Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
Licenciatura en Matemáticas Aplicadas

Modelos multivariados para la clasificación de Enfermedades Crónicas no Transmisibles, en el Estado de Hidalgo

---

T E S I S

Que para obtener el título de

Licenciada en Matemáticas Aplicadas

Presenta

Daira Yalín Hernández Cortés

Bajo la dirección de

Dr. Roberto Ávila Pozos

Mineral de la Reforma, Hidalgo.
Enero de 2026

Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
School of Engineering and Basic Sciences

Mineral de la Reforma, Hgo., a 19 de enero de 2026

Número de control: ICBI-D/075/2026
Asunto: Autorización de impresión.

MTRA. OJUKY DEL ROCÍO ISLAS MALDONADO
DIRECTORA DE ADMINISTRACIÓN ESCOLAR DE LA UAEH

Con Título Quinto, Capítulo II, Capítulo V, Artículo 51 Fracción IX del Estatuto General de nuestra Institución, por este medio, le comunico que el Jurado asignado a la egresada de la Licenciatura en Matemáticas Aplicadas **Daira Yalín Hernández Cortés,** quien presenta el trabajo de titulación **"Modelos multivariados para la clasificación de Enfermedades Crónicas no Transmisibles, en el Estado de Hidalgo",** ha decidido, después de revisar fundamento en lo dispuesto en el Título Tercero, Capítulo I, Artículo 18 Fracción IV; dicho trabajo en la reunión de sinodales, **autorizar la impresión del mismo**, una vez realizadas las correcciones acordadas.

A continuación, firman de conformidad los integrantes del Jurado:

Presidente:	Mtra. Margarita Tetlalmatzi Montiel
Secretario:	Dr. Víctor Gómez Bocanegra
Vocal:	Dr. Roberto Ávila Pozos
Suplente:	Dr. Ronald Richard Jiménez Munguía

Sin otro particular por el momento, reciba un cordial saludo.

Atentamente
"Amor, Orden y Progreso"

Mtro. Gabriel Vergara Rodríguez
Director del ICBI

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
ICBI
DIRECCIÓN

GVR/MMM

Ciudad del Conocimiento, Carretera Pachuca-Tulancingo Km. 4.5 Colonia Carboneras, Mineral de la Reforma, Hidalgo, México. C.P. 42184
Teléfono: 771 71 720 00 Ext. 40001
direccion_icbi@uaeh.edu.mx, vergarar@uaeh.edu.mx

Índice general

Agradecimientos iv

Resumen v

Introducción vi

1. Antecedentes **1**
1.1. Diabetes . . . 1
1.2. Hipertensión . . . 2
1.3. Dislipidemia . . . 2
1.4. Obesidad . . . 3
1.5. Síndrome metabólico . . . 3

2. Modelo Estadístico **5**
2.1. Covarianza y Correlación . . . 5
2.2. Análisis discriminante . . . 6
2.2.1. Análisis discriminante lineal . . . 6
2.2.2. Análisis discriminante cuadrático . . . 10
2.3. Regresión logística . . . 10

3. Resultados **13**
3.1. Base de datos y limpieza . . . 13
3.2. Glosario de variables . . . 15
3.3. Análisis descriptivo . . . 19
3.4. Correlación . . . 21
3.5. Supuestos para los análisis . . . 24
3.5.1. Independencia de observaciones . . . 24
3.5.2. Homogeneidad de matrices de covarianza . . . 24
3.5.3. Normalidad multivariante . . . 25
3.6. Regresión logística general sin metabólicas . . . 26
3.7. Regresión logística por bloques para dislipidemia . . . 28
3.8. Regresión logística por bloques para hipertensión . . . 30
3.9. Regresión logística general con variables metabólicas . . . 32
3.10. Análisis Logit con enfermedades individuales . . . 35

4. Conclusiones **40**
4.1. Trabajos futuros . . . 45

5. Apéndice A. Códigos **46**
5.1. Prueba de homogeneidad de matrices de covarianza 46
5.2. Pruebas de normalidad 46
5.3. Código del análisis discriminante lineal 47
5.4. Código del análisis discriminante cuadrático 48
5.5. Código del análisis de regresión logística 49

6. Apéndice B. Visualización del SIC **50**

7. Apéndice C. Imágenes de correlación **54**

Agradecimientos

Quisiera expresar mi más sincera gratitud a todas las personas que hicieron posible la culminación de este trabajo de tesis.

En primer lugar, extiendo mi agradecimiento más profundo a mi director de tesis, el Dr. Roberto Ávila Pozos, por su invaluable guía, paciencia y apoyo constante durante todo este proceso. Sus enseñanzas y consejos fueron fundamentales para el desarrollo de esta investigación, pero en especial agradezco su entusiasmo, que en muchas ocasiones fue lo que me permitió disfrutar de este trabajo.

No puedo dejar de mencionar a mis compañeros de la carrera, en especial a Alejandro Bigvai, por los momentos de colaboración, desahogo y ánimo mutuo que hicieron de este camino una experiencia más llevadera y memorable.

De manera muy especial, dedico este logro a mi familia. Fui afortunada de tener cinco padres, y a cada uno de ellos les debo una parte fundamental de quien soy. A mi papá chino, quien fue mi principal motivación para estudiar la universidad y no desanimarme ante los tropiezos. Él escuchó mis llantos cuando las cosas no salían bien y me ofreció su ayuda incluso cuando yo misma no sabía que la necesitaba; es el mejor papá, quien siempre se emocionaba con cada uno de mis logros y aunque hoy ya no está en este mundo se fue mil veces más orgulloso al saber que había logrado uno de mis sueños más grandes. A mi mamá Magda, por nunca permitir que su hija tuviera el estómago vacío ni faltara el amor. A mi madre, Eva Yalín, sin cuyo apoyo este trabajo simplemente no se hubiera podido realizar. A mi segundo padre, Jorge Arellano, por su amor incondicional y por ser un ejemplo de perseverancia. A mis hermanos Haziel, Samantha y Carolina, que siempre han sido la luz que tiene mi vida, y a mi pareja, por ser mi fuente constante de motivación y alegría.

Finalmente, agradezco a todas aquellas personas que, de una manera u otra, brindaron su ayuda para que este sueño se hiciera realidad. Muchas gracias.

Resumen

Las enfermedades cardiovasculares como la hipertensión arterial y las enfermedades del corazón, las enfermedades metabólicas como la diabetes y la obesidad, las enfermedades respiratorias como el asma y la enfermedad pulmonar obstructiva crónica, las enfermedades neurodegenerativas como el Alzheimer y el Parkinson, las enfermedades autoinmunes como la esclerosis múltiple y la artritis reumatoide y el Cáncer, son conocidas como Enfermedades Crónicas No Transmisibles.

Las principales enfermedades crónicas en México, que constituyen una causa importante de mortalidad y morbilidad, incluyen enfermedades cardiovasculares, diabetes mellitus, diversos tipos de cáncer, y enfermedades respiratorias crónicas.

De acuerdo a la Organización Panamericana de Salud, las enfermedades crónicas no transmisibles matan a 41 millones de personas cada año, lo que equivale al 71 % de las muertes que se producen en el mundo. En la Región de las Américas, son 5.5 millones las muertes por estas enfermedades cada año.

En este trabajo de analizan datos del estado de Hidalgo, de pacientes con enfermedades crónicas, y particularmente se emplean modelos multivariados para la clasificación de estos padecimiento.

Los resultados aquí presentados pueden constituir una guía para el análisis de la información nacional, además de apoyar en el manejo de grandes volúmenes de información de manera eficiente, para clasificaciones oportunas y, en el mejor de los casos, toma de decisiones de salud pública.

Introducción

La creciente pandemia de obesidad, señalada por la Organización Mundial de la Salud (OMS) desde 1990 como un problema de salud pública en constante agravamiento, representa un desafío epidemiológico global. Esta enfermedad, fuertemente vinculada a hábitos alimenticios inadecuados y sedentarismo, actúa como un eje central que se relaciona e incrementa el riesgo de desarrollar otras condiciones crónicas críticas, como la hipertensión arterial, la dislipidemia, la diabetes mellitus tipo 2 y el síndrome metabólico [1] . Las enfermedades crónicas no transmisibles (ECNT) representan una epidemia creciente en México, puesto que causan el 80 % de las muertes totales [2], mientras que de forma globlal, se registran aproximadamente 41 millones de muertes cada año, lo que equivale al 71 % de las muertes en el mundo. Cada año 15 millones de personas entre 30 y 69 años pierden la vida a causa de enfermedades no transmisibles de forma prematura, fenómeno más frecuente en paises en desarrollo. La gran mayoría de las muertes por ECNT son ocasionadas por enfermedades cardiovasculares (17.9 millones cada año), seguidas del cáncer (9.0 millones), luego, enfermedades respiratorias (3.9 millones) y diabetes (1.6 millones) [3].

La relevancia clínica de estas comorbilidades se vió dramáticamente subrayada durante la pandemia de COVID-19, donde diversos estudios evidenciaron marcadas diferencias en la letalidad [4-6], mientras que pacientes sin enfermedades crónicas presentaron una letalidad del 3.8 %, la presencia de una sola condición crónica elevó significativamente este riesgo (Diabetes: 15.8 %; Hipertensión: 15.6 %; Obesidad: 15.0 %). Más alarmante aún fue el impacto de las comorbilidades múltiples, donde combinaciones como Diabetes + Hipertensión alcanzaron una letalidad del 54.1 % y Diabetes + Obesidad el 36.8 % [7]. La obesidad y la diabetes alteran la respuesta inmunitaria, reduciendo la capacidad para combatir infecciones y aumentar el riesgo de hiperglucemia severa durante la infección. El síndrome metabólico genera un estado proinflamatorio y protrombótico que como consecuencia trae falla multiorgánica [8]. Este es un claro ejemplo de cómo, a través del tiempo, el entender estas enfermedades metabólicas es de gran importancia, tanto por el aumento de casos, como de las consecuencias ligadas a otras enfermedades.

En México, la vigilancia y seguimiento de estas enfermedades se realiza mediante el Sistema de Información en Enfermedades Crónicas (SIC) generado por la Secretaría de Salud con la participación de la fundación Carlos Slim, que hasta julio de 2025 lleva el registro de 1,626,291 pacientes con diagnóstico de, al menos, una enfermedad crónica y al menos una consulta, donde el 71 % de los pacientes son mujeres y el 29 % hombres, 1,025,702 son pacientes con diabetes, 1,067,810 con hipertensión, 584,458 con obesidad y 485,792 con dislipidemia [9]. No obstante, la sistematización de registros clínicos enfrenta muchos retos. Por un lado, se requiere capacitación del registro y seguimiento en la plataforma y a nivel nacional, existe la fragmentación en el sistema de salud, donde existen múltiples subsistemas hospitalarios que operan con estándares desconectados, provocando una limitante en la homogeneidad de la información. La modernización reciente del SIC

ha sido de gran ayuda para esta sistematización, ya que hoy en día toda la información está unificada para el país y se puede analizar facilmente por jurisdicción sanitaria o por entidad federativa.

El presente trabajo de investigación se enfoca en el análisis de datos del estado de Hidalgo mediante modelos multivariados, con el objetivo de identificar patrones de clasificación y factores asociados a la presencia de ECNT. Inicialmente, se planteó la aplicación de técnicas de clasificación supervisada, en particular el Análisis Discriminante Lineal (LDA), por su capacidad para maximizar la separación entre grupos basándose en combinaciones lineales de predictores. Sin embargo, tras la verificación rigurosa de los supuestos estadísticos requeridos por este método —específicamente, la homogeneidad de las matrices de covarianza y la normalidad multivariada— se determinó que los datos no cumplían con estas condiciones, lo que invalidaba la aplicabilidad del LDA.

Como alternativa, se consideró el Análisis Discriminante Cuadrático (QDA), el cual relaja el supuesto de igualdad de matrices de covarianza. No obstante, al evaluar el supuesto de normalidad multivariada, se constató que tampoco era satisfecho, principalmente debido a la naturaleza mixta de las variables (continuas y categóricas) y a la presencia de distribuciones sesgadas en los indicadores metabólicos.

Ante la imposibilidad de aplicar métodos discriminantes basados en supuestos paramétricos estrictos, se optó por el uso de modelos de Regresión Logística, técnica que no requiere normalidad multivariada ni homogeneidad de varianzas-covarianzas, y que resulta adecuada para variables respuesta binarias. A lo largo de este trabajo, se desarrollaron diversos modelos logísticos: generales (con y sin variables metabólicas), por bloques temáticos y específicos para enfermedades individuales, con el fin de evaluar su capacidad predictiva, identificar factores de riesgo y protección, y explorar la interdependencia entre las distintas condiciones crónicas.

Los resultados obtenidos aportan evidencia sobre la utilidad de los modelos multivariados para la clasificación de ECNT en un contexto de datos reales y complejos, como los registrados en el SIC, y ofrecen insumos valiosos para la toma de decisiones en salud pública, la gestión clínica de pacientes crónicos y la priorización de intervenciones preventivas en la población hidalguense.

CAPÍTULO 1

Antecedentes

Las enfermedades crónicas no transmisibles (como la diabetes mellitus, dislipidemia, hipertensión, obesidad y síndrome metabólico) constituyen un problema crítico de salud pública en México y, particularmente, en el estado de Hidalgo. Estos padecimientos, asociados a estilos de vida, hábitos alimenticios y antecedentes heredofamiliares, entre otros, representan las principales causas de mortalidad en el país. Según datos preliminares del INEGI (2024), a nivel nacional, durante el primer semestre de 2023, las enfermedades del corazón y la diabetes mellitus fueron las principales causas de muerte, provocando 97,187 y 55,885 defunciones, respectivamente. [10].

1.1. Diabetes

La diabetes se define como una patología crónica caracterizada por niveles elevados de azúcar (glucosa) en el torrente sanguíneo. Esta molécula representa el combustible esencial para el organismo y se obtiene tanto de la ingesta de alimentos como de procesos metabólicos internos. Para que la glucosa pueda ser aprovechada por las células, requiere de la insulina, una hormona segregada por el páncreas que actúa como facilitadora de este transporte.

Cuando existe un diagnóstico de diabetes, este mecanismo se ve alterado: el páncreas produce niveles insuficientes de insulina, o bien, las células no responden adecuadamente a ella (resistencia a la insulina). Como consecuencia, la glucosa se acumula en la sangre en lugar de nutrir los tejidos. A largo plazo, esta condición genera un impacto sistémico que puede derivar en complicaciones graves en órganos diana como el corazón, los riñones, la vista y el sistema nervioso, además de estar vinculada a ciertos procesos oncológicos. Por ello, el control glucémico y la prevención son fundamentales para mitigar estos riesgos de salud [11].

El incremento sostenido en la prevalencia de la diabetes tipo 2, particularmente en lo que respecta a los casos que aún no han sido diagnosticados, ha impulsado el desarrollo de diversas investigaciones enfocadas en comprender su progresión metabólica y epidemiológica [12].

Dentro de este contexto, se ha identificado que el metabolismo de los oligoelementos desempeña un papel determinante en la evolución de la enfermedad. Un estudio relevante analizó la interrelación entre la glucosa basal en ayunas y diversos elementos traza en suero mediante espectrometría de masas de plasma acoplado inductivamente [13].

Mediante el uso de herramientas de estadística multivariante —tales como el análisis de componentes principales (PCA), el análisis de conglomerados y coeficientes de correlación—, los autores reportaron variaciones significativas en ocho elementos específicos (Zn, Cu, Se, Fe, Mn, Cr, Mg y As) al comparar sujetos diabéticos con individuos sanos.

Este análisis permitió simplificar la dimensionalidad de los datos experimentales, logrando clasificar los oligoelementos de los pacientes en tres grupos o clusters bien definidos [13].

1.2. Hipertensión

La hipertensión arterial se define como la persistencia de niveles elevados de presión en el sistema vascular, con umbrales diagnósticos establecidos generalmente a partir de 140/90 mmHg. Esta condición es a menudo asintomática, lo que subraya la importancia de la monitorización clínica periódica para su detección oportuna. Factores como la edad, la carga genética, el sedentarismo y hábitos alimenticios inadecuados (como el exceso de sodio y alcohol) incrementan significativamente el riesgo de padecerla.

De acuerdo con Tagle (2018), el protocolo diagnóstico requiere la confirmación mediante mediciones en al menos dos sesiones distintas; se establece el diagnóstico cuando la presión sistólica iguala o supera los 140 mmHg y la diastólica los 90 mmHg. El abordaje terapéutico estándar combina modificaciones en el estilo de vida con intervenciones farmacológicas para reducir el riesgo cardiovascular.

La hipertensión arterial representa un desafío crítico para la salud pública global, afectando tanto a naciones desarrolladas como en vías de desarrollo por ser uno de los principales factores de riesgo de mortalidad cardiovascular. En el contexto mexicano, los resultados de la Ensanut 2022 revelan una situación preocupante: la prevalencia de hipertensión en adultos es del 47.8 %, de los cuales una proporción considerable desconocía su diagnóstico previo [14].

A nivel internacional, el estudio de Kishore et al. (2016) [15] en zonas rurales de Delhi aporta evidencia sobre los determinantes de esta condición. Mediante un muestreo aleatorio sistemático en 1,005 sujetos, los autores identificaron una prevalencia del 14.1 %, hallando una correlación significativa con la edad (mayores de 35 años), el consumo de alcohol y la hipercolesterolemia. No obstante, al aplicar un análisis multivariante, se demostró que únicamente la edad, el nivel educativo y los niveles de colesterol actuaron como predictores independientes de la enfermedad.

1.3. Dislipidemia

La dislipidemia se caracteriza por la alteración de los niveles plasmáticos de lípidos, manifestándose a través del aumento de colesterol total y triglicéridos, o bien, por una reducción del colesterol de alta densidad (HDL-C). Esta condición es un precursor crítico de la aterosclerosis y puede derivar de factores genéticos o causas secundarias. De acuerdo con la Guía de Práctica Clínica Mexicana [16], su manejo integral debe basarse en un diagnóstico preciso mediante el perfil lipídico, combinando intervenciones en el estilo de vida con terapia farmacológica hipolipemiante.

La relevancia de esta patología radica en su papel como factor de riesgo modificable ante el incremento global de las enfermedades cardiovasculares. Investigaciones recientes, como el estudio transversal realizado por Xi et al. (2020) [17] en el norte de China con más de 65,000 participantes, subrayan la magnitud de este problema. Mediante análisis de regresión logística multivariante, los autores identificaron que la prevalencia de la dislipidemia está estrechamente vinculada a factores demográficos y hábitos de vida. El estudio concluye que es imperativo focalizar los esfuerzos preventivos en grupos específicos,

tales como varones jóvenes (menores de 55 años) y mujeres en etapa posmenopáusica, para mitigar el impacto de esta enfermedad crónica.

1.4. Obesidad

La obesidad se entiende como una patología crónica de naturaleza compleja, determinada por una acumulación de tejido adiposo que compromete la integridad sistémica del individuo. Su presencia no solo eleva la susceptibilidad a padecer diabetes tipo 2 y afecciones cardiovasculares, sino que también impacta negativamente en la salud ósea, la función reproductiva y aumenta la incidencia de diversos procesos oncológicos. Más allá de lo clínico, esta condición deteriora la calidad de vida al interferir con procesos fundamentales como el ciclo del sueño y la movilidad física.

