricio, Alvaro L Bustamante, Washington R PADILLA.** *Ciencia de Datos. Tecnicas Analiticas y Aprendizaje Estadistico.* Ciudad de Mexico : Alfaomega, 2018.
15. **David J. Domenichini MD, Fred F. Ferri MD.** *Ferri's Clinical Advisor 2022.* s.l. : ELSEVIER, 2021.

16. **Melmed, Shlomo, y otros.** *Williams Textbook of Endocrinology, 14th Edition*. Los Angeles, CA, USA : ELSEVIER, 2020. [63]
17. **McPherson, Richard A y Pincus, Matthew R.** *Henry's clinical diagnosis and management by laboratory methods*. St. Louis, Mo. : ELSEVIER, 2017. [2]
18. *Complicaciones de la diabetes mellitus*. **Bravo, José Javier Mediavilla**. 2001, ScienceDirect (SEMERGEN), págs. 132 - 145. [134]
19. **(CIAD), Centro de Investigación en Alimentación y Desarrollo**. Gobierno de Mexico . [En línea] Oficina de Prensa y Colaboradores, 14 de 11 de 2020. [Citado el: 12 de 11 de 2021.] <https://www.ciad.mx/notas/item/2450-la-pandemia-de-diabetes-en-mexico>. [4]
20. **Instituto Nacional de Estadística, Geografía e Informática**. INEGI. [En línea] Sala de Prensa, 12 de 11 de 2021. [Citado el: 21 de 11 de 2021.] <https://www.inegi.org.mx/app/saladeprensa/noticia.html?id=6923>. [154]
21. **Instituto Nacional de Estadística y Geografía (INEGI), Instituto Nacional de Salud Pública (INSP)**. Instituto Nacional de Estadística y Geografía (INEGI). [En línea] 15 de 03 de 2021. https://www.inegi.org.mx/contenidos/programas/ensanut/2018/doc/ensanut_2018_presentacion_resultados.pdf. [14]
22. **5G-Smart Diabetes: Toward Personalized diabetes Diagnosis With Healthcare Big Data Clouds**. **Min Chen, Jun Yang, Jiehan Zhou; Yixue Hao; Jing Zhang; Chan-Hyun Youn**. 13 de April de 2018, IEEE Access, Vol. 56, págs. 16-23. [126]
23. **Classification Techniques for Disease detection Using Big-Data**. **Jaimin Shah; Raj Patel**. 13 de December de 2019, IEEE Access. [92]
24. **Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data**. **Bhargavi Tragadda; Supriya Kattula; Geetha Guthikonda**. 3, 18 de May de 2018, IEEE Access, Vols. International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). [66]
25. **Diabetic Data Analysis in Big Data With Predictive Method**. **S. Thanga Prasath; S. Sangavi; A. Deepa; F. Sairabani; R. Ragasuda**. 18 de February de 2017, IEEE Access, Vol. IEEE International Conference on Algorithms Methodology Models and Applications in Emerging Technologies (ICAMMAET). [150]
26. **Predicting the Risk of Diabetes in Big Data Electronic Health Records by Using Scalable Random Forest Classification Algorithm**. **Sreekanth Rallapalli; Sreekanth Rallapalli**. 19 de October de 2017, IEEE Access, Vol. International Conference on Advances in Computing and Communication Engineering (ICACCE). [60]
27. **Diabetes prediction using machine learning algorithms**. **Aishwarya Mujumdar; Dr Vaidehi V. 165**, s.l. : ELSEVIER, 2019, ScienceDirect, Vol. International Conference on Recent Trends in Advanced Computing (ICRTAC), págs. 292-299. [18]
28. **Apache Software Fundation**. Apache.org. *Hadoop*. [En línea] 22 de August de 2019. Http://Hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#purpose. [110]