Clínicamente, el diagnóstico se establece mediante el Índice de Masa Corporal (IMC), el cual relaciona el peso con la estatura elevada al cuadrado:

$$IMC = \frac{\text{peso (kg)}}{\text{estatura (m)}^2} \tag{1.1}$$

No obstante, al ser un indicador indirecto, se recomienda complementarlo con mediciones como la circunferencia abdominal para obtener una evaluación más precisa de la adiposidad. Es importante notar que los umbrales de estas categorías se ajustan según el sexo y la etapa del desarrollo, especialmente en poblaciones pediátricas y adolescentes [18].

A partir de la segunda mitad del siglo XX, el IMC y el perfil lipídico se consolidaron como biomarcadores esenciales para predecir el riesgo de hipertensión arterial. Investigaciones recientes han profundizado en la interacción de estos factores; por ejemplo, Tang et al. [19] demostraron, mediante modelos de regresión logística, que existe una sinergia determinante entre el exceso de peso y la dislipidemia. Su análisis reveló que los pacientes que presentan ambas condiciones tienen un riesgo de hipertensión significativamente más alto que aquellos con un solo factor, mientras que un peso bajo parece actuar como un elemento protector.

Por otro lado, se ha cuestionado cuál de estos factores tiene un peso mayor en eventos críticos. Estudios de cohortes en Asia, utilizando herramientas de aprendizaje automático (*machine learning*) y análisis multivariante, sugieren que, aunque la obesidad es un factor relevante, la dislipidemia ejerce una influencia más determinante en el pronóstico tras un infarto cerebral isquémico. Estos hallazgos subrayan la necesidad clínica de priorizar el control de los lípidos en sangre para mejorar los resultados desfavorables en pacientes con antecedentes cerebrovasculares.

1.5. Síndrome metabólico

El síndrome metabólico (SM) se define como un conglomerado de anomalías fisiológicas y metabólicas que representan un desafío crítico para la salud pública contemporánea. Esta condición se caracteriza por la coexistencia de obesidad central, hipertensión arterial, hiperglucemia y un perfil lipídico adverso (dislipidemia), factores que en conjunto elevan exponencialmente la morbimortalidad asociada a patologías crónicas no transmisibles [20].

Desde una perspectiva diagnóstica, el SM se identifica mediante la presencia de obesidad abdominal, niveles elevados de triglicéridos, una reducción del colesterol de lipoproteínas de alta densidad (HDL-C) y alteraciones en la presión arterial y la glucosa basal. El

incremento global en su prevalencia se atribuye fundamentalmente a la transición hacia estilos de vida sedentarios y al consumo excesivo de dietas hipercalóricas. La relevancia de su detección temprana radica en que actúa como un precursor determinante de enfermedades cardiovasculares, diabetes tipo 2, enfermedad renal crónica, esteatosis hepática no alcohólica y diversos procesos oncológicos [21].

A nivel epidemiológico, la incidencia del síndrome metabólico presenta variaciones significativas dependiendo de los criterios diagnósticos aplicados, así como de variables demográficas como la edad, el sexo, el origen étnico y el nivel socioeconómico de la población analizada. No obstante, la evidencia clínica y epidemiológica coincide en que la obesidad constituye, frecuentemente, el evento iniciador en la cascada de estos trastornos metabólicos [8].

CAPÍTULO 2

Modelo Estadístico

Muchos fenómenos son influenciados por múltiples factores de modo que es necesario analizar más de una variable a la vez y evaluar el efecto de varios factores al mismo tiempo, teniendo en cuenta las posibles interacciones entre ellos.

El análisis multivariable es un conjunto de técnicas estadísticas que analizan simultaneamente múltiples variables, mediciones u objetos bajo investigación [22]. Las técnicas más conocidas de análisis multivariable son:

- Regresión Múltiple y correlación múltiple
- Análisis discriminante múltiple
- Análisis de conglomerados
- Componentes principales y análisis factorial
- Análisis multivariable de Varianza y covarianza
- Correlación canónica
- Análisis de correspondencias

2.1. Covarianza y Correlación

Se define como covarianza a la medida estadística que nos dice la relación lineal que existe entre dos variables y se representa como:

$$\sigma_{ij} = Cov(X_i, X_j) = E((X_i - E(X_i))(X_j - E(X_j))) \tag{2.1}$$

donde X_i y X_j son variables aleatorias diferentes y $E(X_k)$ es el valor esperado de la variable numérica X_k [23].

Así la matriz de covarianza para las variables aleatorias $X_1, X_2, \cdots, X_n$ será:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} \tag{2.2}$$

Un valor de covarianza positivo indica que las variables tienden a aumentar o disminuir juntas, mientras que un valor negativo indica que tienden a moverse en direcciones opuestas. Una covarianza cercana a cero sugiere que no hay una relación lineal clara entre las variables [24]. El valor de la covarianza va a depender del valor que tienen las variables numéricas, de modo que dificulta su interpretación.

Por otro lado, la correlación es una es una medida estandarizada de la covarianza que facilita la comparación entre diferentes pares de variables. La correlación ρ_{ij} se define

como:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{2.3}$$

donde $\sigma_i = \sigma_{ii}$ [25].

De forma práctica, la correlación es un número que se encuentra entre -1 y 1, si es 0 no existe una relación lineal entre las variables, si ρ_{ij} es -1 significa que hay una correlación negativa perfecta, es decir, que mientras el valor de una variable aumenta, el valor de la otra disminuye, y que si ρ_{ij} es +1, hay una correlación positiva perfecta, es decir, el valor de una variable aumenta mientras la otra también lo hace [26].

2.2. Análisis discriminante

Una problemática muy común en el análisis de datos es la clasificación de elementos u observaciones en determinadas categorías dependiendo sus características. El problema central es reconocer a las variables que más discriminan a los elementos [27].

El análisis discriminante es una técnica estadística mediante la cual se propone identificar las variables predictoras o explicativas que describen de mejor forma la atribución de una observación a un grupo o categoría específica [28]. A través de este análisis se busca maximizar la separación entre los grupos, identificando aquellas características que mejor discriminen entre ellos, y permitiendo, a su vez, clasificar nuevas observaciones con base en dichas variables, siendo además una técnica de reducción de dimensionalidad.

Es importante enfatizar que el análisis discriminante es útil cuando poseemos como variables dependientes variables tipo nominales mientras que puede trabajar con variables independientes de diversos tipos, es esencial considerar cuidadosamente las implicaciones de esta elección en la interpretación y validez de los resultados.

El análisis discriminante puede clasificarse según el enfoque y las suposiciones sobre los datos, los que nos interesan en este trabajo son:

- Análisis Discriminante Lineal (ADL).
- Análisis Discriminante Cuadrático (ADC).

2.2.1. Análisis discriminante lineal

El ADL fue propuesto por primera vez en 1936 por Ronald Aylmer Fisher cuando publicó un artículo titulado "The Use of Multiple Measurements in Taxonomic Problems"[27]. Aunque su formulación original se limita a dos grupos, este trabajo sentó las bases para generalizar el ADL a más de dos grupos. La idea central es proyectar los datos en una dirección donde la separación entre los grupos sea máxima, mientras se minimiza la dispersión dentro de cada grupo. Esta dirección óptima se define mediante un vector de pesos $\mathbf{w}$, que se obtiene maximizando la razón entre la varianza entre grupos y la varianza dentro de los grupos [29].

El ADL se basa en los siguientes supuestos:

- Normalidad multivariada: Las variables predictoras deben seguir una distribución normal multivariada dentro de cada grupo.
- Homogeneidad de matrices de covarianza: Las matrices de covarianza poblacionales deben ser iguales entre los grupos.
- Independencia de las observaciones: Las observaciones deben ser independientes entre sí.

Estos supuestos son fundamentales para garantizar que el ADL funcione correctamente y proporcione resultados confiables [30].

Análisis discriminante lineal (LDA) para dos grupos

Supongamos que tenemos dos grupos y p variables predictoras. La proyección se obtiene al multiplicar las variables predictoras por un vector de pesos $\mathbf{w}$, que define la dirección en la que los grupos están mejor separados. Fisher propuso una solución explícita para este caso, la cual se conoce como la **función discriminante de Fisher**. Esta función es una combinación lineal de las variables predictoras que maximiza la separación entre los dos grupos [27]. La función discriminante de Fisher para dos grupos se expresa como:

$$D = \mathbf{w}^T \mathbf{X}$$

donde:

- $\mathbf{w}$: Vector de pesos que define la dirección de proyección.
- $\mathbf{X}$: Vector de variables predictoras.

Si $\mathbf{w} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ y $\mathbf{X} = (x_1, x_2, \ldots, x_p)^T$, entonces:

$$D = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

La función discriminante lineal D puede interpretarse como una rotación de los datos en el espacio original, donde la nueva dirección $\mathbf{w}$ maximiza la separación entre los grupos [31].

Fisher demostró que la dirección óptima $\mathbf{w}$ se obtiene maximizando la razón entre la varianza entre clases ($\mathbf{S}_B$) y la varianza dentro de las clases ($\mathbf{S}_W$). Esto se logra resolviendo el siguiente problema de optimización:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

1. **Varianza dentro de las clases ($\mathbf{S}_W$)**: Mide la dispersión de los datos dentro de cada grupo. Se calcula como:

$$\mathbf{S}_W = \sum_{\mathbf{x} \in \text{Clase 1}} (\mathbf{X} - \boldsymbol{\mu}_1)(\mathbf{X} - \boldsymbol{\mu}_1)^T + \sum_{\mathbf{x} \in \text{Clase 2}} (\mathbf{x} - \boldsymbol{\mu}_2)(\mathbf{x} - \boldsymbol{\mu}_2)^T$$

   donde μ_1 es la media aritmética del grupo uno y μ_2 la media aritmética del grupo dos. Esta matriz estima la matriz de covarianzas común $\boldsymbol{\Sigma}$ [24].
2. **Varianza entre clases ($\mathbf{S}_B$)**: Mide la separación entre las medias de los dos grupos. En el caso de dos grupos, $\mathbf{S}_B$ se simplifica a:

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$$

La solución se obtiene resolviendo la ecuación de valores propios:

$$\mathbf{S}_W^{-1} \mathbf{S}_B \mathbf{w} = \lambda \mathbf{w}$$

En el caso de dos grupos, esta ecuación tiene una solución explícita: $\mathbf{w}$ es proporcional a $\mathbf{S}_W^{-1}(\mu_1 - \mu_2)$. Por lo tanto, la función discriminante de Fisher se obtiene al proyectar los datos $\mathbf{X}$ en esta dirección óptima:

$$D = \mathbf{w}^T \mathbf{X} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{S}_W^{-1} \mathbf{X}$$

Análisis discriminante lineal (LDA) para más de dos grupos

Aunque la función discriminante de Fisher fue desarrollada originalmente para dos grupos, el ADL se puede generalizar para más de dos grupos. En este caso, se busca encontrar múltiples direcciones de proyección (en lugar de una sola) que maximicen la separación entre los grupos. Estas direcciones se obtienen resolviendo un problema de valores propios generalizado, similar al caso de dos grupos.

Sin embargo, en lugar de una sola función discriminante, se obtienen varias funciones discriminantes, cada una correspondiente a una dirección óptima de proyección.

El análisis discriminante lineal (ADL) no solo se utiliza para separar grupos, sino también para clasificar nuevas observaciones en uno de los grupos predefinidos. Una vez que se ha encontrado la función discriminante óptima, esta se puede usar para asignar un nuevo dato $\mathbf{X}$ a uno de los grupos existentes [29].

El ADL está relacionado con el teorema de Bayes. Bajo las suposiciones de distribución normal multivariada y matrices de covarianzas iguales, la probabilidad posterior de que un dato $\mathbf{X}$ pertenezca a un grupo k se puede expresar como:

$$P(G = k \mid \mathbf{X}) = \frac{P(G = k)P(\mathbf{X} \mid G = k)}{\sum_{i=1}^{K} P(\mathbf{X} \mid G = i)P(G = i)}$$

donde:

- $P(G = k)$ es la probabilidad a priori de pertenecer al grupo k.
- $P(\mathbf{X} \mid G = k)$ es la densidad de probabilidad de $\mathbf{X}$ dado que pertenece al grupo k [23].

Asumimos que $\mathbf{X}$ sigue una distribución normal para cada grupo k, es decir que:

$$P(\mathbf{X} \mid G = k) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1}(\mathbf{X} - \boldsymbol{\mu}_k)\right)$$

Donde:

- $\boldsymbol{\mu}_k$ es el vector de medias del grupo k.
- Σ_k es la matriz de covarianza del grupo k.
- p es la dimensionalidad de $\mathbf{X}$, o bien, el número de variables predictoras.

Además asumimos que todos los grupos tienen la misma matriz de covarianza $\mathbf{\Sigma}$, entonces:

$$\mathbf{\Sigma}_k = \mathbf{\Sigma} \quad \forall k$$

de este modo:

$$P(\mathbf{X} \mid G = k) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_k)\right)$$

Podemos notar que el denominador será el mismo para todas las clases, por lo que no afectará la clasificación, de modo que solo se trabajará con la probabilidad posterior no normalizada, es decir:

$$P(G = k \mid \mathbf{X}) \propto P(G = k) \cdot P(\mathbf{X} \mid G = k)$$

tomando el logaritmo natural de ambos lados podemos expresar que:

$$\log P(G = k \mid \mathbf{X}) = \log\left(P(G = k) \cdot P(\mathbf{X} \mid G = k)\right)$$

así,

$$\log P(G = k \mid \mathbf{X}) = \log P(G = k) + \log P(\mathbf{X} \mid G = k)$$

ya que asumimos que $P(\mathbf{X} \mid G = k)$ sigue una distribución normal:

$$\log P(\mathbf{X} \mid G = k) = \log\left(\frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}\right) + \log\left(\exp\left(-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_k)\right)\right)$$

simplificando:

$$\log P(\mathbf{X} \mid G = k) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_k)$$

de este modo:

$$\log P(G = k \mid \mathbf{X}) = \log P(G = k) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_k)$$

Los términos que no dependen de k no afectan la decisión de clasificación, por lo que pueden ser eliminados. Por lo tanto, la función discriminante lineal $\delta_k(\mathbf{X})$ se define como:

$$\delta_k(\mathbf{X}) = \log P(G = k) - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}_k).$$

Expandiendo el término cuadrático:

$$\delta_k(\mathbf{X}) = \log P(G = k) - \frac{1}{2}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X} + \boldsymbol{\mu}_k^T\mathbf{\Sigma}^{-1}\mathbf{X} - \frac{1}{2}\boldsymbol{\mu}_k^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k$$

Eliminamos el término $\mathbf{X}^T\Sigma^{-1}\mathbf{X}$ (que no depende de k):

$$\delta_k(\mathbf{X}) = \log P(G = k) + \boldsymbol{\mu}_k^T\mathbf{\Sigma}^{-1}\mathbf{X} - \frac{1}{2}\boldsymbol{\mu}_k^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k$$

Finalmente, la función discriminante lineal para el grupo k es:

$$\delta_k(\mathbf{X}) = \mathbf{w}_k^T\mathbf{X} + b_k$$

Donde:

- $\mathbf{w}_k = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k$ es el vector de pesos.
- $b_k = \log P(G = k) - \frac{1}{2}\boldsymbol{\mu}_k^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k$ es el término de sesgo [30].

Recordemos que la expresión $D = \mathbf{w}^T\mathbf{X}$ es una proyección lineal de $\mathbf{X}$ en la dirección de $\mathbf{w}$. Para el grupo k:

$$D_k = \mathbf{w}_k^T\mathbf{X}$$

Por lo tanto, la función discriminante lineal se puede expresar como:

$$\delta_k(\mathbf{X}) = D_k + b_k$$

Si $\mathbf{w} = (\alpha_1, \alpha_2, \ldots, \alpha_p)^T$ y $\mathbf{X} = (x_1, x_2, \ldots, x_p)^T$, entonces:

$$D = \mathbf{w}^T\mathbf{X} = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_p x_p$$

Para el grupo k, el vector de pesos $\mathbf{w}_k$ se define como:

$$\mathbf{w}_k = \mathbf{\Sigma}^{-1}\boldsymbol{\mu}_k$$

Por lo tanto, la proyección D_k para el grupo k es:

$$D_k = \mathbf{w}_k^T \mathbf{X} = (\mathbf{\Sigma}^{-1} \boldsymbol{\mu}_k)^T \mathbf{X}.$$

Entonces, para clasificar un nuevo punto $\mathbf{X}$, se calcula la función discriminante lineal para cada grupo k:

$$\delta_k(\mathbf{X}) = \mathbf{w}_k^T \mathbf{X} + b_k.$$

Por lo que es importante recalcar que una vez entrenado el modelo, se puede usar la función discriminante para clasificar nuevas observaciones en uno de los grupos [31]. Luego, se asigna $\mathbf{X}$ al grupo k que maximiza $\delta_k(\mathbf{X})$.

2.2.2. Análisis discriminante cuadrático

A diferencia del Análisis Discriminante Lineal (LDA), que asume que las matrices de covarianza de los grupos son iguales, el análisis discriminante cuadrático (QDA) no hace esta suposición, lo que le permite ser más flexible y adaptarse a situaciones donde los grupos tienen varianzas y covarianzas diferentes [29]. El QDA se basa en la función discriminante de Bayes, que asigna una observación al grupo que maximiza la probabilidad posterior de pertenecer a ese grupo. La forma de la función discriminante cuadrática es la siguiente:

$$\delta_k(x) = -\frac{1}{2}\log(|\mathbf{\Sigma}_k|) - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{X} - \boldsymbol{\mu}_k) + \log(\pi_k)$$

donde $\delta_k(x)$ es la función discriminante para el grupo k, $\mathbf{X}$ es el vector de variables predictoras, $\mathbf{\Sigma}_k$ es la matriz de covarianza del grupo k, $\boldsymbol{\mu}_k$ es el vector de medias del grupo k, π_k es la probabilidad a priori de pertenecer al grupo k [28].

2.3. Regresión logística

Otro método muy utilizado es la regresión logística, también llamado análisis logit. Es un método discriminante que se limita a dos grupos y a diferencia del ADL, este es mucho más flexible en cuanto a los supuestos. La regresión logística es similar a la regresión múltiple en muchos aspectos, pero difiere en la forma en que se estiman los coeficientes [32].

La regresión múltiple busca modelar la relación entre una variable dependiente (Y) y varias variables independientes ($X_1, X_2, \ldots, X_n$). El objetivo es predecir el valor de Y en función de las variables independientes. La ecuación de regresión múltiple tiene la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

donde:

- Y: Variable dependiente.
- $X_1, X_2, \ldots, X_n$: Variables independientes.
- β_0: Intercepto (valor de Y cuando todas las X son cero).
- $\beta_1, \beta_2, \ldots, \beta_n$: Coeficientes que indican el efecto de cada variable independiente sobre Y.
- ϵ: Término de error.

Los coeficientes se estiman utilizando el método de mínimos cuadrados. Este método minimiza la suma de los cuadrados de los errores (las diferencias entre los valores observados y los valores predichos por el modelo) [25].

La regresión logística es un modelo estadístico diseñado específicamente para variables de resultado binarias (por ejemplo, 0/1, éxito/fracaso). A diferencia de la regresión múltiple que se usa para variables dependientes continuas, en lugar de predecir directamente el valor de Y, la regresión logística predice la probabilidad de que Y pertenezca a una categoría específica [33].

Deriva su nombre de la transformación logística (función logit) aplicada a la probabilidad de la categoría de interés. Esta transformación permite modelar la relación de forma lineal:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

donde:

- p: Probabilidad de que Y sea 1 (o pertenezca a la categoría de interés).
- $\ln\left(\frac{p}{1-p}\right)$: Logaritmo del odds (logit) [32].

En lugar de mínimos cuadrados, la regresión logística utiliza el método de máxima verosimilitud para estimar los coeficientes. Este método busca maximizar la probabilidad de que los datos observados sean explicados por el modelo [33].

Ya que la regresión logística es más flexible no se requiere normalidad de los errores ni homocedasticidad.

Para interpretar los coeficientes de manera más intuitiva, se utiliza el concepto de *odds ratio*. Al exponenciar el coeficiente β_i de una variable, se obtiene su *odds ratio*:

$$\text{Odds Ratio (OR)} = e^{\beta_i}$$

Este *odds ratio* representa cuántas veces se multiplica el *odds* del evento por cada incremento de una unidad en la variable X_i, manteniendo constantes las demás variables. La relación entre la probabilidad del evento y las variables predictoras se puede expresar directamente como:

$$\frac{\text{Prob}_{\text{evento}}}{\text{Prob}_{\text{no evento}}} = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n}$$

Así, los coeficientes β_i se interpretan a través de su *odds ratio*:

- Si $\beta_i > 0$, entonces $e^{\beta_i} > 1$. Esto indica que un aumento en X_i incrementa el *odds* del evento, por lo que X_i actúa como un **factor de riesgo**.
- Si $\beta_i < 0$, entonces $e^{\beta_i} < 1$. Esto significa que un aumento en X_i disminuye el *odds* del evento, por lo que X_i actúa como un **factor de protección**.
- Si $\beta_i = 0$, entonces $e^{\beta_i} = 1$, lo que indica que X_i no tiene efecto sobre el *odds* del evento [32, 33].

Hay que considerar las medidas de ajuste $-2LL$ (menos dos veces el logaritmo de la verosimilitud), el cual es un valor pequeño que indica un buen ajuste, y **Pseudo** R^2, que mide el ajuste global del modelo y se calcula como:

$$R^2_{\text{logit}} = \frac{-2LL_{\text{nulo}} - (-2LL_{\text{model}})}{-2LL_{\text{nulo}}}$$

donde:

- $-2LL_{\text{nulo}}$: Valor de $-2LL$ para el modelo nulo.
- $-2LL_{\text{model}}$: Valor de $-2LL$ para el modelo con variables independientes [32].

Un valor más alto de pseudo R^2 indica un mejor ajuste, y por último, el contraste Chi-Cuadrado, que evalúa la mejora del modelo al añadir variables. Se compara un modelo nulo (que solo incluye la media de la variable dependiente) con un modelo más complejo. La reducción en $-2LL$ (menos dos veces el logaritmo de la verosimilitud) indica cuánto mejora el modelo [33].

La interpretación de los coeficientes β_i está intrínsecamente ligada al concepto de factores de riesgo y protección. Cuando una variable independiente tiene un coeficiente positivo y estadísticamente significativo, su aumento incrementa el *odds ratio* de que ocurra el evento. En este contexto, esa variable se interpreta como un factor de riesgo para la categoría de interés (por ejemplo, contraer una enfermedad). Por el contrario, si el coeficiente es negativo, un aumento en la variable reduce el *odds ratio*, actuando entonces como un factor de protección [32]. Por ejemplo, en un modelo que predice la probabilidad de tener una enfermedad cardíaca, el coeficiente para la variable "edad" podría ser positivo (factor de riesgo), mientras que el de "horas de ejercicio semanal" podría ser negativo (factor de protección).

Para evaluar la capacidad predictiva de los modelos de clasificación, ya sea LDA, QDA o Regresión Logística, se utilizan las tablas de confusión. Esta es una matriz que contrasta las clasificaciones reales con las predicciones del modelo. En el caso de este trabajo, las tablas de confusión se construyen específicamente con un subconjunto de prueba.

Una tabla de confusión para dos grupos se estructura de la siguiente manera:

		Clasificación Predicha	
		Grupo 1	**Grupo 2**
Clasificación Real	**Grupo 1**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Grupo 2	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Cada celda representa un resultado crucial:

- **Verdaderos Positivos (VP):** Casos que son del Grupo 1 y fueron correctamente clasificados como Grupo 1.
- **Verdaderos Negativos (VN):** Casos que son del Grupo 2 y fueron correctamente clasificados como Grupo 2.
- **Falsos Positivos (FP):** Casos que son del Grupo 2 pero fueron incorrectamente clasificados como Grupo 1 (error Tipo I).
- **Falsos Negativos (FN):** Casos que son del Grupo 1 pero fueron incorrectamente clasificados como Grupo 2 (error Tipo II).

A partir de estas cifras, se calculan métricas clave como la Precisión (VP / (VP+FP)), que mide la fiabilidad de una clasificación positiva, y la Sensibilidad o Tasa de Verdaderos Positivos (VP / (VP+FN)), que mide la capacidad del modelo para detectar los casos positivos reales.

CAPÍTULO 3

Resultados

3.1. Base de datos y limpieza

La base de datos se conformó mediante la integración de los doce conjuntos de datos correspondientes a cada jurisdicción sanitaria, provenientes de la Gerencia del Sistema de Información en Crónicas (SIC) del Estado de Hidalgo. Este sistema constituye una herramienta diseñada para el registro, seguimiento y gestión de pacientes diagnosticados con enfermedades crónicas como diabetes, hipertensión, obesidad, dislipidemia y síndrome metabólico, las cuales requieren atención continua. Su objetivo principal es mejorar la calidad de la atención médica, facilitar el monitoreo de los pacientes y proporcionar datos para la toma de decisiones en políticas de salud.

En el proceso de preprocesamiento, inicialmente se creó una variable numérica denominada "juris"para identificar la jurisdicción de pertenencia de cada registro antes de proceder a la unión de los doce conjuntos de datos. Una vez integrada la base, se generó un identificador único ("id") de tipo numérico a partir de la Clave Única de Registro de Población (CURP), considerando que R no admite identificadores no numéricos para los análisis. Cabe destacar que un mismo paciente puede contar con múltiples registros. Posteriormente, se identificaron errores en los registros, lo que permitió detectar y eliminar un caso sin CURP y cinco registros con datos inconsistentes. Como resultado, se obtuvo un conjunto de datos inicial compuesto por 191 variables y 350,506 registros.

Las variables no métricas se transformaron en variables categóricas numéricas según su naturaleza, mientras que las variables de tipo fecha y caracter se conservaron sin modificaciones. Los valores faltantes en estas variables se codificaron como "NA"para indicar la ausencia de información. A continuación, se realizó un conteo de valores faltantes por variable. Se eliminaron las variables relacionadas con fechas de registro, fechas de diagnóstico y datos personales, conservando únicamente el identificador numérico único previamente creado. Se mantuvieron aquellas variables que no superaran el 20 % de valores faltantes del total de datos.

Dado que la base de datos integrada no incluía de manera explícita variables categóricas para las cinco principales enfermedades crónicas, se crearon cinco nuevas variables: "Diabetes", "HT"(Hipertensión), "Dislipidemia", "Sindrome"(Síndrome Metabólico) y "Obesidad". Estas variables se generaron a partir de los indicadores de control de cada enfermedad ("ControlHT", "ControlDiabetes", "ControlDislipidemia", "ControlSindrome" y "ControlObesidad"), los cuales especificaban si el paciente presentaba la enfermedad controlada ("1"), no controlada ("0") o si no la padecía ("NA").

Con el fin de facilitar el análisis y la visualización, las variables se clasificaron en cuatro categorías: información general, medicamentos, complicaciones, y antecedentes familiares y personales. La organización de las variables se realizó con base en esta categorización.

Para este estudio, se seleccionó un único registro por paciente, optando por aquel que

presentara la mayor completitud de datos. De esta manera, la base de datos final quedó constituida por 104 variables y 33,672 observaciones.

El SIC presenta limitaciones inherentes a los datos registrados, lo que define un rango mínimo y máximo posible para las variables. En este conjunto de datos, se identificaron errores de captura en las variables de peso y talla. Por consiguiente, se excluyeron los registros de pacientes con un peso superior a 150 kg y una talla inferior a 1.3 m. Asimismo, se eliminaron 43 registros correspondientes a menores de 18 años con el propósito de enfocar el análisis en la población adulta.

Se realizó un análisis exhaustivo de valores faltantes para cada variable, identificándose las siguientes frecuencias de NA: sobrepeso (18,296), colesHDL (17,028), colesLDL (16,579), creatinina (13,303), hbA1c1 (10,357), colesTotal (8,443), trigliceridos (8,315), glucemia (3,122), etnia (2,406), cc (318), imc (138), peso (138), sistolica (114), diastolica (115), influenza (4). Las variables restantes presentaron cero valores faltantes. Considerando este patrón de completitud y los objetivos del estudio, se seleccionó el siguiente conjunto final de variables para el análisis:

- "id"
- "sexo"
- "edad"
- "etnia"
- "talla"
- "peso"
- "imc"
- "cc"
- "glucemia"
- "hbA1c1"
- "colesHDL"
- "colesLDL"
- "colesTotal"
- "trigliceridos"
- "creatinina"
- "sobrepeso"
- "Juris"
- "sistolica"
- "diastolica"
- "albuminuria"
- "prediabetico"
- "influenza"
- "AC_alimenCorrecta"
- "AF_activFisica"
- "HA_habAlcoholico"
- "HT_habTabaquico"
- "antecPersSedentarismo"
- "antecPersSobrepeso"
- "antecPersAlcoholismo"
- "antecPersTabaquismo"
- "antecPersVIH"
- "antecPersTuberculosis"
- "antecPersPostmenop"
- "antecPersTeraRemplazo"
- "antecPersEnfCardiovascular"
- "antecPersEnfCerebroVascular"
- "ABUELOShta"
- "PADREShta"
- "TIOShta"
- "HERMANOShta"
- "NINGUNOhta"
- "ABUELOSdm"
- "PADRESdm"
- "TIOSdm"
- "HERMANOSdm"
- "NINGUNOdm"
- "ABUELOSdislip"
- "PADRESdislip"
- "TIOSdislip"
- "HERMANOSdislip"
- "NINGUNOdislip"
- "ABUELOSobes"
- "PADRESobes"
- "TIOSobes"
- "HERMANOSobes"
- "NINGUNOobes"
- "ABUELOSenfCardio"
- "PADRESenfCardio"
- "TIOSenfCardio"
- "HERMANOSenfCardio"
- "NINGUNOenfCardio"
- "ABUELOSenfCerebroVasc"
- "PADRESenfCerebroVasc"
- "TIOSenfCerebroVasc"

- “HERMANOSenfCerebroVasc”
- “NINGUNOenfCerebroVasc”
- “cRetinopatia”
- “cNeuropatia”
- “cPieDiabetico”
- “cEnfermedadRenal”
- “cEnfCardiovasc”
- “cEnfCerebrovasc”
- “cApneadelSueo”
- “mAcidoAcetil”
- “mMetformina”
- “mGlibenclamida”
- “mLinagliptina”
- “mAcarbosa”
- “mInsulinaRap”
- “mInsulinaGlarg”
- “mInsulinaNPH”
- “mInsulinaLispro”
- “mComplejoB”
- “mCaptopril”
- “mEnalapril”
- “mNifedipino”
- “mHidroclorotiaz”
- “mClortalidona”
- “mMetoprolol”
- “mPropranolol”
- “mTelmisartan”
- “mLosartan”
- “mIrbesartan”
- “mAlopurinol”
- “mPravastatina”
- “mAtorvastatina”
- “mBezafibrato”
- “mVerapamil”
- “mFurosemida”

3.2. Glosario de variables

Las tablas adjuntas presentan un glosario de las variables contenidas en la base de datos, excluyendo variables relacionadas con fechas de registro, fechas de diagnóstico, datos personales sensibles y variables categóricas con mas de 2 categorías. La base de datos se ha dividido en dos categorías principales para facilitar su interpretación:

La Tabla 3.1 incluye las variables de tipo métrico (continuas), donde se especifica el nombre original de cada variable, los valores posibles que puede tomar y una descripción detallada de su significado clínico.

La Tabla 3.2 presenta las variables dicotómicas y categóricas. Para todas las variables dicotómicas en esta tabla, se utiliza la codificación estándar donde la presencia de la característica se representa con el valor 1 y la ausencia con el valor 0. Cada variable incluye su nombre original, la abreviatura utilizada y una descripción de su significado clínico según el manual de uso del SIC [9].

Esta división permite una mejor organización y comprensión de los diferentes tipos de variables utilizadas en el análisis.

Tabla 3.1: Variables métricas en la base de datos.

Variable	**Valores posibles**	**Descripción**
id	Numérico Único	Identificación única del paciente a partir del CURP.
edad	≥ 0	Edad del paciente en años.
talla	≥ 0 (metros), con un entero y dos decimales	Estatura del paciente.
peso	≥ 0 (kg), mínimo 20, máximo 750	Peso del paciente.

Continúa en la siguiente página

Variable	Valores posibles	Descripción
imc	≥0	Índice de masa corporal calculado.
cc	≥0 (cm), mínimo 40, máximo 250	Medida de la circunferencia de la cintura.
glucemia	≥0 (mg/dL), mínimo 20, máximo 999	Niveles de glucosa en sangre.
hbA1c1	≥0	Hemoglobina glicosilada.
colesHDL	≥0	Colesterol HDL.
colesLDL	≥0	Colesterol LDL.
colesTotal	≥0	Colesterol total.
trigliceridos	≥0	Triglicéridos.
creatinina	≥0	Creatinina.
sobrepeso	≥0	Indicador de sobrepeso.
juris	Numérico	Número de Jurisdicción al que pertenece.
sistolica	≥0 (mmHg), mínimo 50, máximo 300	Presión arterial sistólica.
diastolica	≥0 (mmHg), mínimo 20, máximo 150	Presión arterial diastólica.

Tabla 3.2: Variables dicotómicas y categóricas en la base de datos.

Variable	Abreviación	Descripción
DATOS PERSONALES		
sexo	sexo	Género del paciente. 1 = Mujer, 0 = Hombre.
etnia	etnia	Indica si el paciente declara pertenecer a un pueblo indígena.
ANTECEDENTES PERSONALES		
Sindrome	Sindrome	Diagnóstico de síndrome metabólico.
Dislipidemia	Dislipidemia	Diagnóstico de dislipidemia.
Obesidad	Obesidad	Diagnóstico de obesidad.
HTA	HTA	Diagnóstico de hipertensión arterial.
DM	DM	Diagnóstico de diabetes mellitus.
prediabetico	prediabetico	Indica si el paciente está en estado prediabético.
influenza	influenza	Indica si el paciente ha recibido la vacuna contra la influenza.
AC_alimenCorrecta	ACalimenCorrecta	Indica si el paciente lleva alimentación correcta.
AF_activFisica	AFactivFisica	Indica si el paciente realiza actividad física.
HT_habTabaquico	HThabTabaquico	Indica si el paciente es fumador.

Continúa en la siguiente página

Variable	Abreviación	Descripción
HA_habAlcoholico	HAhabAlcoholico	Indica si el paciente consume alcohol.
antecPersSedentarismo	aPersSedentarismo	Indica si el paciente ha llevado una vida sedentaria.
antecPersSobrepeso	aPersSobrepeso	Indica si el paciente ha tenido sobrepeso.
antecPersAlcoholismo	aPersAlcoholismo	Indica si el paciente ha tenido antecedentes de alcoholismo.
antecPersTabaquismo	antecPersTabaquismo	Indica si el paciente ha tenido antecedentes de tabaquismo.
antecPersVIH	aPersVIH	Indica si el paciente ha sido diagnosticado con VIH.
antecPersTuberculosis	aPersTuberculosis	Indica si el paciente ha sido diagnosticado con tuberculosis.
antecPersPostmenop	aPersPostmenop	Indica si el paciente está en postmenopausia. Solo se habilita si el sexo es femenino.
antecPersTeraRemplazo	aPersTeraRemplazo	Indica si el paciente ha recibido terapia de reemplazo hormonal. Solo se habilita si el sexo es femenino.
antecPersEnfCardiovascular	aPersEnfCarvascular	Indica antecedentes personales de enfermedad cardiovascular.
antecPersEnfCerebroVascular	aPersEnfCerVascular	Indica antecedentes personales de enfermedad cerebrovascular.
albuminuria	albuminuria	Indica la presencia de albuminuria.
ANTECEDENTES FAMILIARES		
ABUELOShta	ABUELOShta	Antecedentes de HTA en abuelos.
PADREShta	PADREShta	Antecedentes de HTA en padres.
TIOShta	TIOShta	Antecedentes de HTA en tíos.
HERMANOShta	HERMANOShta	Antecedentes de HTA en hermanos.
NINGUNOhta	NINGUNOhta	Sin antecedentes familiares de HTA.
ABUELOSdm	ABUELOSdm	Antecedentes de DM en abuelos.
PADRESdm	PADRESdm	Antecedentes de DM en padres.
TIOSdm	TIOSdm	Antecedentes de DM en tíos.
HERMANOSdm	HERMANOSdm	Antecedentes de DM en hermanos.
NINGUNOdm	NINGUNOdm	Sin antecedentes familiares de DM.
ABUELOSdislip	ABUELOSdislip	Antecedentes de dislipidemia en abuelos.
PADRESdislip	PADRESdislip	Antecedentes de dislipidemia en padres.
TIOSdislip	TIOSdislip	Antecedentes de dislipidemia en tíos.
HERMANOSdislip	HERMANOSdislip	Antecedentes de dislipidemia en hermanos.
NINGUNOdislip	NINGUNOdislip	Sin antecedentes familiares de dislipidemia.
ABUELOSobes	ABUELOSobes	Antecedentes de obesidad en abuelos.
PADRESobes	PADRESobes	Antecedentes de obesidad en padres.
TIOSobes	TIOSobes	Antecedentes de obesidad en tíos.
HERMANOSobes	HERMANOSobes	Antecedentes de obesidad en hermanos.