- 65
29. *Real-Time Machine Learning for Early Detection of Heart disease Using Big Data Approach.* **Abderrahmane Ed-Daoudy ; Khalil Maalmi.** 3 de April de 2019, IEEE Access, Vols. International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS).
30. **Python software fundation.** Python.org. Python. [En línea] 2001-2020. <https://www.python.org/about/>.
31. **The Apache Software Fundation.** Apache Spark. spark.apache.org. [En línea] 2018. <https://spark.apache.org>.
- 50
32. **Saha, Sumit.** Towards Data Science. *A Comprehensive Guide to Convolutional Neural Networks - The ELI5 way.* [En línea] 15 de December de 2018. www.towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.
- 101
33. *Understanding and Formulation of Various Kernel Techniques for Support Vector Machines.* **Prayashi Bohra; Hemant Palivela.** 2015, 21 de March de 2016, IEEE Access, Vol. International Conference on Computational Intelligence and Computing Research (ICCIC).
- 22
34. **Rocio Erandi Martinez; Nicandro Cruz Ramirez; Hector Gabriel Acosta Mesa; Ivonne Rebatta Suarez; Maria del Carmen Gogesacoechea Trejo; Patricia Pavon Leon; Sobeida L. Blanquez Morales.** Decision trees as tool in the medical diagnosis. *Articulo Original.* 18 de September de 2009.
- 152
35. *K-Nearest Neighbour Classifiers.* **Padraig Cunningham; Sarah Jane Delany.** 27 de March de 2007, ResearchGate, Vols. Technical Report UCD-CSI-2007-4.
36. **Desarda, Akash.** Towards Data Science. *Understanding AdaBoost.* [En línea] 17 de January de 2019. <http://www.towardsdatascience.com/understanding-adaboost-2f94f22d5bfe>.
37. **Balakrishnama, S. y Ganapathiraju, A.** *Linear Discriminant Analysis - A Brief Tutorial.* Mississippi State : Institute for Signal and Information Processing.
- 91
38. *Logistic Regression Model Optimization and Case Analysis.* **Xiaonan Zou; Youn Hu; Zhewen Tian; Kaiyuan Shen.** 19 de October de 2019, IEEE Access, Vol. International Conference on Computer Science and Network Technology.
- 61
39. *Comparision of Naive Bayes and SVM Algorithm Based on Sentiment Analysis.* **Abdul Mohaimin Rahat; Abdul Kahir; Abu Kaisar Mohammad Masum.** 23 de November de 2019, IEEE Access, Vol. International Conference System Modeling and Advancement in research Trends (SMART).
- 40
40. *Extended K-Means Algorithm.* **Faliu Yi; Inkyu Moon.** 24 de October de 2013, IEEE Access, Vols. International Conference on Intelligent Human-Machine Systems and Cybernetics.
41. **McPeak, Paul Wilton and Jeremy.** *Beginning JavaScript.* s.l. : John Wiley & Sons, Incorporated, 2009. 4ta edicion.
42. **Inc., Facebook.** Reactjs.org. [En línea] Facebook Open Souce, 2010. <https://es.reactjs.org/docs/getting-started.html>.
43. **Inc., Vercel.** Nextjs.org. [En línea] Vercel Inc., 10 de 2016. <https://nextjs.org/docs/getting-started>.

44. **Collective, Mongoose Open.** Mongoose.com. [En línea] Open Collective. <https://mongoosejs.com/docs/guide.html>.
45. **Foundation, OpenJS.** Expressjs.com. [En línea] OpenJS Foundation, 2017. <https://expressjs.com/en/resources/glossary.html>.
46. **Dwight Merriman, Eliot Horowitz y Kevin Ryan.** mongodb.com. *MongoDB Inc.* [En línea] 2007. <https://www.mongodb.com/es/company>.
47. **James Lindenbaum, Adam Wiggins, and Orion Henry.** Heroku.com. [En línea] Salesforce company, 2007. <https://www.heroku.com/about>.
- 54
48. *Disease Prediction by Machine Learning Over Big Data From Health*²¹⁷ *Communities.* Chen, Min; Hao, Yixue; Hwang, Kai; Wang, Lu; Wang, Lin. 26 de April de 2017, *IEEE Access*, Vol. 5, págs. 8869 - 8879.
49. **The Apache Software Fundation.** cassandra.apache.org. *Apache Casandra.* [En línea] 2016. <https://cassandra.apache.org>.

Tesis

ORIGINALITY REPORT



PRIMARY SOURCES

| | | |
|---|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------|
| 1 | qdoc.tips Internet Source | 1% |
| 2 | www.elsevier.es Internet Source | 1% |
| 3 | idoc.pub Internet Source | 1% |
| 4 | Submitted to Universidad Anahuac México Sur Student Paper | <1% |
| 5 | Denis Cedeno-Moreno, Miguel Vargas-Lombardo. "Application of Machine Learning with Supervised Classification Algorithms: In the Context of Health", 2019 7th International Engineering, Sciences and Technology Conference (IESTEC), 2019 Publication | <1% |
| 6 | www.coursehero.com Internet Source | <1% |
| 7 | repositorio.unan.edu.ni Internet Source | <1% |