Continúa en la siguiente página

Variable	Abreviación	Descripción
NINGUNOobes	NINGUNOobes	Sin antecedentes familiares de obesidad.
ABUELOSenfCardio	ASenfCardio	Antecedentes de enfermedad cardiovascular en abuelos.
PADRESenfCardio	PADRESenfCardio	Antecedentes de enfermedad cardiovascular en padres.
TIOSenfCardio	TIOSenfCardio	Antecedentes de enfermedad cardiovascular en tíos.
HERMANOSenfCardio	HSenfCardio	Antecedentes de enfermedad cardiovascular en hermanos.
NINGUNOenfCardio	NINGUNOenfCardio	Sin antecedentes familiares de enfermedad cardiovascular.
ABUELOSenfCerebroVasc	AenfCerebroVasc	Antecedentes de enfermedad cerebrovascular en abuelos.
PADRESenfCerebroVasc	PenfCerebroVasc	Antecedentes de enfermedad cerebrovascular en padres.
TIOSenfCerebroVasc	TenfCerebroVasc	Antecedentes de enfermedad cerebrovascular en tíos.
HERMANOSenfCerebroVasc	HenfCerebroVasc	Antecedentes de enfermedad cerebrovascular en hermanos.
NINGUNOenfCerebroVasc	NenfCerebroVasc	Sin antecedentes familiares de enfermedad cerebrovascular.
COMPLICACIONES		
cRetinopatia	cRetinopatia	Diagnóstico de retinopatía.
cNeuropatia	cNeuropatia	Diagnóstico de neuropatía.
cPieDiabetico	cPieDiabetico	Diagnóstico de pie diabético.
cEnfermedadRenal	cEnfermedadRenal	Diagnóstico de enfermedad renal.
cEnfCardiovasc	cEnfCardiovasc	Diagnóstico de enfermedad cardiovascular.
cEnfCerebrovasc	cEnfCerebrovasc	Diagnóstico de enfermedad cerebrovascular.
cApneadelSueño	cApneadelSueño	Diagnóstico de apnea del sueño.
MEDICAMENTOS		
mAcidoAcetil	mAcidoAcetil	Consumo de Ácido acetil salicílico 300 mg
mMetformina	mMetformina	Consumo de Metformina 850 mg
mGlibenclamida	mGlibenclamida	Consumo de Glibenclamida 5 mg
mLinagliptina	mLinagliptina	Consumo de Linagliptina 5 mg
mAcarbosa	mAcarbosa	Consumo de Acarbosa 50 mg
mInsulinaRap	mInsulinaRap	Consumo de Insulina rápida
mInsulinaGlarg	mInsulinaGlarg	Consumo de Insulina glargina
mInsulinaNPH	mInsulinaNPH	Consumo de Insulina NPH
mInsulinaLispro	mInsulinaLispro	Consumo de Insulina lispro protamina
mComplejoB	mComplejoB	Consumo de Complejo B
mCaptopril	mCaptopril	Consumo de Captopril 25 mg
mEnalapril	mEnalapril	Consumo de Enalapril 10 mg
mNifedipino	mNifedipino	Consumo de Nifedipino 30 mg

Continúa en la siguiente página

Variable	Abreviación	Descripción
mHidroclorotiaz	mHidroclorotiaz	Consumo de Hidroclorotiazida 25 mg
mClortalidona	mClortalidona	Consumo de Clortalidona 50 mg
mMetoprolol	mMetoprolol	Consumo de Metoprolol 100 mg
mPropranolol	mPropranolol	Consumo de Propranolol 40 mg
mTelmisartan	mTelmisartan	Consumo de Telmisartán 40 mg
mLosartan	mLosartan	Consumo de Losartán 50 mg
mIrbesartan	mIrbesartan	Consumo de Irbesartán 150 mg
mAlopurinol	mAlopurinol	Consumo de Alopurinol 100 mg
mPravastatina	mPravastatina	Consumo de Pravastatina 10 mg
mAtorvastatina	mAtorvastatina	Consumo de Atorvastatina 20 mg
mBezafibrato	mBezafibrato	Consumo de Bezafibrato 200 mg
mVerapamil	mVerapamil	Consumo de Verapamil 80 mg
mFurosemida	mFurosemida	Consumo de Furosemida 40 mg

3.3. Análisis descriptivo

En esta sección se presenta un análisis descriptivo de la población de estudio, organizado en variables numéricas y categóricas. El análisis incluye un total de 33,517 individuos.

Estadísticos descriptivos de variables numéricas

Estadístico	CC	HDL	LDL	Col. Total	Creatinina	Diastólica	Edad
Media	95.43	50.48	82.33	162.35	1.04	73.21	57.92
Moda	90.00	40.00	90.00	100.00	1.00	70.00	56.00
Desv. Est.	11.49	22.73	38.39	60.54	1.33	10.24	12.21
Mínimo	40.00	10.00	10.00	10.00	0.20	20.00	18.00
Máximo	190.00	200.00	493.00	1912.00	9.00	150.00	120.00

Estadístico	Glucemia	HbA1c	IMC	Peso	Sistólica	Talla	Triglicéridos
Media	144.86	7.42	27.96	64.76	122.31	1.52	177.50
Moda	98.00	6.00	25.33	60.00	120.00	1.50	149.00
Desv. Est.	69.54	2.34	5.04	13.89	17.04	0.08	104.91
Mínimo	22.00	3.00	8.22	20.00	60.00	1.30	10.00
Máximo	991.00	19.90	73.51	150.00	245.00	2.00	1984.00

CC: Circunferencia de cintura (cm), HDL/LDL: Colesterol (mg/dL), Col. Total: Colesterol Total (mg/dL)
HbA1c: Hemoglobina glicosilada (%), IMC: Índice de Masa Corporal (kg/m²), Peso (kg), Talla (m)
Presiones en mmHg, Triglicéridos (mg/dL), Glucemia (mg/dL)

Distribución de frecuencias de variables categóricas

Variable	**Categoría 0**		**Categoría 1**		**Total**
	Frecuencia	%	Frecuencia	%	N
Sexo	8,480	25.30	25,037	74.70	33,517
Etnia	22,609	72.67	8,502	27.33	33,111
Síndrome	23,438	69.93	10,079	30.07	33,517
Dislipidemia	22,477	67.06	11,040	32.94	33,517
DM	13,126	39.16	20,391	60.84	33,517
HTA	10,366	30.93	23,151	69.07	33,517
Obesidad	21,581	64.39	11,936	35.61	33,517

Nota: Todas las variables incluyen el 100 % de los casos (Total: 33,517), excepto Etnia que tiene 33,111 casos.

El análisis de las variables numéricas revela una población con edad promedio de 57.9 años, presentando valores medios de IMC (27.96 kg/m^2) y presión arterial (122.31/73.21 mmHg) que sugieren una prevalencia importante de condiciones metabólicas. En las variables categóricas, se observa una distribución predominantemente femenina (74.7 %) y altas prevalencias de hipertensión arterial (69.1 %) y diabetes mellitus (60.8 %), mientras que la obesidad se presenta en el 35.6 % de la población. Estos patrones descriptivos establecen el perfil basal para los análisis subsecuentes.

El análisis de comorbilidades en la población de estudio (n = 33,517) revela un perfil de alta complejidad clínica, con la mayoría de los pacientes presentando múltiples condiciones simultáneamente. Solo el 1.70 % de la población (569 individuos) no presentaba ninguna de las cinco enfermedades analizadas.

Tabla 3.3: Casos con diagnóstico único por enfermedad

Enfermedad	**Casos Únicos**	**Total con la enfermedad**	**% Casos Únicos**
Síndrome Metabólico	0	10,079	0.00 %
Dislipidemia	222	11,040	2.01 %
Obesidad	351	11,936	2.94 %
Hipertensión Arterial	6,304	23,151	27.23 %
Diabetes Mellitus	4,666	20,391	22.88 %

La Tabla 3.3, evidencia la alta carga de comorbilidad en esta población, donde el 63.86 % de los pacientes presenta dos o más enfermedades simultáneamente. Destaca particularmente que el 28.59 % de la población tiene cuatro o cinco comorbilidades, indicando un subgrupo de alta complejidad clínica que requiere manejo integral multidisciplinario. Esto representa la situación de gravedad clínica de quienes están dentro del sistema de información, y por lo tanto el reto que representa la atención que requieren.

Tabla 3.4: Distribución del número de enfermedades por paciente

Número de enfermedades	Pacientes (n)	Porcentaje
0	569	1.70 %
1	11,543	34.44 %
2	11,187	33.38 %
3	635	1.89 %
4	7,140	21.30 %
5	2,443	7.29 %

La Tabla 3.4 evidencia la alta carga de comorbilidad en esta población, donde el 68.86 % de los pacientes presenta dos o más enfermedades simultáneamente. Destaca particularmente que el 28.59 % de la población tiene cuatro o cinco comorbilidades, indicando un subgrupo de alta complejidad clínica que requiere manejo integral multidisciplinario.

3.4. Correlación

Como primer paso en el análisis exploratorio, se calculó la matriz de correlación de todas las variables del glosario. Debido a la cantidad de variables, la matriz se presenta en seis secciones para facilitar su visualización. El resultado se muestra en la figura 3.1. Se observa que la edad se asocia positivamente con la hipertensión arterial, mientras que las mediciones de glucosa en sangre y hemoglobina glicosilada muestran una correlación moderada. Las condiciones crónicas presentan interconexiones importantes, particularmente entre el síndrome metabólico, la dislipidemia y la hipertensión. Un hallazgo relevante es la correlación negativa entre diabetes e hipertensión, sugiriendo que estas condiciones pueden presentarse de manera diferencial en la población estudiada. Las variables antropométricas muestran relaciones esperadas, como la fuerte asociación entre peso e índice de masa corporal. El análisis muestra conexiones importantes entre los antecedentes de las personas y su salud actual. Quienes tienen familiares directos con diabetes (padres) tienen mayor probabilidad de desarrollar diabetes también. Quienes ya tienen problemas cardíacos tienden a tener la presión arterial más alta. Quienes están en estado prediabético muestran valores más bajos de diabetes y azúcar en la sangre, lo que sugiere que es una etapa previa diferenciada. En general, los resultados indican que tanto los antecedentes familiares como los hábitos de vida influyen en el desarrollo de enfermedades crónicas como diabetes y problemas del corazón.

El análisis muestra conexiones importantes entre los medicamentos, las complicaciones de salud y las condiciones crónicas. Los medicamentos para la diabetes como metformina muestran una fuerte relación con niveles más altos de diabetes (r=0.53), lo que indica que se recetan a personas con la enfermedad más avanzada. De manera similar, los medicamentos para la presión arterial como losartán e hidroclorotiazida se asocian con hipertensión (r=0.31 y r=0.33), confirmando que se usan en pacientes que realmente necesitan estos tratamientos.

Las complicaciones como neuropatía y pie diabético se relacionan con diabetes más avanzada, mientras que problemas cardíacos y cerebrales se asocian más con presión arterial alta. Los medicamentos para el colesterol como atorvastatina y bezafibrato muestran conexión con niveles altos de colesterol, lo que confirma su uso adecuado.

En general, los resultados demuestran que los tratamientos médicos se están dirigiendo correctamente a las condiciones que pretenden controlar, y que existe coherencia entre los diagnósticos y los medicamentos recetados. El análisis muestra importantes conexiones entre los diferentes tipos de antecedentes. Se observa una relación muy fuerte entre los hábitos de tabaquismo y alcoholismo (r=0.99), lo que indica que estas conductas suelen presentarse juntas. También existe una alta correlación entre la alimentación correcta y la actividad física (r=0.75), sugiriendo que las personas con buenos hábitos alimenticios también tienden a ser más activas.

Los antecedentes familiares muestran patrones esperados: los antecedentes de diabetes, hipertensión y dislipidemia en familiares cercanos tienden a presentarse de manera agrupada. Por ejemplo, los antecedentes de diabetes en abuelos se relacionan con antecedentes de hipertensión en la misma línea familiar.

Las enfermedades personales como VIH y tuberculosis muestran una correlación moderada (r=0.42), posiblemente reflejando factores de riesgo compartidos. Los antecedentes de enfermedades cardiovasculares y cerebrovasculares también presentan cierta relación, aunque menos fuerte.

Estos resultados destacan cómo los estilos de vida y los antecedentes familiares tienden a agruparse, mostrando patrones de riesgo que pueden ser útiles para identificar grupos de población que requieren intervenciones preventivas más intensivas.

El análisis muestra las conexiones entre diferentes complicaciones de salud y los medicamentos utilizados. Se observa que algunas complicaciones tienden a presentarse juntas, como el pie diabético y la apnea del sueño (r=0.32), y la neuropatía con enfermedades cardiovasculares (r=0.17).

En cuanto a los medicamentos, se identifican patrones de prescripción lógicos. Los medicamentos para diabetes como metformina y glibenclamida muestran una correlación moderada (r=0.30), indicando que a veces se recetan juntos. De manera similar, los medicamentos para la presión arterial como hidroclorotiazida y losartán presentan una correlación positiva (r=0.29), sugiriendo que se usan en combinación para el tratamiento de la hipertensión.

Las estatinas para el colesterol (atorvastatina y bezafibrato) también muestran cierta relación (r=0.12), reflejando su uso en el manejo de lípidos. En general, los medicamentos que tratan condiciones similares tienden a tener correlaciones positivas, lo que indica patrones coherentes de prescripción médica según las necesidades de tratamiento de los pacientes.

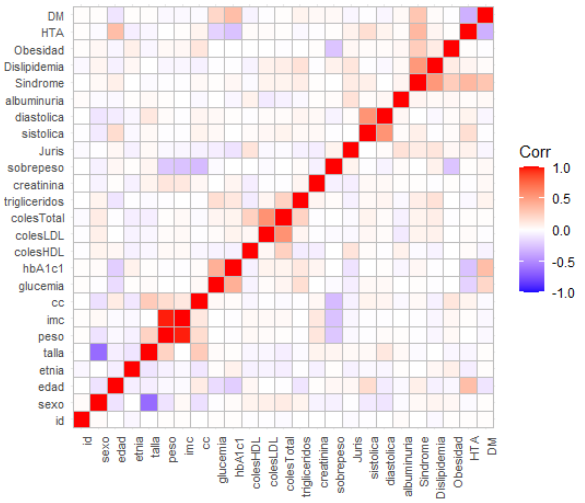
El análisis muestra conexiones importantes entre los antecedentes de salud y los medicamentos que reciben las personas. Se observa que los pacientes con antecedentes personales de enfermedades cardiovasculares tienen mayor probabilidad de recibir medicamentos para la presión arterial como metoprolol (r=0.067) y nifedipino (r=0.036).

Las personas con antecedentes familiares de diabetes (especialmente en padres) muestran mayor uso de medicamentos para la diabetes como metformina (r=0.081) y diferentes tipos de insulina. Quienes reportan no tener antecedentes familiares de diabetes (NINGUNOdm) tienen menor uso de estos medicamentos.

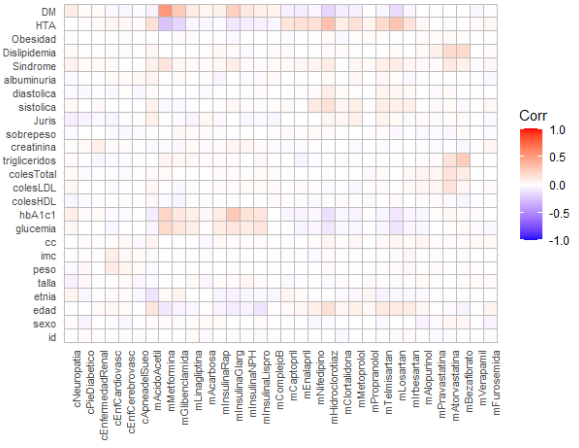
Los hábitos de vida también se relacionan con ciertos tratamientos. Las personas con alimentación correcta y actividad física muestran patrones de medicación más consistentes. Los antecedentes de tabaquismo y alcoholismo presentan correlaciones positivas con varios medicamentos para la presión arterial.

Estos resultados reflejan que los tratamientos médicos se ajustan a los antecedentes y factores de riesgo de los pacientes, mostrando una práctica clínica que considera el

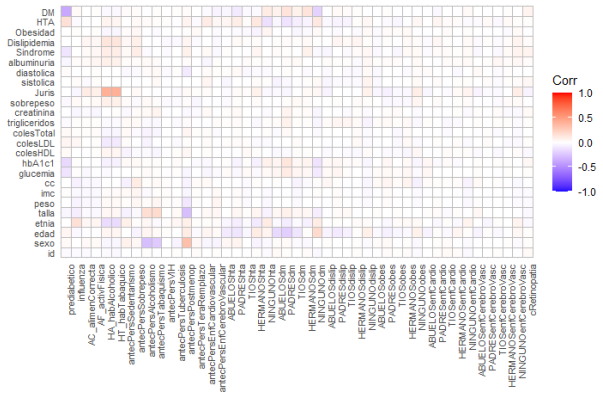
historial personal y familiar en la prescripción de medicamentos.



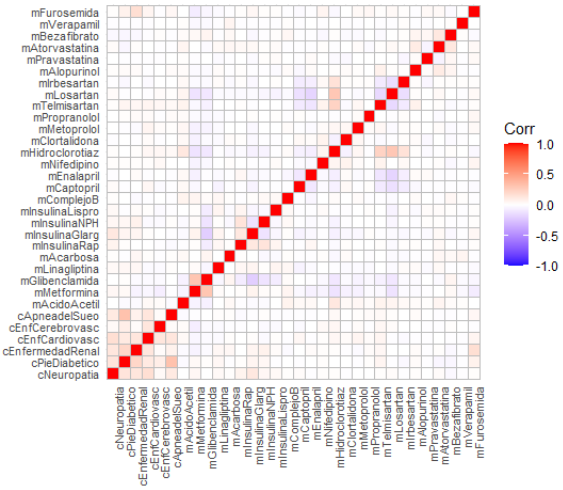
(a) Inf. gral. vs Inf. gral



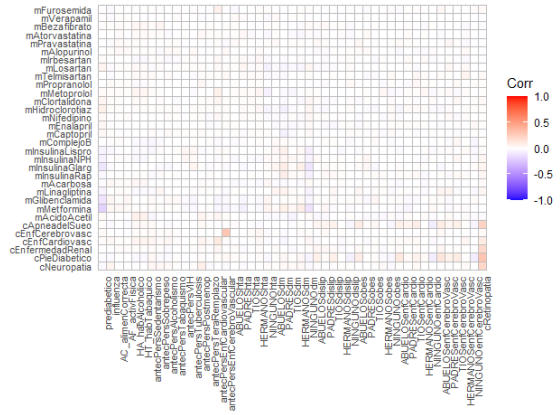
(b) Medicamentos y complicaciones vs Inf. Gral



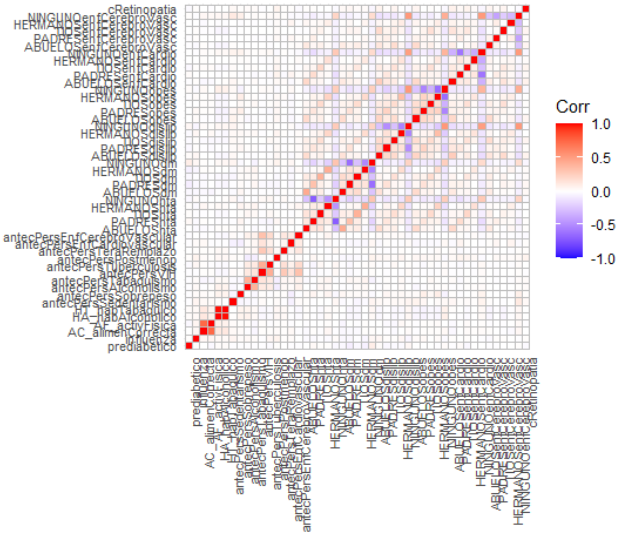
(c) Antecedentes familiares y personales vs Inf. Gral.



(d) Medicamentos y complicaciones vs Medicamentos y complicaciones



(e) Medicamentos y complicaciones vs Antecedentes familiares y personales



(f) Antecedentes familiares y personales

Figura 3.1: Matriz de correlación dividida en 6 secciones

Además del análisis de correlación general, se realizaron cinco análisis adicionales focalizados en subgrupos de pacientes con enfermedades crónicas específicas. En cada uno de estos análisis se seleccionaron exclusivamente los casos donde la variable binaria correspondiente a la condición crónica de interés presentaba valor 1 (presencia confirmada de la enfermedad). Esta aproximación permitió examinar los patrones de correlación particulares en cada población clínica definida, en contraste con el análisis general que incluía todos los valores. Los resultados de estos análisis se encuentran disponibles en el Apéndice.

3.5. Supuestos para los análisis

Los análisis que se buscan realizar preferentemente en este trabajo son Análisis discriminante lineal y regresión logística.

El estudio consideró tres supuestos fundamentales para la aplicación de análisis discriminante y regresión logística (normalidad multivariante, homogeneidad de matrices de covarianza e independencia entre las observaciones).

3.5.1. Independencia de observaciones

Es fundamental destacar que la base de datos utilizada en este estudio ha sido depurada para garantizar la independencia estadística de las observaciones. Cada registro corresponde a un paciente único, tras eliminar todas las repeticiones y consultas múltiples. Esta depuración asegura que no existan mediciones repetidas ni dependencias temporales entre los casos analizados.

3.5.2. Homogeneidad de matrices de covarianza

Para validar el supuesto de igualdad de matrices de covarianza necesario en el Análisis Discriminante Lineal (LDA), se realizó la prueba estadística M de Box para cada una de las cinco condiciones crónicas bajo estudio. Los resultados se muestran en la Tabla 3.5.

Tabla 3.5: Resultados de la prueba M de Box para homogeneidad de matrices de covarianza por condición crónica

Condición	**Estadístico M**	**Valor p**	**Conclusión**
Síndrome Metabólico		$p < 4.81068 \times 10^{-291}$	Se rechaza H_0
Dislipidemia		$p < 9.762373 \times 10^{-80}$	Se rechaza H_0
Obesidad		$p < 1.00 \times 10^{-300}$	Se rechaza H_0
Hipertensión Arterial		$p < 3.670099 \times 10^{-126}$	Se rechaza H_0
Diabetes Mellitus		$p < 1.00 \times 10^{-300}$	Se rechaza H_0

Las pruebas de Box's M para las cinco condiciones crónicas ($p < 0.001$ en todos los casos) proporcionan evidencia contundente contra la hipótesis nula de igualdad de matrices de covarianza entre grupos. Estos hallazgos justifican el uso del Análisis Discriminante Cuadrático (QDA) para todos los modelos, ya que esta técnica no requiere el supuesto de homogeneidad de covarianzas.

Por lo anterior, no es posible aplicar el análisis discriminante lineal para ninguna de las condiciones estudiadas. En su lugar, se propone análisis discriminante cuadrático para los cinco modelos.

3.5.3. Normalidad multivariante

La base de datos analizada contiene 104 variables, de las cuales 88 son dicotómicas y, por su naturaleza categórica, no siguen una distribución normal. Las 14 variables restantes corresponden a mediciones métricas continuas (excluyendo identificadores y jurisdicción), con una muestra que supera las 30,000 observaciones.

El análisis gráfico de normalidad univariante mediante gráficos Q-Q (Figuras 3.2 y 3.3) revela patrones de desviación significativos en la mayoría de las variables. Se observan sesgos, particularmente en variables como colesterol total, triglicéridos y creatinina, lo que sugiere distribuciones sesgadas.

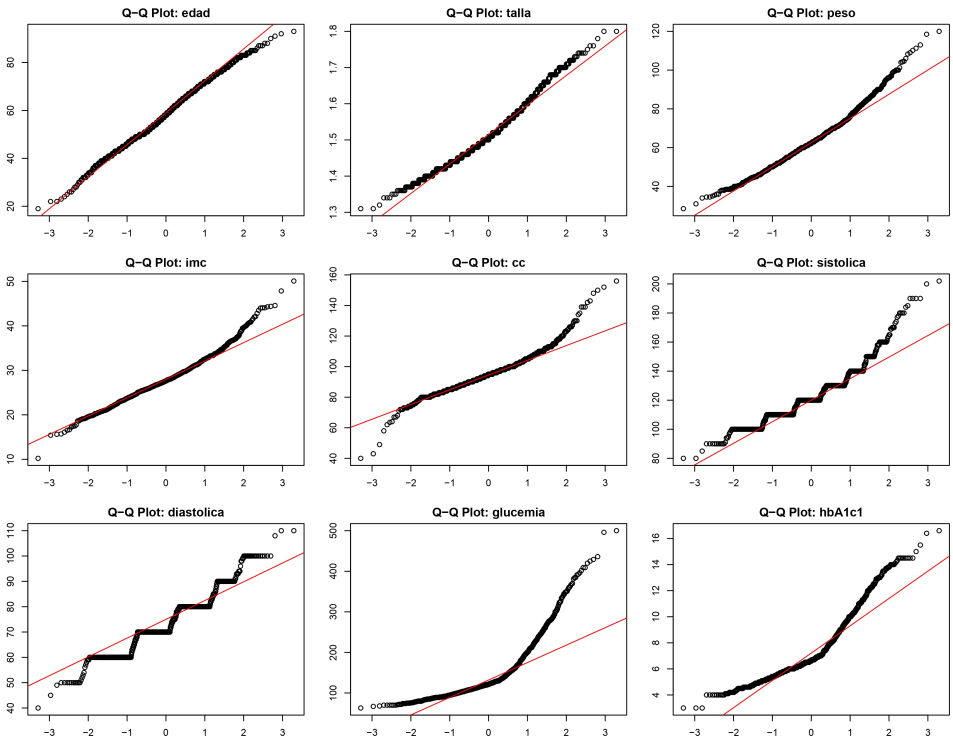


Figura 3.2: Gráfico Q-Q de normalidad para edad, talla, peso, imc, cc, presión sistólica y diastólica, glucemia y hemoglobina glicosilada.

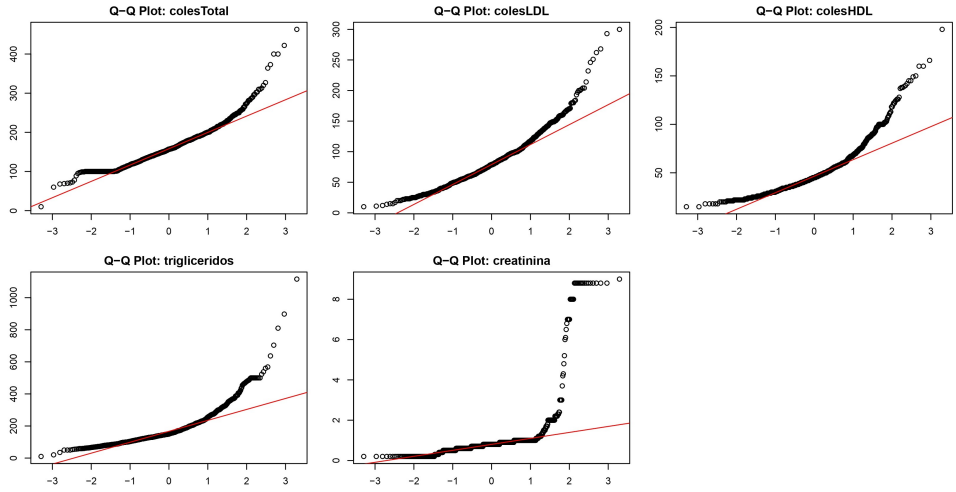


Figura 3.3: Gráfico Q-Q de normalidad para colesterol total, LDL, HDL, triglicéridos y creatinina

En cuanto al análisis multivariante, realizamos una prueba para verificar el supuesto de que tenga una distribución normal p-variada. Para ello, hicimos la siguiente prueba de hipótesis:

$$H_0 : X \sim N_p(\mu, \Sigma)$$

$$H_a : X \text{ no tiene distribución normal } p-\text{variada}$$

Los resultados de la prueba de Henze-Zirkler (Tabla 3.6) confirman el rechazo de la hipótesis de normalidad multivariada:

Test	**Estadístico**	**p-valor**	**Método**	**Normalidad**
Henze-Zirkler	5.316	<0.001	asintótico	No normal

Tabla 3.6: Resultados del test de normalidad multivariante

El p-valor significativo ($p < 0.001$) y el elevado valor del estadístico de prueba, sugieren rechazar la hipótesis nula acerca del supuesto de normalidad multivariada en el conjunto de datos. Esta falta de normalidad justifica el uso de regresión logística.

3.6. Regresión logística general sin metabólicas

Se desarrollaron modelos de regresión logística múltiple para predecir cinco condiciones médicas crónicas: síndrome metabólico, dislipidemia, obesidad, hipertensión arterial y diabetes mellitus. Los modelos incorporaron todas las variables clínicas disponibles, excluyendo únicamente las variables metabólicas que son glucemia, hbA1c1, colesHDL, colesLDL, colesTotal, triglicéridos y creatinina. La población de estudio comprendió 30,814 casos completos tras el manejo de valores faltantes, con una validación mediante partición 90 %-10 % (27,732 observaciones para entrenamiento y 3,082 para prueba).

Los modelos de regresión logística demostraron un excelente desempeño de clasificación, particularmente para diabetes mellitus (88.58 % de precisión) e hipertensión arterial

(87.90 % de precisión). Los modelos para obesidad, síndrome metabólico y dislipidemia mostraron precisiones del 82.64 %, 75.89 % y 70.73 % respectivamente. Es notable el alto balance entre sensibilidad y especificidad alcanzado en diabetes mellitus (92.52 % y 81.80 %) e hipertensión arterial (91.83 % y 79.30 %), indicando una capacidad robusta para identificar correctamente tanto casos positivos como negativos. La inclusión de todas las variables clínicas disponibles, permitió desarrollar modelos con una capacidad dee clasificación consistente a través de las cinco condiciones crónicas estudiadas.

Tabla 3.7: Tablas de confusión del Análisis de regresión logística múltiple para cada enfermedad crónica

Síndrome metabólico

	Predicho		
Real	**0**	**1**	**Sum**
0	1909	189	2098
1	554	430	984
Sum	2463	619	3082

Dislipidemia

	Predicho		
Real	**0**	**1**	**Sum**
0	1789	225	2014
1	677	391	1068
Sum	2466	616	3082

Obesidad

	Predicho		
Real	**0**	**1**	**Sum**
0	1762	177	1939
1	358	785	1143
Sum	2120	962	3082

Hipertensión arterial

	Predicho		
Real	**0**	**1**	**Sum**
0	774	202	976
1	172	1934	2106
Sum	946	2136	3082

Diabetes mellitus

	Predicho		
Real	**0**	**1**	**Sum**
0	935	208	1143
1	145	1794	1939
Sum	1080	2002	3082

Tabla 3.8: Desempeño Predictivo del modelo de regresión logística con todas las variables clínicas

Enfermedad	**Precisión Global**	**Tasa de Error**	**Sensibilidad**	**Especificidad**
Síndrome	75.89 %	24.11 %	43.70 %	90.99 %
Dislipidemia	70.73 %	29.27 %	36.61 %	88.83 %
Obesidad	82.64 %	17.36 %	68.68 %	90.87 %
Hipertensión	87.90 %	12.10 %	91.83 %	79.30 %
Diabetes	88.58 %	11.42 %	92.52 %	81.80 %

3.7. Regresión logística por bloques para dislipidemia

Dado que las variables originales incluyen HbA1c, colesterol total, colesterol LDL, colesterol HDL, triglicéridos y creatinina, resulta pertinente realizar nuevos análisis de regresión logística para dislipidemia utilizando los bloques de categorías propuestos inicialmente.

A continuación, se presentan seis tablas de confusión:

1. La primera tabla corresponde a un modelo que considera únicamente las variables predictoras de la categoría "Información General", es decir, edad, sexo, etnia, talla, peso, IMC, cc, sistolica, diastolica, glucemia, hbA1c1, colesTotal, colesLDL, colesHDL, trigliceridos, creatinina, albuminuria, síndrome metabólico, obesidad, HTA y DM.
2. El segundo análisis incluye las mismas variables de "Información General", excepto síndrome metabólico, obesidad, HTA y DM.
3. El tercer análisis considera únicamente las variables de información general sin metabólicas pero incluyendo síndrome metabólico, obesidad, HTA y DM.
4. El cuarto análisis utiliza las variables de "Información General", añadiendo exclusivamente la variable "HTA".
5. El quinto análisis se basa en la categoría de "Medicamentos y complicaciones".
6. El sexto análisis incorpora las variables de "Antecedentes personales y familiares".

Es importante destacar que al considerar las variables de HbA1c, colesterol total, colesterol LDL, colesterol HDL, triglicéridos y creatinina, se dispone de 9,700 pacientes sin valores faltantes, de los cuales el modelo utiliza 8,730 pacientes (90 %) para entrenamiento y 970 pacientes (10 %) para prueba. Por otro lado, para las categorías de antecedentes personales y familiares, así como para medicamentos y complicaciones, se cuenta con 33,513 y 33,517 pacientes completos respectivamente, utilizando aproximadamente 30,161 y 30,165 pacientes para entrenamiento y 3,352 para prueba en ambos casos.

Tabla 3.9: Desempeño Predictivo del modelo logit en diferentes configuraciones

Conf.	**Precisión Global**	**Tasa de Error**	**Sensibilidad**	**Especificidad**
1	77.94 %	22.06 %	78.98 %	76.79 %
2	59.38 %	40.62 %	73.08 %	44.25 %
3	78.62 %	21.38 %	57.41 %	89.61 %
4	61.03 %	38.97 %	74.85 %	45.77 %
5	70.64 %	29.36 %	32.22 %	89.69 %
6	66.62 %	33.38 %	3.37 %	98.70 %

Info. gral. con metabólicas incluyendo SM, Obesidad, HT y DM.
Información general con metabólicas sin ECs.
Información general sin metabólicas incluyendo SM, Obesidad, HT y DM.
Solo HT.
Medicamentos y Complicaciones.
Antecedentes personales y familiares.

Tabla 3.10: Tablas de confusión para dislipidemia con diferentes conjuntos de predictores donde la categoría de información general incluye variables metabólicas

Clasificación con categoría de información general con variables metabólicas incluyendo síndrome metabólico, obesidad, hipertensión y diabetes.

	Predicho		
Real	**0**	**1**	**Sum**
0	354	107	461
1	107	402	509
Sum	461	509	970

Clasificación con categoría de información general con metabólicas sin ECNT.

	Predicho		
Real	**0**	**1**	**Sum**
0	204	257	461
1	137	372	509
Sum	341	629	970

Clasificación con categoría de información general sin variables metabólicas pero incluyendo síndrome metabólico, obesidad, hipertensión y diabetes.

	Predicho		
Real	**0**	**1**	**Sum**
0	1819	211	2030
1	448	604	1052
Sum	2267	815	3082

Clasificación con categoría de información general incluyendo únicamente hipertensión.

	Predicho		
Real	**0**	**1**	**Sum**
0	211	250	461
1	128	381	509
Sum	339	631	970

Clasificación con categorías de medicamentos y complicaciones

	Predicho		
Real	**0**	**1**	**Sum**
0	2010	231	2241
1	753	358	1111
Sum	2763	589	3352

Clasificación con categorías de antecedentes personales y familiares

	Predicho		
Real	**0**	**1**	**Sum**
0	2195	29	2224
1	1090	38	1128
Sum	3285	67	3352

3.8. Regresión logística por bloques para hipertensión

Además, se realizó un análisis de regresión logística para hipertensión siguiendo la misma metodología aplicada en el caso de la dislipidemia. A continuación, se presentan seis tablas de confusión, cada una correspondiente a un modelo distinto:

1. La primera tabla corresponde a un modelo que considera únicamente las variables predictoras de la categoría "Información General"donde también se toman en cuenta las variables metabólicas. Incluye síndrome metabólico, dislipidemia, obesidad y diabetes.
2. El segundo análisis considera únicamente las variables de información general incluyendo variables metabólicas pero sin enfermedades crónicas.
3. El tercer análisis incluye las variables de información general sin variables metabólicas y considerando síndrome metabólico, dislipidemia, obesidad y diabetes.
4. El cuarto análisis utiliza las variables de "Información General", añadiendo exclusivamente la variable "dislipidemia".
5. El quinto análisis se basa en la categoría de "Medicamentos y complicaciones".
6. El sexto análisis incorpora las variables de "Antecedentes personales y familiares".

Para estos nuevos análisis, en el caso de la categoría de información general tenemos 9,700 pacientes sin NA, con 8,730 pacientes para entrenamiento y 970 para prueba. Para el análisis excluyendo enfermedades crónicas se cuenta con 30,818 pacientes sin NA, utilizando 27,736 para entrenamiento y 3,082 para prueba. Mientras que para las categorías de antecedentes y medicamentos hay 33,513 y 33,517 pacientes sin NA respectivamente, utilizando aproximadamente 30,161 y 30,165 pacientes para entrenamiento y 3,352 para prueba.

Tabla 3.11: Desempeño de clasificación del modelo logit para Hipertensión

Conf.	**Precisión Global**	**Tasa de Error**	**Sensibilidad**	**Especificidad**
1	85.15 %	14.85 %	91.98 %	70.55 %
2	74.74 %	25.26 %	91.83 %	38.19 %
3	85.24 %	14.76 %	92.52 %	68.75 %
4	74.64 %	25.36 %	91.53 %	38.51 %
5	86.49 %	13.51 %	89.52 %	80.13 %
6	70.44 %	29.56 %	92.95 %	20.38 %

[1] Info. gral. con metabólicas Incluyendo SM, Dislipidemia, Obesidad y DM.
[2] Información general con metabólicas sin ECs.
[3] Información general sin metabólicas e incluyendo SM, Dislipidemia, Obesidad y DM.
[4] Solo Dislipidemia.
[5] Medicamentos y Complicaciones.
[6] Antecedentes personales y familiares.

Tabla 3.12: Matriz de confusión del modelo de regresión logística para Hipertensión con diferentes conjuntos de predictores.

Análisis con categoría de información general con variables metabólicas incluyendo síndrome metabólico, obesidad, dislipidemia y diabetes .

	0	**1**	**Sum**
0	218	91	309
1	53	608	661
Sum	271	699	970

Clasificación con categoría de información general con variables metabólicas sin enfermedades crónicas.

	0	**1**	**Sum**
0	118	191	309
1	54	607	661
Sum	172	798	970

Clasificación con categoría de información general sin variables metabólicas e incluyendo síndrome metabólico, dislipidemia, obesidad y diabetes.

	0	**1**	**Sum**
0	649	295	944
1	160	1978	2138
Sum	809	2273	3082

Clasificación con categoría de información general incluyendo únicamente dislipidemia.

	0	**1**	**Sum**
0	119	190	309
1	56	605	661
Sum	175	795	970

Clasificación con categorías de medicamentos y complicaciones.

	0	**1**	**Sum**
0	867	215	1082
1	238	2032	2270
Sum	1105	2247	3352

Clasificación con categorías de antecedentes personales y familiares.