| | | |
|----|------------------------------------------------------------------|------|
| 8 | sedici.unlp.edu.ar | <1 % |
| 9 | www.clubensayos.com | <1 % |
| 10 | repositorio.ug.edu.ec | <1 % |
| 11 | doku.pub | <1 % |
| 12 | Submitted to CONACYT | <1 % |
| 13 | Submitted to Universitat Politècnica de València | <1 % |
| 14 | www.ijeat.org | <1 % |
| 15 | www.theibfr.com | <1 % |
| 16 | Submitted to University of Wales Swansea | <1 % |
| 17 | Submitted to Universidad Autónoma de Nuevo León | <1 % |
| 18 | www.ijitee.org | <1 % |

| | | |
|----|------------------------------------------------------------------------|------|
| 19 | upcommons.upc.edu Internet Source | <1 % |
| 20 | www.fib.upc.edu Internet Source | <1 % |
| 21 | Submitted to Universidad Carlos III de Madrid Student Paper | <1 % |
| 22 | www.ilae.edu.co Internet Source | <1 % |
| 23 | repositorio.utn.edu.ec Internet Source | <1 % |
| 24 | Submitted to Universidad de Jaén Student Paper | <1 % |
| 25 | documentop.com Internet Source | <1 % |
| 26 | dokumen.pub Internet Source | <1 % |
| 27 | Submitted to Universidad Internacional de la Rioja Student Paper | <1 % |
| 28 | es.reactjs.org Internet Source | <1 % |
| 29 | pesquisa.bvsalud.org Internet Source | <1 % |
| 30 | www.debate.com.mx | |

Internet Source

<1 %

31 www.mdsaud.com

Internet Source

<1 %

32 www.consulta.com.mx

Internet Source

<1 %

33 Submitted to Universidad de Almeria

Student Paper

<1 %

34 livrosdeamor.com.br

Internet Source

<1 %

35 Submitted to BENEMERITA UNIVERSIDAD
AUTONOMA DE PUEBLA BIBLIOTECA

Student Paper

<1 %

36 hdl.handle.net

Internet Source

<1 %

37 mejorconsalud.as.com

Internet Source

<1 %

38 pt.scribd.com

Internet Source

<1 %

39 decon.edu.uy

Internet Source

<1 %

40 kidshealth.org

Internet Source

<1 %

41 www.pinterest.com

Internet Source

<1 %

42 diabetestalk.net <1 %
Internet Source

43 moam.info <1 %
Internet Source

44 www.slideshare.net <1 %
Internet Source

45 Submitted to Universidad de Málaga - Tii <1 %
Student Paper

46 repositorio.uncp.edu.pe <1 %
Internet Source

47 es.wikihow.com <1 %
Internet Source

48 terra.drtango.com <1 %
Internet Source

49 www.eluniverso.com <1 %
Internet Source

50 bdm.unb.br <1 %
Internet Source

51 docplayer.es <1 %
Internet Source

52 Irma Araceli Aburto López. "Principales problemas de Salud Pública en México", <1 %

Universidad Nacional Autonoma de Mexico, 2018

Publication

| | | |
|----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| 53 | Submitted to Universidad Estatal a Distancia Student Paper | <1 % |
| 54 | Submitted to University of Warwick Student Paper | <1 % |
| 55 | nutricioni.com Internet Source | <1 % |
| 56 | clasesantoniolc.files.wordpress.com Internet Source | <1 % |
| 57 | coggle.it Internet Source | <1 % |
| 58 | repository.ucatolica.edu.co Internet Source | <1 % |
| 59 | revista.condusef.gob.mx Internet Source | <1 % |
| 60 | www.scilit.net Internet Source | <1 % |
| 61 | Abdul Mohaimin Rahat, Abdul Kahir, Abu Kaisar Mohammad Masum. "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset", 2019 8th International Conference System | <1 % |