	0	**1**	**Sum**
0	212	828	1040
1	163	2149	2312
Sum	375	2977	3352

3.9. Regresión logística general con variables metabólicas

Con el objetivo de aprovechar las variables metabólicas (glucemia, hbA1c1, colesHDL, colesLDL, colesTotal, trigliceridos, creatinina), se realizó un nuevo análisis de regresión logística, esta vez incluyendo estas variables como predictoras, además de las utilizadas en el primer modelo de regresión logística presentado en este trabajo. Para este análisis, se contó con datos de 9,700 pacientes, divididos en 8,730 para entrenamiento y 970 para prueba. A continuación, se presentan las tablas de confusión correspondientes a cada enfermedad crónica evaluada.

Tabla 3.13: Tablas de confusión del modelo de regresión logística para cada enfermedad crónica donde la categoría de información general incluye variables metabólicas.

Clasificación donde síndrome es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	401	129	530
1	150	290	440
Sum	551	419	970

Clasificación donde dislipidemia es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	284	177	461
1	163	346	509
Sum	447	523	970

Clasificación donde obesidad es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	508	69	577
1	100	293	393
Sum	608	362	970

Clasificación donde hipertensión es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	244	65	309
1	58	603	661
Sum	302	668	970

Clasificación donde diabetes mellitus es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	176	57	233
1	32	705	737
Sum	208	762	970

Tabla 3.14: Desempeño de clasificación de los modelos logit para enfermedades crónicas incluyendo variables metabólicas

Enfermedad	Precisión Global	Tasa de Error	Sensibilidad	Especificidad
Síndrome	71.24 %	28.76 %	65.91 %	75.66 %
Dislipidemia	64.95 %	35.05 %	67.98 %	61.61 %
Obesidad	82.58 %	17.42 %	74.55 %	88.04 %
Hipertensión	87.32 %	12.68 %	91.22 %	78.96 %
Diabetes	90.82 %	9.18 %	95.66 %	75.54 %

[1] Precisión Global: (VP+VN)/Total
[2] Tasa de Error: (FP+FN)/Total
[3] Sensibilidad : VP/(VP+FN)
[4] Especificidad: VN/(VN+FP)

Con el fin de revisar los factores de protección y de riesgo se presentan los coeficientes de los modelos en la tabla 3.15.

Tabla 3.15: Coefficientes del modelo de regresión logística para enfermedades crónicas incluyendo variables metabólicas

Variable	Síndrome	Dislipidemia	Obesidad	HTA	DM
Información general incluyendo las variables metabólicas					
(Intercept)	0.3386	-2.1491	0.6686	-3.7322	2.2880
edad	0.0055	-0.0102	-0.0090	0.0446	0.0021
sexo	0.1919	0.2334	0.1642	0.0418	0.1395
etnia	0.1284	0.2315	0.0259	-0.0954	0.1425
talla	-5.6238	0.8665	-11.1100	-0.0932	-1.5560
peso	0.0561	-0.0050	0.1064	-0.0051	0.0185
imc	0.0004	0.0264	0.1807	0.0284	-0.0848
cc	0.0098	-0.0074	0.0297	0.0097	0.0058
sistolica	0.0034	0.0047	-0.0032	0.0071	0.0029
diastolica	-0.0103	-0.0132	-0.0049	-0.0076	-0.0105
glucemia	-0.0011	-0.0007	-0.0010	-0.0006	-0.0019
hbA1c1	0.0091	-0.0340	0.0359	-0.0645	0.2142
colesTotal	0.0003	0.0005	-0.0002	-0.0001	0.0016
colesLDL	0.0019	0.0021	-0.0016	0.0014	-0.0022
colesHDL	-0.0024	-0.0038	0.0016	-0.0033	-0.0023
trigliceridos	0.0010	0.0020	-0.0006	0.0008	-0.0002
creatinina	0.0051	-0.0029	-0.0186	-0.0243	0.0401
prediabetico	-1.2857	-0.1323	-0.1841	1.2087	-17.8000
albuminuria	0.2825	0.1106	0.3837	0.1101	0.2627
influenza	-0.1294	-0.2384	0.0271	-0.1392	-0.0795
AC_alimenCorrecta	-0.2732	-0.0637	-0.4722	-0.2652	-0.1469
AF_activFisica	0.2842	0.1409	0.5395	-0.1913	-1.0610
HT_habTabaquico	-0.0347	0.4943	1.0970	-1.8291	0.2089
HA_habAlcoholico	0.3497	0.0160	-1.0960	1.9529	-0.1781

Continúa en la siguiente página

Tabla 3.15 continuada

Variable	Síndrome	Dislipidemia	Obesidad	HTA	DM
mAcidoAcetil	0.1422	0.0949	0.1933	0.4001	-0.1175
mMetformina	0.7923	0.0730	0.3313	-0.6913	2.4180
mGlibenclamida	0.2645	0.0133	0.0083	-0.3407	1.5630
mLinagliptina	0.3536	0.0719	0.2051	-0.0545	1.4230
mAcarbosa	0.0905	-0.2145	-0.0124	-0.5116	1.9700
mInsulinaRap	0.0876	0.4202	-0.1461	-0.1018	-0.2320
mInsulinaGlarg	0.4914	0.0947	-0.1187	-0.3192	2.2540
mInsulinaNPH	0.2742	-0.1746	-0.0167	-0.2669	1.9140
mInsulinaLispro	0.2383	0.0836	-0.0221	-0.3873	1.3300
mComplejoB	0.0846	0.1713	-0.1966	-0.1688	0.3728
mCaptopril	0.8974	-0.0109	0.1770	2.6483	-0.6695
mEnalapril	0.8903	0.1578	0.2175	2.4905	-0.4096
mNifedipino	0.7101	0.1018	0.4173	2.6204	-0.0022
mHidroclorotiaz	0.1346	0.0466	0.1501	1.2417	-0.4766
mClortalidona	0.5737	0.3490	0.4746	2.1090	-0.0440
mMetoprolol	0.6680	0.0692	0.3587	2.6539	-0.0581
mPropranolol	0.8819	0.2736	0.7378	2.0351	0.1889
mTelmisartan	0.9551	0.1054	0.1665	2.8008	-0.1137
mLosartan	0.9384	0.0885	0.0696	2.5126	-0.2523
mIrbesartan	0.8459	0.0608	0.0822	2.8178	-0.4614
mAlopurinol	-0.1772	0.2985	-0.3841	-0.1706	0.1615
mPravastatina	0.4894	1.0166	0.1968	-0.1457	-0.1487
mAtorvastatina	0.4917	0.9882	0.1196	-0.0354	0.1036
mBezafibrato	0.3957	0.8720	0.1315	-0.0255	-0.0522
mVerapamil	0.4300	0.1553	0.7878	0.9787	-0.9167
mFurosemida	0.2189	0.0711	-0.0973	0.7923	-0.3065
antecPersSedentarismo	0.2238	0.1537	0.4028	-0.0458	0.0549
antecPersSobrepeso	0.2815	0.2031	0.4101	0.1664	0.1208
antecPersAlcoholismo	0.1334	0.0837	0.2696	0.3231	-0.1916
antecPersTabaquismo	-0.1269	-0.1636	-0.0615	-0.1302	0.0081
antecPersVIH	-1.7466	-0.6324	-4.8500	-0.1469	-1.4860
antecPersTuberculosis	0.2982	-0.0725	-0.2269	-0.9128	3.3510
antecPersPostmenop	0.0811	0.0309	0.1294	0.0261	0.0297
antecPersTeraRemplazo	-0.1907	-0.1963	0.0261	-0.5589	0.0620
antecPersEnfCardiovascular	0.3785	0.2341	0.4670	1.0236	-0.1218
antecPersEnfCerebroVascular	-0.4399	-0.4994	0.1171	0.3988	-0.6783
ABUELOShta	0.0118	-0.1599	-0.0848	0.4922	-0.0277
PADREShta	0.0035	-0.0076	0.1120	0.4139	-0.1586
TIOShta	0.1457	0.1587	0.1511	0.0835	-0.1912
HERMANOShta	0.1741	0.0844	0.0867	0.5587	-0.1549
NINGUNOhta	0.0339	-0.0150	0.1138	-0.1572	0.1728
ABUELOSdm	0.1083	0.0410	-0.0329	-0.0892	0.3723
PADRESdm	-0.0223	-0.0313	0.0755	-0.2493	0.2465
TIOSdm	-0.0648	-0.0485	-0.0016	-0.0362	0.1545

Continúa en la siguiente página

Tabla 3.15 continuada

Variable	Síndrome	Dislipidemia	Obesidad	HTA	DM
HERMANOSdm	0.0752	0.1370	-0.0525	-0.2242	0.3501
NINGUNOdm	-0.1566	-0.0318	0.0008	-0.0157	-0.2201
ABUELOSdislip	-0.3673	0.1984	-0.0974	0.2062	-0.1590
PADRESdislip	0.2668	0.4308	-0.0971	0.0110	-0.1881
TIOSdislip	0.3649	0.6068	0.3334	-0.4574	0.1241
HERMANOSdislip	0.1631	0.5035	0.0697	-0.1031	0.1588
NINGUNOdislip	0.3784	0.2408	0.1179	0.2589	0.0470
ABUELOSobes	-0.0014	-0.1932	-0.0328	0.1811	0.0611
PADRESobes	0.1957	0.1444	0.2615	-0.0820	0.0701
TIOSobes	0.0352	0.1093	0.2823	-0.1499	0.0026
HERMANOSobes	0.2353	0.0141	0.3079	0.0764	0.0416
NINGUNOobes	0.0067	0.0730	-0.1450	0.0393	0.2205
ABUELOSenfCardio	-0.1026	-0.0349	0.4882	0.2441	-0.7168
PADRESenfCardio	0.5055	0.3731	0.6615	0.2463	0.1474
TIOSenfCardio	0.1264	-0.0585	-0.1537	0.2396	0.1582
HERMANOSenfCardio	-0.0639	0.1195	0.2656	0.1033	-0.2100
NINGUNOenfCardio	0.1524	0.1085	0.4600	0.2051	-0.0952
ABUELOSenfCerebroVasc	-0.1603	-0.1311	-0.4739	-0.4393	-0.1074
PADRESenfCerebroVasc	-0.2200	0.1892	0.0358	-0.1246	-0.0796
TIOSenfCerebroVasc	-0.2039	0.1085	-0.9109	0.2574	0.3316
HERMANOSenfCerebroVasc	0.4450	0.3401	0.0464	-0.3196	0.0211
NINGUNOenfCerebroVasc	0.1845	0.3450	0.0979	-0.2343	-0.1118

3.10. Análisis Logit con enfermedades individuales

Además se realizó un análisis de regresión logística para la clasificación de enfermedades crónicas individuales, la cual consistió en comparar grupos puros de cada enfermedad (individuos que presentaban únicamente una condición específica) contra individuos sanos (sin ninguna de las enfermedades consideradas). Este enfoque permite identificar patrones específicos asociados a cada enfermedad de manera aislada, minimizando la interferencia de comorbilidades.

Se seleccionaron cuatro enfermedades crónicas para el análisis: Dislipidemia, Obesidad, Hipertensión Arterial (HTA) y Diabetes Mellitus (DM), ya que por la naturaleza del Síndrome metabólico no puede pertenecer a este análisis. Para cada enfermedad, se construyó un modelo de regresión logística utilizando como predictores todas las variables incluyendo metabólicas. Los modelos fueron evaluados mediante validación cruzada, utilizando 90 % de los datos para entrenamiento y 10 % para prueba, con el objetivo de determinar su capacidad de clasificación. Las tablas de confusión de estos resultados se muestran en el cuadro 3.16

Tabla 3.16: Tablas de confusión del modelo de regresión logística para cada enfermedad crónica donde la categoría de información general incluye variables metabólicas.

Clasificación donde Dislipidemia es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	39	5	44
1	11	7	18
Sum	50	12	62

Clasificación donde Obesidad es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	28	6	34
1	9	30	39
Sum	37	36	73

Clasificación donde Hipertensión es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	23	22	45
1	6	557	563
Sum	29	579	608

Clasificación donde Diabetes Mellitus es la variable dependiente

	Predicho		
Real	**0**	**1**	**Sum**
0	2	39	41
1	0	420	420
Sum	2	459	461

Tabla 3.17: Desempeño de clasificación de los modelos logit para enfermedades crónicas incluyendo variables metabólicas.

Enfermedad	Precisión Global	Tasa de Error	Sensibilidad	Especificidad
Dislipidemia	74.19 %	25.81 %	38.89 %	88.64 %
Obesidad	79.45 %	20.55 %	76.92 %	82.35 %
Hipertensión	95.39 %	4.61 %	98.93 %	51.11 %
Diabetes	91.54 %	8.46 %	100.00 %	4.88 %

[1] Precisión Global: (VP+VN)/Total
[2] Tasa de Error: (FP+FN)/Total
[3] Sensibilidad: VP/(VP+FN)
[4] Especificidad: VN/(VN+FP)

Con el fin de analizar los factores de riesgo y de protección, el cuadro 3.18 muestra los coeficientes de los modelos.

Tabla 3.18: Coeficientes de los modelos de regresión logística para enfermedades crónicas individuales (grupos puros vs sanos).

Variable	Dislipidemia	Obesidad	HTA	DM
Información general				
(Intercept)	2.9480	-29.9300	1.8960	-0.7823

Continúa en la siguiente página

Tabla 3.18 continuada

Variable	Dislipidemia	Obesidad	HTA	DM
sexo	-0.0549	0.1191	-0.0469	0.2334
edad	0.0059	-0.0293	0.0454	0.0126
etnia	-0.1041	-1.0410	-0.1209	0.4596
talla	-1.3410	12.8000	-3.4440	1.0290
peso	0.0274	-0.1746	0.0345	-0.0145
imc	-0.1243	0.7700	-0.0131	0.0539
cc	-0.0264	0.0150	-0.0222	-0.0206
sistolica	0.0101	0.0001	-0.0064	-0.0082
diastolica	-0.0187	-0.0168	0.0120	0.0070
albuminuria	1.0420	0.3474	0.1567	0.7988
prediabetico	0.7963	1.4940	1.1840	-20.4300
influenza	0.3307	0.5728	0.2493	0.2148
AC_alimenCorrecta	-0.5412	-0.6706	1.3610	0.6674
AF_activFisica	1.5630	-0.0167	-0.4112	0.1003
HA_habAlcoholico	0.6856	3.5250	0.0603	0.4187
HT_habTabaquico	-0.9056	-3.4460	-0.1053	-0.5444
antecPersSedentarismo	-0.4557	0.5087	0.1850	0.0475
antecPersSobrepeso	0.7052	0.6618	0.1077	0.2576
antecPersAlcoholismo	-0.9989	-0.3041	0.1213	0.2455
antecPersTabaquismo	0.6201	1.3700	0.5607	0.5110
antecPersVIH	NA	NA	11.8500	12.3300
antecPersTuberculosis	-17.3600	-12.8200	-0.6893	1.2220
antecPersPostmenop	1.1160	1.0970	0.3484	0.3901
antecPersTeraRemplazo	19.6300	14.9800	13.9900	13.7800
antecPersEnfCardiovascular	1.2790	-0.0344	1.8360	0.8624
antecPersEnfCerebroVascular	-17.0000	1.3850	0.8899	-0.3857
ABUELOShta	-0.8770	0.0638	0.2714	-0.2351
PADREShta	-0.0082	0.4979	1.2250	0.1366
TIOShta	-0.6553	1.2280	0.8298	-0.0972
HERMANOShta	-1.0560	1.0690	1.1070	0.0897
NINGUNOhta	-1.1100	0.2853	0.2714	-0.1446
ABUELOSdm	-1.1800	0.0674	0.5068	0.3891
PADRESdm	-0.5969	-0.2543	-0.0607	0.3376
TIOSdm	-0.5643	-0.8735	-0.4762	-0.1675
HERMANOSdm	0.1429	-1.1690	-0.3542	-0.0615
NINGUNOdm	-0.3692	0.1185	0.8554	0.2633
ABUELOSdislip	-11.7600	-24.3500	-0.0418	-0.8943
PADRESdislip	2.6550	-1.1580	0.8876	-0.1133
TIOSdislip	28.3700	17.5000	16.8500	14.2700
HERMANOSdislip	3.0780	0.7626	1.0550	0.6820
NINGUNOdislip	1.5390	0.1105	0.2138	0.0694
ABUELOSobes	-18.1200	1.3610	1.0490	0.2550
PADRESobes	-1.5470	0.5631	0.0408	0.3549
TIOSobes	-0.5516	0.1517	-0.5266	-0.3423

Continúa en la siguiente página

Tabla 3.18 continuada

Variable	Dislipidemia	Obesidad	HTA	DM
HERMANOSobes	0.5932	0.9812	0.6047	0.2931
NINGUNOobes	-0.3750	0.0178	0.2889	0.3542
ABUELOSenfCardio	1.6830	-0.6477	0.5657	0.2146
PADRESenfCardio	-0.0308	0.3681	0.8406	0.1306
TIOSenfCardio	-5.4540	14.9200	2.4870	1.2270
HERMANOSenfCardio	-0.5329	0.1629	0.8657	0.1480
NINGUNOenfCardio	0.9878	0.3999	1.5220	0.4795
ABUELOSenfCerebroVasc	-14.8500	1.7910	-0.5778	0.4867
PADRESenfCerebroVasc	-0.4986	-0.1102	-1.0460	-1.0140
TIOSenfCerebroVasc	1.4870	-15.2600	0.2070	0.0484
HERMANOSenfCerebroVasc	-17.4500	0.8240	-1.1550	-0.5724
NINGUNOenfCerebroVasc	0.3740	0.7533	-1.1660	-0.2447
cRetinopatia	-0.9303	-2.3360	-0.4196	0.1719
cNeuropatia	-2.5100	2.1960	-0.5467	0.3380
cPieDiabetico	0.6295	-0.3455	-0.0178	1.0360
cEnfermedadRenal	2.4270	3.1170	0.6971	1.5750
cEnfCardiovasc	-0.8501	-1.9180	0.1815	-1.4350
cEnfCerebrovasc	-15.3200	-17.1300	-0.7609	-1.0990
cApneadelSueo	0.0489	3.7660	0.3348	0.0717
mAcidoAcetil	-0.1500	1.0160	0.3199	-0.2485
mMetformina	-0.6098	-0.3985	-1.6650	0.5018
mGlibenclamida	-1.4880	-0.9671	-1.1420	0.1146
mLinagliptina	-1.1720	-2.4850	-1.2110	0.2852
mAcarbosa	-16.5400	-2.2630	-1.6380	-0.2954
mInsulinaRap	0.5332	0.4220	-0.1937	-0.2420
mInsulinaGlarg	-1.2210	-1.3690	-1.7200	0.2711
mInsulinaNPH	-0.3134	-0.9206	-1.5260	0.5309
mInsulinaLispro	-0.6260	-1.5080	-2.2480	-0.3143
mComplejoB	-0.6151	-0.5214	-0.1651	-0.2144
mCaptopril	-0.2481	-1.7200	2.0710	0.4087
mEnalapril	-0.2451	-0.3992	1.8680	-0.1278
mNifedipino	-18.8000	-1.1440	1.3820	-0.3158
mHidroclorotiaz	-1.5910	-0.9501	1.1470	-0.0688
mClortalidona	-14.6900	-15.7000	3.4590	0.6094
mMetoprolol	-1.7990	2.9800	1.8390	-0.0647
mPropranolol	-17.5400	-15.5000	1.2120	0.2122
mTelmisartan	-1.7590	0.8534	2.3490	0.0030
mLosartan	-0.1852	-1.8780	1.7380	-0.5123
mIrbesartan	-17.9500	-1.9290	2.2570	-0.2329
mAlopurinol	1.4340	-0.7495	-0.5630	0.6304
mPravastatina	2.3570	0.3579	1.1460	0.5672
mAtorvastatina	2.4660	-0.2460	-0.1708	0.4039
mBezafibrato	2.7180	-0.1013	0.6517	0.3702
mVerapamil	-1.4980	-12.7500	-1.0820	13.7300

Continúa en la siguiente página

Tabla 3.18 continuada

Variable	Dislipidemia	Obesidad	HTA	DM
mFurosemida	1.4040	-2.7480	1.0960	-0.3304

CAPÍTULO 4

Conclusiones

Comparación del Desempeño Predictivo entre Modelos

Se elaboraron las Tablas 4.1 y 4.2 para comparar el desempeño de clasificación de los modelos de regresión logística generales. La Tabla 4.1 contrasta el modelo que incluye variables metabólicas con aquel que no las considera, mientras que la Tabla 4.2 presenta una comparación directa entre los modelos específicos para dislipidemia e hipertensión arterial.