Modeling and Advancement in Research Trends (SMART), 2019

Publication

| | | |
|----|---------------------------------------------------|------|
| 62 | Submitted to National University College - Online | <1 % |
| 63 | conitec.gov.br | <1 % |
| 64 | ichi.pro | <1 % |
| 65 | ijesc.org | <1 % |
| 66 | jyx.jyu.fi | <1 % |
| 67 | vsip.info | <1 % |
| 68 | Submitted to Associatie K.U.Leuven | <1 % |
| 69 | Submitted to Universidad de Guadalajara | <1 % |
| 70 | Submitted to Universidad de Lima | <1 % |
| 71 | Submitted to UNIV DE LAS AMERICAS | <1 % |

diabetesenelanciano.blogspot.com

| | | | |
|----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|------|
| 72 | Internet Source | <1 % | |
| 73 | es.slideshare.net | <1 % | |
| 74 | Internet Source | repository.unap.edu.pe | <1 % |
| 75 | www.easp.es | <1 % | |
| 76 | Internet Source | www.eumed.net | <1 % |
| 77 | www.studocu.com | <1 % | |
| 78 | Danilo Mendoza, Nelson Piedra. "TutNorBD: Assistant for teaching and learning process of relational database normalization up to 3NF from a universal table", 2020 XV Conferencia Latinoamericana de Tecnologias de Aprendizaje (LACLO), 2020 | Publication | <1 % |
| 79 | Submitted to International Baccalaureate Ministry of Education of Ecuador | Student Paper | <1 % |
| 80 | Submitted to Unviersidad de Granada | Student Paper | <1 % |
| 81 | nacionmx.com | Internet Source | <1 % |

<1 %

82 vitaminad.mx <1 %
Internet Source

83 www.eird.org <1 %
Internet Source

84 www.icfes.gov.co <1 %
Internet Source

85 www.msc.es <1 %
Internet Source

86 "プログラム", Folia Endocrinologica Japonica,
2012 <1 %
Publication

87 Submitted to Universidad de Alcalá <1 %
Student Paper

88 gaceta.diputados.gob.mx <1 %
Internet Source

89 openaccess.uoc.edu <1 %
Internet Source

90 www.diabetes.org <1 %
Internet Source

91 "Front Matter", 2019 IEEE 7th International
Conference on Computer Science and
Network Technology (ICCSNT), 2019 <1 %
Publication

- 92 Bhargavi Chatragadda, Supriya Kattula, Geetha Guthikonda. "Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data", 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2018 <1 %
Publication
-
- 93 Submitted to Universidad Católica de Santa María <1 %
Student Paper
-
- 94 Submitted to Universidad Pontificia de Salamanca <1 %
Student Paper
-
- 95 arxiv.org <1 %
Internet Source
-
- 96 es.wikipedia.org <1 %
Internet Source
-
- 97 habilidandopormery.blogspot.com <1 %
Internet Source
-
- 98 infotiti.com <1 %
Internet Source
-
- 99 www.elreportero.net <1 %
Internet Source
-
- 100 www.piediabeticoaped.com <1 %
Internet Source

- 101 "Index", 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2015 <1 %
Publication
-
- 102 Submitted to Consorcio CIXUG <1 %
Student Paper
-
- 103 Martín Adrián Bolívar-Rodríguez, Marcel Antonio Cazarez-Aguilar, Eduardo Esaú Luna-Madrid, Fred Morgan-Ortiz. "Infected jejunal mesenteric pseudocyst: A case report", Cirugía y Cirujanos (English Edition), 2015 <1 %
Publication
-
- 104 Submitted to Universidad de San Martín de Porres <1 %
Student Paper
-
- 105 www.es.kayak.com <1 %
Internet Source
-
- 106 www.metas.org <1 %
Internet Source
-
- 107 www.nobleprog.co.cr <1 %
Internet Source
-
- 108 www.taringa.net <1 %
Internet Source
-
- 109 Submitted to Fundación Universitaria CEIPA <1 %
Student Paper

| | | |
|-----|------------------------------------------------------------------------------|------|
| 110 | Submitted to Southwestern Oklahoma State University Student Paper | <1 % |
| 111 | Submitted to Universidad Catolica San Antonio de Murcia Student Paper | <1 % |
| 112 | Submitted to Universidad Internacional Isabel I de Castilla Student Paper | <1 % |
| 113 | elinversionista.tv Internet Source | <1 % |
| 114 | reddeacceso.org Internet Source | <1 % |
| 115 | tech.gobetech.com Internet Source | <1 % |
| 116 | www.armeria-alvarez.es Internet Source | <1 % |
| 117 | documents.mx Internet Source | <1 % |
| 118 | medlineplus.gov Internet Source | <1 % |
| 119 | patents.google.com Internet Source | <1 % |
| 120 | vbook.pub Internet Source | <1 % |

- 121 www.anahuac.mx <1 %
Internet Source
- 122 www.medicalprobeauty.com <1 %
Internet Source
- 123 www.monografias.com <1 %
Internet Source
- 124 www.terra.es <1 %
Internet Source
- 125 FIDEL SALAS VICENTE. "Investigación y modelización de la adherencia, el desgaste y la fenomenología de daño asociada a la rodadura en contactos rueda-carril de aceros al carbono y bainíticos.", Universitat Politecnica de Valencia, 2015 <1 %
Publication
- 126 Jaimin Shah, Raj Patel. "Classification techniques for Disease detection using Big-data", 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2019 <1 %
Publication
- 127 Silvia Marzal Romeu. "Concepción e integración de arquitecturas y protocolos de comunicación dentro de sistemas de supervisión y control de microrredes" <1 %

inteligentes", Universitat Politecnica de Valencia, 2019

Publication

| | | |
|-----|---------------------------------|------|
| 128 | aprendeia.com | <1 % |
| 129 | cideprospectiva.blogspot.com | <1 % |
| 130 | citic.ucr.ac.cr | <1 % |
| 131 | guelcom.net | <1 % |
| 132 | lookformedical.com | <1 % |
| 133 | prezi.com | <1 % |
| 134 | proposalcentral.com | <1 % |
| 135 | searchdatacenter.techtarget.com | <1 % |
| 136 | umc.minedu.gob.pe | <1 % |
| 137 | worldcat.org | <1 % |
| 138 | www.39ymas.com | <1 % |

-
- 139 www.anodis.com <1 %
Internet Source
- 140 www.mdpi.com <1 %
Internet Source
- 141 www.nutriologo.net <1 %
Internet Source
- 142 www.powershow.com <1 %
Internet Source
- 143 www.tandfonline.com <1 %
Internet Source
- 144 wwwils.nlm.nih.gov <1 %
Internet Source
- 145 "Inventarios Forestales Nacionales de América Latina y el Caribe", Food and Agriculture Organization of the United Nations (FAO), 2021 <1 %
Publication
- 146 1library.co <1 %
Internet Source
- 147 Submitted to Colegio Internacional SEK Quito <1 %
Student Paper
- 148 El Colegio De San Luis. "Texto Completo del No.13 - Revista de El Colegio de San Luis Nueva Época", Revista de El Colegio de San Luis, 2017 <1 %

- 149 Folia Endocrinologica Japonica, 2015 <1 %
Publication
- 150 Jayasri N.P., R. Aruna. "Big data analytics in health care by data mining and classification techniques", ICT Express, 2021 <1 %
Publication
- 151 Pablo Cigliuti. "Proceso de Identificación de Comportamiento de Comunidades Educativas Basado en Resultados Académicos", Revista Latinoamericana de Ingenieria de Software, 2014 <1 %
Publication
- 152 citeseerx.ist.psu.edu <1 %
Internet Source
- 153 diariomedico.recoletos.es <1 %
Internet Source
- 154 el-blog-del-narco.blogspot.com <1 %
Internet Source
- 155 eprints.ucm.es <1 %
Internet Source
- 156 expansion.mx <1 %
Internet Source
- 157 iesalqueriasmatematicas.blogspot.com <1 %
Internet Source

| | | |
|-----------------|---------------------------|------|
| 158 | jmcoach-mx.blogspot.com | <1 % |
| Internet Source | | |
| 159 | omasarai1724.blogspot.com | <1 % |
| Internet Source | | |
| 160 | pecaqo.sejose.com | <1 % |
| Internet Source | | |
| 161 | repositorio.comillas.edu | <1 % |
| Internet Source | | |
| 162 | revistamedica.imss.gob.mx | <1 % |
| Internet Source | | |
| 163 | sefh.interguias.com | <1 % |
| Internet Source | | |
| 164 | uvadoc.uva.es | <1 % |
| Internet Source | | |
| 165 | wapa.pe | <1 % |
| Internet Source | | |
| 166 | www.bibliopsquis.com | <1 % |
| Internet Source | | |
| 167 | www.bioeticaparatodos.com | <1 % |
| Internet Source | | |
| 168 | www.cenetec-difusion.com | <1 % |
| Internet Source | | |
| 169 | www.dpw.state.pa.us | <1 % |
| Internet Source | | |