El presente estudio analizó, mediante modelos multivariados, las enfermedades crónicas no transmisibles registradas en el Sistema de Información en Crónicas (SIC) del Estado de Hidalgo. Es fundamental precisar que la base de datos del SIC contiene exclusivamente pacientes con diagnóstico previo de alguna ECNT. Por lo tanto, los modelos desarrollados no estiman el riesgo poblacional de enfermar, sino que describen los patrones de coexistencia clínica y la capacidad de distintos conjuntos de variables para diferenciar, dentro de una población ya enferma, a aquellos pacientes que presentan una condición específica de quienes no la tienen.

Para síndrome metabólico, el modelo con variables metabólicas presenta una sensibilidad notablemente superior (65.91 % vs 43.70 %), lo que indica una mejor capacidad para detectar casos positivos, aunque con menor especificidad (75.66 % vs 90.99 %) y precisión global (71.24 % vs 75.89 %). Un patrón similar se observa en dislipidemia, donde la sensibilidad mejora sustancialmente (67.98 % vs 36.61 %), pero con deterioro en especificidad (61.61 % vs 88.83 %) y precisión global (64.95 % vs 70.73 %). Este incremento en la sensibilidad al incluir indicadores bioquímicos refleja que variables como la glucosa, HbA1c, colesterol, triglicéridos y creatinina actúan como marcadores centrales del estado metabólico del paciente, permitiendo una identificación más certera de las alteraciones subyacentes a las ECNT dentro del SIC.

Para obesidad, ambos modelos muestran rendimiento comparable en precisión global (82.58 % vs 82.64 %), pero el modelo con variables metabólicas logra mayor sensibilidad (74.55 % vs 68.68 %) con mínima pérdida de especificidad. En hipertensión arterial, las diferencias son marginales, manteniéndose alta sensibilidad en ambos casos. Destaca diabetes mellitus, donde el modelo con variables metabólicas alcanza el mejor desempeño general con precisión global del 90.82 % y sensibilidad del 95.66 %, superando claramente al modelo base. Cabe resaltar que, incluso sin las mediciones metabólicas, los modelos generales para diabetes e hipertensión mostraron precisiones superiores al 87 %. Esto demuestra que la combinación de variables clínicas, antropométricas y de comorbilidad permite distinguir adecuadamente a los pacientes con estas enfermedades dentro del sistema, lo cual es relevante en contextos donde los exámenes bioquímicos no están disponibles de rutina.

En conjunto, el modelo con variables metabólicas mejora sistemáticamente la detección

de casos positivos (sensibilidad), aunque a expensas de una reducción en la especificidad y, en algunos casos, en la precisión global. Esto sugiere que, si bien los indicadores metabólicos fortalecen la identificación de pacientes afectados, también introducen un mayor número de falsos positivos en este contexto clínico específico de pacientes ya diagnosticados.

Tabla 4.1: Comparación de modelos de regresión logística con y sin variables metabólicas

Enfermedad	Variables	Precisión	Error	Sensibilidad	Especificidad
Síndrome	Sin metabólicas	75.89 %	24.11 %	43.70 %	90.99 %
	Con metabólicas	71.24 %	28.76 %	65.91 %	75.66 %
Dislipidemia	Sin metabólicas	70.73 %	29.27 %	36.61 %	88.83 %
	Con metabólicas	64.95 %	35.05 %	67.98 %	61.61 %
Obesidad	Sin metabólicas	82.64 %	17.36 %	68.68 %	90.87 %
	Con metabólicas	82.58 %	17.42 %	74.55 %	88.04 %
Hipertensión	Sin metabólicas	87.90 %	12.10 %	91.83 %	79.30 %
	Con metabólicas	87.32 %	12.68 %	91.22 %	78.96 %
Diabetes	Sin metabólicas	88.58 %	11.42 %	92.52 %	81.80 %
	Con metabólicas	90.82 %	9.18 %	95.66 %	75.54 %

Tabla 4.2: Comparación de diferentes configuraciones para Dislipidemia e Hipertensión

Configuración	Precisión	Error	Sensibilidad	Especificidad
Dislipidemia				
Info general + ECs	77.94 %	22.06 %	78.98 %	76.79 %
Info general sin ECs	59.38 %	40.62 %	73.08 %	44.25 %
Info general sin met + ECs	78.62 %	21.38 %	57.41 %	89.61 %
Solo HTA	61.03 %	38.97 %	74.85 %	45.77 %
Medicamentos	70.64 %	29.36 %	32.22 %	89.69 %
Antecedentes	66.62 %	33.38 %	3.37 %	98.70 %
Hipertensión				
Info general + ECs	85.15 %	14.85 %	91.98 %	70.55 %
Info general sin ECs	74.74 %	25.26 %	91.83 %	38.19 %
Info general sin met + ECs	85.24 %	14.76 %	92.52 %	68.75 %
Solo Dislipidemia	74.64 %	25.36 %	91.53 %	38.51 %
Medicamentos	86.49 %	13.51 %	89.52 %	80.13 %
Antecedentes	70.44 %	29.56 %	92.95 %	20.38 %

El análisis comparativo de los modelos por bloques para dislipidemia e hipertensión (Tabla 4.2) arroja conclusiones relevantes sobre la naturaleza de estas enfermedades dentro del SIC. En general, los modelos que incorporan variables clínicas y metabólicas directas demuestran un rendimiento predictivo superior para distinguir entre pacientes enfermos, mientras que aquellos basados principalmente en antecedentes presentan limitaciones importantes.

Para la dislipidemia, el modelo que incluye todas las variables de información general junto con síndrome metabólico, obesidad, HTA y DM (Configuración 1) muestra el mejor equilibrio global. La dificultad para identificar la dislipidemia de forma aislada radica en que actúa como un condicionante fisiopatológico común a casi todas las ECNT y es un componente central del síndrome metabólico. Por ello, los bloques que incluyen otras enfermedades metabólicas logran una clasificación más precisa, reflejando la coexistencia de estas condiciones en la población clínicamente afectada. Los modelos basados únicamente en información general o antecedentes familiares tienen una capacidad discriminante menor, ya que en una población donde todos los pacientes tienen al menos una ECNT, estas variables no logran diferenciar efectivamente a los subgrupos.

En el análisis de hipertensión arterial se observa un patrón elucidante. El modelo con medicamentos y complicaciones (Configuración 5) ofrece el mejor desempeño (86.49 % de precisión). Esto se debe a que, dentro del SIC, la hipertensión rara vez se presenta como una entidad aislada, sino que se manifiesta acompañada de otras comorbilidades, tratamientos farmacológicos y complicaciones. Los medicamentos antihipertensivos y la presencia de complicaciones actúan así como marcadores directos de la enfermedad establecida y de su grado de control clínico, permitiendo una mejor distinción entre los pacientes que la padecen. Este hallazgo responde a la pregunta de qué información permite distinguir mejor, dentro de una población ya enferma del SIC, a los pacientes hipertensos, destacando el valor de los datos de tratamiento y evolución clínica.

Comparando transversalmente ambos grupos de modelos, resulta evidente la marcada interdependencia entre las condiciones metabólicas. La exclusión de las enfermedades crónicas relacionadas deteriora significativamente el desempeño, confirmando que su inclusión como predictores es crucial para reflejar la realidad clínica de los pacientes del SIC. La selección del modelo ideal dependerá, por tanto, del objetivo específico: si se prioriza identificar el mayor número de casos (alta sensibilidad), se optará por modelos con variables metabólicas o de medicamentos; si se busca minimizar falsos positivos (alta especificidad), los modelos más conservadores podrían ser preferibles.

Interpretación del Modelo de Análisis Logit con Enfermedades Individuales

El análisis de regresión logística para enfermedades individuales constituye un ejercicio metodológico complementario que busca aislar los patrones asociados a cada condición crónica cuando se presenta de forma pura, es decir, en pacientes que padecen únicamente una de las enfermedades analizadas (dislipidemia, obesidad, hipertensión arterial o diabetes mellitus), comparándolos con individuos sanos dentro del SIC. Este enfoque permite observar la "firma" específica de cada enfermedad, minimizando el ruido estadístico introducido por la multimorbilidad, que es la norma en esta población.

Los resultados de este modelo ofrecen varias interpretaciones clave:

- **Alta capacidad discriminante en formas puras**: La hipertensión arterial y la diabetes mellitus lograron sensibilidades muy altas (98.93 % y 100 %, respectivamente), indicando que los modelos pueden identificar casi todos los casos positivos cuando estas enfermedades no están enmascaradas por otras condiciones. Sin embargo, la baja especificidad observada (especialmente del 4.88 % en diabetes) sugiere que muchos pacientes sanos son clasificados incorrectamente, revelando que los perfiles clínicos en el SIC, incluso en ausencia de un diagnóstico específico, comparten características con los enfermos.

- **El caso particular de la dislipidemia aislada**: Este modelo confirma la dificultad de identificar una dislipidemia pura. Con una precisión global del 74.19 % y una sensibilidad moderada (38.89 %), se evidencia que, incluso en ausencia de otras ECNT, el perfil clínico de la dislipidemia no es suficientemente distintivo utilizando las variables disponibles. Esto refuerza la noción de que la dislipidemia actúa más como un **sustrato metabólico compartido** que como una entidad clínica aislada con marcadores únicos claros.

- **Relevancia en el contexto del SIC**: Si bien este modelo proporciona información valiosa sobre la fisiopatología aislada, su utilidad práctica directa en el SIC es limitada, dado que la realidad clínica predominante es la multimorbilidad (solo el 1.70 % de la población no tenía ECNT y 34.44 % no tenía comorbilidades). No obstante, sirve como **referencia teórica** para entender qué variables serían más relevantes si las enfermedades se presentaran de forma aislada, y confirma que la fuerza predictiva de muchas variables (como los medicamentos) se deriva precisamente de su uso en contextos de enfermedad compleja y establecida.

Análisis de Coeficientes

El análisis de los coeficientes del modelo de regresión logística general utilizando variables metabólicas revela patrones epidemiológicos significativos sobre los factores asociados a la presencia de enfermedades crónicas en la población del SIC. Es importante recordar que estos coeficientes describen asociaciones dentro de un universo de pacientes ya diagnosticados, por lo que reflejan la gravedad, el manejo y la coexistencia de condiciones más que factores de riesgo etiológicos tradicionales.

Para el síndrome metabólico, se identifican efectos protectores asociados con la talla (-5.62), hábitos alimenticios correctos (-0.27) y niveles elevados de colesterol HDL (-0.0024). Por el contrario, el peso (0.056), la circunferencia de cintura (0.0098) y el uso de medicamentos como captopril (0.897) y metformina (0.792) se asocian con un mayor riesgo, lo que probablemente indica la presencia de un cuadro clínico más complejo y que requiere tratamiento farmacológico.

En dislipidemia, el colesterol LDL (0.0021) y los triglicéridos (0.0020) muestran el efecto esperado como marcadores de la enfermedad. El uso de estatinas como pravastatina (1.016) y atorvastatina (0.988) presenta coeficientes positivos elevados, lo que confirma que estos medicamentos se prescriben a pacientes con dislipidemia confirmada, actuando como un indicador de la condición más que como su causa.

La obesidad presenta el patrón más definido, donde la talla (-11.11) ejerce un efecto protector marcado, mientras el peso (0.106) e IMC (0.1807) son los principales factores de riesgo. El hábito tabáquico (1.097) también muestra una asociación positiva, un hallazgo que merece mayor investigación para entender su relación con el metabolismo en esta población.

Para hipertensión arterial, la edad (0.0446) emerge como el principal factor asociado. Todos los medicamentos antihipertensivos muestran coeficientes positivos muy altos (e.g., telmisartán 2.800, irbesartán 2.817), lo que refuerza la noción de que, dentro del SIC, el uso de estos fármacos es un marcador robusto de la presencia de la enfermedad.

La diabetes mellitus muestra el patrón más extremo, con la hemoglobina glicosilada (0.2142) como principal marcador. El estado prediabético (-17.8) aparece como un factor "protector"paradójico, lo que muy probablemente refleja una separación completa en los

datos (los prediabéticos en el estudio no progresaron a diabetes) o el efecto de intervenciones preventivas efectivas, y no una relación causal inversa.

Los antecedentes familiares presentan efectos diferenciados según la enfermedad y el parentesco, siendo más consistentes para la dislipidemia. Los estilos de vida muestran relaciones complejas que a veces contradicen la intuición (e.g., tabaquismo como factor de riesgo para obesidad pero "protector"para hipertensión), lo que probablemente refleja sesgos de indicación, factores de confusión no medidos o las complejas interacciones entre conductas y fisiología en una población bajo tratamiento médico.

El análisis de regresión logística para grupos puros (pacientes con una sola enfermedad vs. sanos) revela patrones epidemiológicos aún más específicos, aunque con la limitante de tamaños muestrales reducidos que pueden generar coeficientes extremos. Este análisis refuerza la idea de la centralidad de los indicadores metabólicos y el tratamiento farmacológico como elementos definitorios del perfil clínico de los pacientes dentro del SIC.

Conclusiones Generales

Este estudio permitió analizar mediante modelos multivariados las enfermedades crónicas no transmisibles registradas en el SIC del Estado de Hidalgo. La evidencia demuestra consistentemente que los modelos que incorporan variables clínicas y metabólicas directas ofrecen el mejor equilibrio predictivo para distinguir, dentro de esta población ya enferma, a los pacientes con una condición específica. La inclusión de variables metabólicas (glucemia, HbA1c, perfil lipídico) mejora sustancialmente la sensibilidad, especialmente para síndrome metabólico y dislipidemia, confirmando que estos indicadores son marcadores centrales del estado fisiopatológico subyacente.

Los hallazgos subrayan la marcada interdependencia entre las condiciones metabólicas. La hipertensión y la dislipidemia, en particular, no se presentan como entidades aisladas, sino insertas en un complejo de comorbilidades, tratamientos y complicaciones. Por ello, los modelos que incorporan estas variables relacionadas obtienen un mejor desempeño. La dislipidemia se revela como una condición de base difícil de aislar, actuando como un sustrato común para otras ECNT.

La selección final del enfoque de modelado óptimo dependerá críticamente del objetivo clínico-operativo específico, balanceando el costo de los falsos positivos versus los falsos negativos. Para fines de identificación y tamizaje dentro del sistema de salud, los modelos con alta sensibilidad (que incluyen variables metabólicas) pueden ser más útiles. Para confirmación diagnóstica o estudios de precisión, podrían preferirse modelos con mayor especificidad.

Finalmente, es crucial reiterar la naturaleza descriptiva de estos modelos. Al provenir de una base de datos de pacientes previamente diagnosticados (SIC), los resultados no deben interpretarse como predictores de riesgo en la población general, sino como herramientas poderosas para entender los patrones de coexistencia, gravedad y manejo clínico de las ECNT en un contexto real de atención médica. Esto valida la utilidad del SIC como fuente de información para la toma de decisiones en salud pública y la gestión de pacientes crónicos, destacando la importancia de registrar de manera sistemática tanto las variables metabólicas como los datos de tratamiento y evolución clínica.

4.1. Trabajos futuros

Dado que el análisis realizado en esta investigación se llevó a cabo de manera general sobre las bases de datos disponibles, un posible desarrollo futuro sería profundizar el estudio a nivel jurisdiccional o por localidad. Este enfoque permitiría obtener resultados más precisos y específicos, considerando las particularidades de cada región.

Asimismo, sería valioso realizar un análisis estratificado por grupos de edad limitada, ya que en ciertas jurisdicciones existen inconsistencias o faltantes en algunas variables, lo que se relaciona con haber encontrado que hay mayor sensibilidad en los modelos de QDA cuando se compara la población adulta y con niños.

Adicionalmente, incorporar la búsqueda de estados de remisión de las enfermedades podría enriquecer los análisis de interés, permitiendo evaluar no solo la prevalencia sino también la dinámica temporal de las condiciones de salud y la efectividad de intervenciones médicas o políticas públicas en diferentes contextos geográficos o demográficos.