| | | |
|-----------------|--------------------------------------------------------------------------------|------|
| 170 | www.elcolombiano.net | <1 % |
| Internet Source | | |
| 171 | www.estad-admin.netfirms.com | <1 % |
| Internet Source | | |
| 172 | www.fao.org | <1 % |
| Internet Source | | |
| 173 | www.finanzas.df.gob.mx | <1 % |
| Internet Source | | |
| 174 | www.gacetamedica.com | <1 % |
| Internet Source | | |
| 175 | www.informatica-juridica.com | <1 % |
| Internet Source | | |
| 176 | www.itc.mx | <1 % |
| Internet Source | | |
| 177 | www.medynet.com | <1 % |
| Internet Source | | |
| 178 | www.researchgate.net | <1 % |
| Internet Source | | |
| 179 | www.revistaespacios.com | <1 % |
| Internet Source | | |
| 180 | www.sciencedirect.com | <1 % |
| Internet Source | | |
| 181 | www.ti-pin.com | <1 % |
| Internet Source | | |

- 182 www2.eckerd.com <1 %
Internet Source
- 183 xdocs.net <1 %
Internet Source
- 184 Diego Basso. "Propuesta de Métricas para Proyectos de Explotación de Información", Revista Latinoamericana de Ingenieria de Software, 2014 <1 %
Publication
- 185 ENRIQUE RUIZ MARTÍNEZ. "Herramienta de gestión integral en innovación en imagen médica", Universitat Politecnica de Valencia, 2017 <1 %
Publication
- 186 Matías Marín Falco. "Estudio de la heterogeneidad regulatoria en cáncer y sus implicaciones en la medicina personalizada", Universitat Politecnica de Valencia, 2021 <1 %
Publication
- 187 Submitted to Pontificia Universidad Católica Madre y Maestra PUCMM <1 %
Student Paper
- 188 cdigital.uv.mx <1 %
Internet Source
- 189 ftp.riken.jp <1 %
Internet Source

| | | |
|-----------------|------------------------------------|------|
| 190 | homeopatasenelalambre.blogspot.com | <1 % |
| Internet Source | | |
| 191 | modulodemateematicas.wixsite.com | <1 % |
| Internet Source | | |
| 192 | nutriologosdejalisco.org.mx | <1 % |
| Internet Source | | |
| 193 | oa.upm.es | <1 % |
| Internet Source | | |
| 194 | red.uao.edu.co | <1 % |
| Internet Source | | |
| 195 | repositorio.ucsg.edu.ec | <1 % |
| Internet Source | | |
| 196 | repository.javeriana.edu.co | <1 % |
| Internet Source | | |
| 197 | revistabyte.es | <1 % |
| Internet Source | | |
| 198 | revistas.unal.edu.co | <1 % |
| Internet Source | | |
| 199 | seminaroumgsabadoxela.home.blog | <1 % |
| Internet Source | | |
| 200 | www.cardperu.edu.pe | <1 % |
| Internet Source | | |
| 201 | www.cdc.gov | <1 % |
| Internet Source | | |

| | | |
|-----------------|----------------------------------------------------------------------------------------------------|------|
| 202 | www.censon.ca | <1 % |
| Internet Source | | |
| 203 | www.cfgbiotech.com | <1 % |
| Internet Source | | |
| 204 | www.imsersomayores.csic.es | <1 % |
| Internet Source | | |
| 205 | www.jove.com | <1 % |
| Internet Source | | |
| 206 | www.karger.com | <1 % |
| Internet Source | | |
| 207 | www.losandes.com.ar | <1 % |
| Internet Source | | |
| 208 | www.medicointernistadrherminiocruz.com | <1 % |
| Internet Source | | |
| 209 | www.noticias-oax.com.mx | <1 % |
| Internet Source | | |
| 210 | www.repository.usac.edu.gt | <1 % |
| Internet Source | | |
| 211 | www.saludymedicinas.com | <1 % |
| Internet Source | | |
| 212 | www.uninorte.edu.co | <1 % |
| Internet Source | | |
| 213 | zaguan.unizar.es | <1 % |
| Internet Source | | |

| | | |
|-----------------|--------------------------------------------------------------|------|
| 214 | www.timetoast.com | <1 % |
| Internet Source | | |
| 215 | Stephen W. Moore. "G", Elsevier BV, 2011 | <1 % |
| Publication | | |
| 216 | archive.org | <1 % |
| Internet Source | | |
| 217 | eprints.lancs.ac.uk | <1 % |
| Internet Source | | |
| 218 | www.gnp.com.mx | <1 % |
| Internet Source | | |

Exclude quotes Off

Exclude bibliography Off

Exclude matches Off