CAPÍTULO 5

Apéndice A. Códigos

5.1. Prueba de homogeneidad de matrices de covarianza

Se aplicó la prueba M de Box para evaluar la homogeneidad de las matrices de covarianza en la base de datos *dat*. El análisis se realizó sobre las variables cuantitativas *metricas*, agrupadas según la presencia (1) o ausencia (0) de la enfermedad crónica. Para mejores resultados se eliminaron NA´s, se balancearon los grupos, las variables de enfermedades se convirtieron a factores y se estandarizaron las variables predictoras.

```
# Limpieza de datos y aplicación de la prueba
dat_clean = na.omit(dat[, c(metricas, "Enfermedad")])
# Prueba M de Box
boxM(dat[, metricas], dat[, "Enfermedad"])
```

5.2. Pruebas de normalidad

La prueba de normalidad gráfica realizada fue Q-Q plot de normalidad. El siguiente código hace la prueba únicamente para las variables métricas que se encuentran en un vector y guarda los gráficos en un pdf.

```
guardar_qq_pdf <- function(data, variables, n_muestra = 1000,
    filename = "qq_plots.pdf") {
  pdf(filename, width = 11, height = 8.5)  # Tamaño carta

  # Configurar layout para 9 gráficos por página
  par(mfrow = c(3, 3))
  par(mar = c(3, 3, 2, 1))

  for (var in variables) {
    datos_completos <- na.omit(data[[var]])
    if (length(datos_completos) > n_muestra) {
      muestra <- sample(datos_completos, n_muestra)
    } else {
      muestra <- datos_completos
    }

    qqnorm(muestra, main = paste("Q-Q Plot:", var))
    qqline(muestra, col = "red")
  }
```

```
19
20        dev.off()
21        cat("Gráficos guardados en:", filename, "\n")
22      }
```

Además se realizó el siguiente código que nos regresa un análisis Henze-Zirkler multivariado y Anderson-Darling univariado en donde la base de datos se llama *dat* y el vector con las variables métricas se llama *metricas*.

```
1     library(MVN)
2     dat_metricas <- dat[metricas]
3     mvn(data = dat_metricas)
```

5.3. Código del análisis discriminante lineal

La función *lda_analysis* realiza un análisis de discriminante lineal para predecir una variable categórica usando un conjunto de predictores y la base de datos *dat*. Primero, filtra los datos para incluir solo la variable objetivo y las predictoras, eliminando filas con valores faltantes. Luego, divide los datos en 90 % para entrenamiento y 10 % para prueba, con semilla *set.seed*(4646). El modelo LDA se construye con la función lda del paquete MASS, usando la fórmula variable predictoras. Después, predice las clases en el conjunto de prueba y genera una matriz de confusión para comparar valores reales vs. predichos. Finalmente, imprime los coeficientes de discriminación.

```
1  library(MASS)
2
3  lda_analysis = function(variable, columnas_predictoras) {
4    # Filtrar solo las columnas necesarias y eliminar NAs
5    datos_analisis = na.omit(dat[, c(variable, columnas_predictoras
       )])
6
7    # División de datos
8    set.seed(4646)
9    tamaño = nrow(datos_analisis)
10   casos_entrena = floor(0.9 * tamaño)
11   indices = sample(1:tamaño, size = casos_entrena)
12
13   entrena = datos_analisis[indices, ]
14   prueba = datos_analisis[-indices, ]
15
16   formula_str = as.formula(paste(variable, "~",
17   paste(columnas_predictoras, collapse = "+")))
18   lda_model = lda(formula_str, data = entrena)
19
20   # Predicciones y evaluación
21   predicciones = predict(lda_model, newdata = prueba)
22   tabla_confusion = table(Real = prueba[[variable]], Predicho =
       predicciones$class)
23
24   print(addmargins(tabla_confusion))
25   print(lda_model$scaling)
```

```
}
```

5.4. Código del análisis discriminante cuadrático

La función *qda_analysis* realiza un análisis de discriminante cuadrático para predecir una variable categórica usando un conjunto de predictores y la base de datos *dat* . Primero, filtra los datos para incluir solo la variable objetivo y las predictoras, eliminando filas con valores faltantes. Luego, divide los datos en 90 % para entrenamiento y 10 % para prueba, con semilla *set.seed*(4646) . El modelo QDA se construye con la función *qda* del paquete MASS, usando la fórmula *variablepredictoras* , que estima matrices de covarianza separadas para cada grupo. Después, predice las clases en el conjunto de prueba, genera una matriz de confusión para comparar valores reales vs. predichos, y muestra las proporciones previas de los grupos. A diferencia del LDA, este enfoque no asume igualdad de covarianzas entre clases ni imprime coeficientes lineales.

```
library(MASS)

# Función para análisis QDA
qda_analysis = function(variable, columnas_predictoras) {

    # Filtrar datos completos
    datos_analisis = na.omit(dat[, c(variable, columnas_
       predictoras)])

    # División entrenamiento/prueba (90%/10%)
    set.seed(4646)
    tamaño = nrow(datos_analisis)
    casos_entrena = floor(0.9 * tamaño)
    indices = sample(1:tamaño, size = casos_entrena)

    entrena = datos_analisis[indices, ]
    prueba = datos_analisis[-indices, ]

    # Ajustar modelo QDA
    formula_str = as.formula(paste(variable, "~", paste(columnas_
       predictoras, collapse = "+")))
    qda_model = qda(formula_str, data = entrena)

    # Predicciones y evaluación
    predicciones = predict(qda_model, newdata = prueba)
    tabla_confusion = table(Real = prueba[[variable]], Predicho =
        predicciones$class)

    print(addmargins(tabla_confusion))

    print(qda_model$prior)
    invisible(list(
        modelo = qda_model,
        datos_entrena = entrena,
        datos_prueba = prueba,
```

```
        confusion = tabla_confusion,
        precision = precision,
        nas_originales = nas_por_columna
    ))
}
```

5.5. Código del análisis de regresión logística

La función *logit_analysis* en R realiza un análisis de regresión logística para predecir una variable binaria (presencia o ausencia de la enfermedad) usando un conjunto de predictores llamado *columnas_pedictoras* y la base de datos *dat*. Primero, filtra los datos para incluir solo la variable objetivo y las predictoras, eliminando filas con valores faltantes. Luego, divide los datos en un 90 % para entrenamiento y 10 % para prueba, con semilla *set.seed*(4646). Construye el modelo logístico usando la función glm con la fórmula variable . (donde el punto representa todas las predictoras), ajustándolo a los datos de entrenamiento. Después, predice las probabilidades en el conjunto de prueba y las convierte en clases binarias (0 o 1) usando un umbral de 0.5. Finalmente, imprime la matriz de confusión (comparando valores reales vs. predichos) y los coeficientes del modelo.

```
logit_analysis = function(variable, columnas_predictoras) {
  # Filtrar el conjunto de datos
  datos_analisis = dat[, c(variable, columnas_predictoras)]

  # Eliminar filas con NA
  datos_analisis = na.omit(datos_analisis)

  # División de datos
  set.seed(4646)
  tamaño = nrow(datos_analisis)
  casos = floor(0.9 * tamaño)
  indices = sample(1:tamaño, size = casos)
  entrena = datos_analisis[indices, ]
  prueba = datos_analisis[-indices, ]

  formula_str = as.formula(paste(variable, "~ ."))
  logit_model = glm(formula_str, data = entrena, family =
     binomial)

  # Predicciones
  prob_pred = predict(logit_model, newdata = prueba, type = "
     response")
  clas = ifelse(prob_pred > 0.5, 1, 0)

  # Matriz de confusión
  tabla = table(Real = prueba[[variable]], Predicho = clas)
  print(addmargins(tabla))
  print(coef(logit_model))
}
```

CAPÍTULO 6

Apéndice B. Visualización del SIC

La Secretaría de Salud, en un esfuerzo por optimizar el manejo de información, utiliza el Sistema Nominal de Información en Crónicas (SIC). Este sistema, en su versión 3.0, permite el registro nominal de datos de pacientes y su interfaz se visualiza de la siguiente forma: La imagen 6.1 muestra una captura de pantalla del Sistema Nominal de Información en Crónicas (SIC) que corresponde al módulo de "Alta de Paciente", donde se registran los datos personales del paciente durante su primer ingreso al sistema. En la figura 6.2 se observa específicamente el apartado de antecedentes familiares y personales. Cuando un nuevo paciente se registra en el sistema, se requiere completar una sección de datos basales que no fue incluida en este análisis debido al alto porcentaje de información faltante en estos rubros. Esto se debe a que, por ejemplo, si un paciente tenía 15 consultas registradas, faltarían 14 registros en las variables de tipo basal. No obstante, el sistema cuenta con un módulo específico para el registro de mediciones donde, en cada consulta subsiguiente, se actualizan estos datos y se lleva el control de las cinco enfermedades crónicas que constituyen el objetivo principal de este sistema como se puede observar en la figura 6.3. Adicionalmente, en esta misma pestaña se encuentra el apartado de tratamiento no farmacológico. Por último, en la figura 6.4, para el registro del tratamiento farmacológico, el sistema contiene pestañas con listas desplegables de medicamentos que el paciente está tomando al momento de la consulta. Este módulo también permite registrar complicaciones asociadas, grado de adicción al tabaco, estado de vacunación antiinfluenza, referencias a otros centros médicos para el control de las enfermedades y procesos de baja del sistema. Cada uno de estos elementos se documenta de manera sistemática para garantizar un seguimiento integral del paciente dentro del sistema de salud.

El nombre que aparece en las figuras recientes es ficticio.

Figura 6.1: Alta de paciente

Figura 6.2: Antecedentes familiares y personales

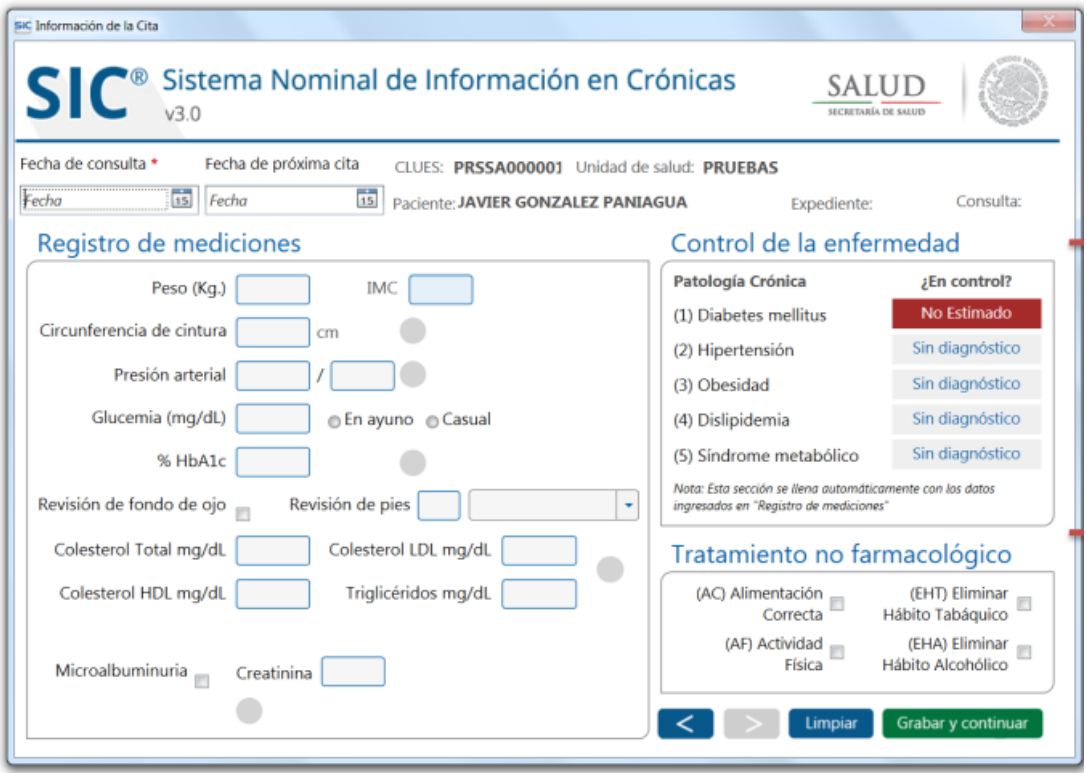


Figura 6.3: Registro de mediciones, control de enfermedad y tratamiento no farmacológico

Figura 6.4: Tratamiento farmacológico, complicaciones

Universidad Autónoma del Estado de Hidalgo
Instituto de Ciencias Básicas e Ingeniería
School of Engineering and Basic Sciences
Área Académica de Matemáticas y Física
Department of Physics and Mathematics

Mineral de la Reforma, Hgo., a 1 de agosto de 2024

Número de control: ICBI-AAMyF/2740/2024
Asunto: Carta de Confidencialidad

Los que suscriben, Dr. Roberto Ávila-Pozos, profesor investigador del Área Académica de Matemáticas y Física del Instituto de Ciencias Básicas e Ingeniería, perteneciente a la Universidad Autónoma del Estado de Hidalgo, y Daira Yalín Hernández Cortés, estudiante de la Licenciatura en Matemáticas Aplicadas de la UAEH, hacemos constar, en relación al protocolo titulado: **Análisis Multivariado de factores asociados a enfermedades crónicas** que nos comprometemos a resguardar, mantener la confidencialidad y no hacer mal uso de la base de datos del Sistema Nominal de Información en Crónicas (SIC) del Estado de Hidalgo o bien, cualquier otro registro o información relacionada con el estudio mencionado a mi cargo, o en el cual participo como investigador, así como a no difundir, distribuir o comercializar los datos personales contenidos en los sistemas de información, desarrollados en la ejecución del mismo.
Estando en conocimiento de que en caso de no dar cumplimiento se procederá acorde a las sanciones civiles, penales o administrativas que procedan de conformidad con lo dispuesto en la Ley Federal de Transparencia y Acceso a la Información Pública Gubernamental, la Ley Federal de Protección de Datos Personales en Posesión de los Particulares y el Código Penal en las entidades federativas, a la Ley Federal de Protección de Datos Personales en Posesión de los Particulares, y demás disposiciones aplicables en la materia.

Atentamente
"Amor, Orden y Progreso"

Dr. Roberto Ávila Pozos
Jefe del Área Académica de Matemáticas y Física

Daira Yalín Hernández Cortés
Estudiante de la Licenciatura en Matemáticas Aplicadas.

Dirección General
02 AGO 2024
RECIBIDO
NOMBRE:
HORA: 11:38
ANEXO:

Ciudad del Conocimiento, Carretera Pachuca-Tulancingo Km. 4.5 Colonia Carboneras, Mineral de la Reforma, Hidalgo, México. C.P. 42184
Teléfono: 52 (771) 71 720 00 Ext. 40124, 40119
aamyf_icbi@uaeh.edu.mx, ravila@uaeh.edu.mx

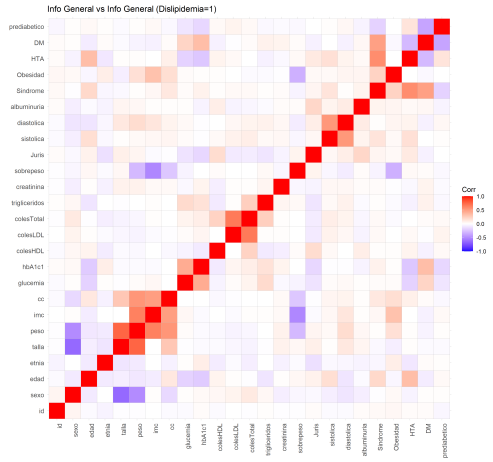
uaeh.edu.mx

Figura 6.5: Carta de confidencialidad entregada a la Secretaría de Salud

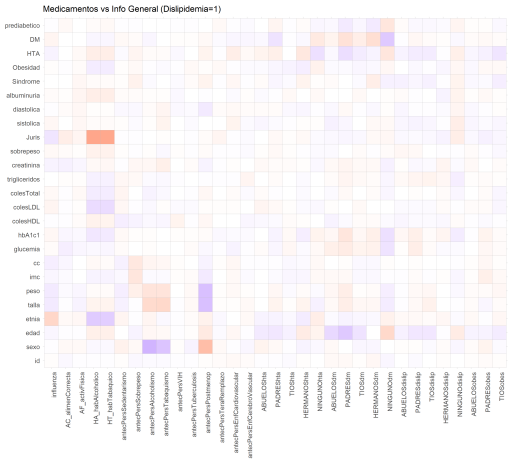
CAPÍTULO 7

Apéndice C. Imágenes de correlación

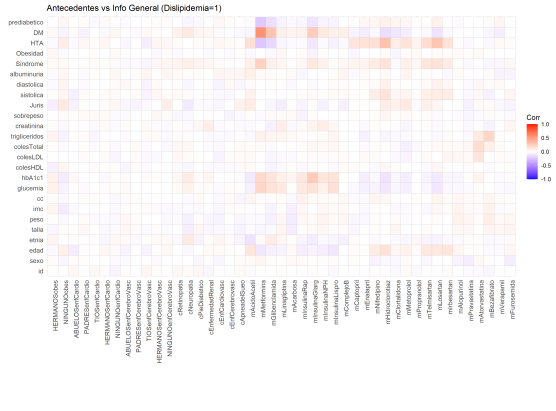
A continuación se agregan las imagenes de los análisis de correlación donde en cada enfermedad se seleccionan los que la presentan como población total.



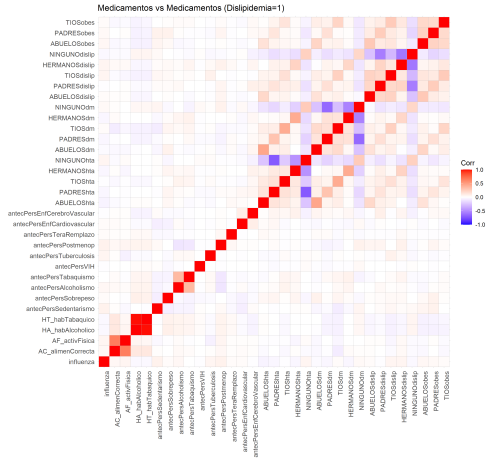
(a) Info General vs Info General



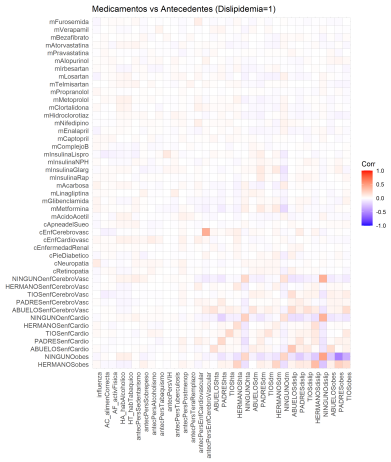
(b) Medicamentos vs Info General



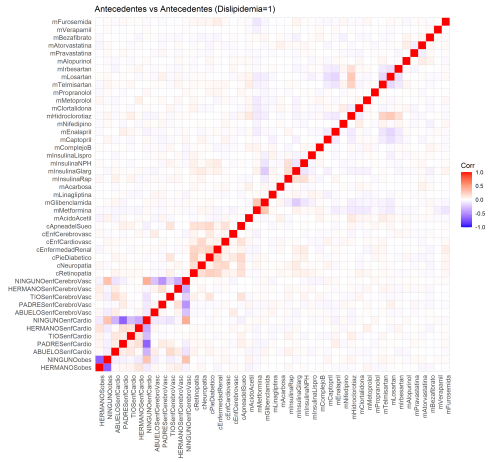
(c) Antecedentes vs Info General



(d) Medicamentos vs Medicamentos

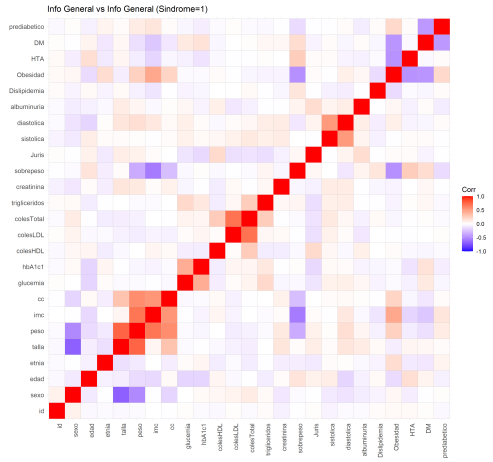


(e) Medicamentos vs Antecedentes

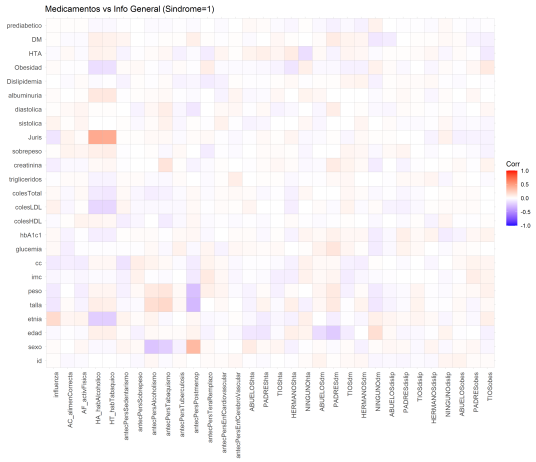


(f) Antecedentes vs Antecedentes

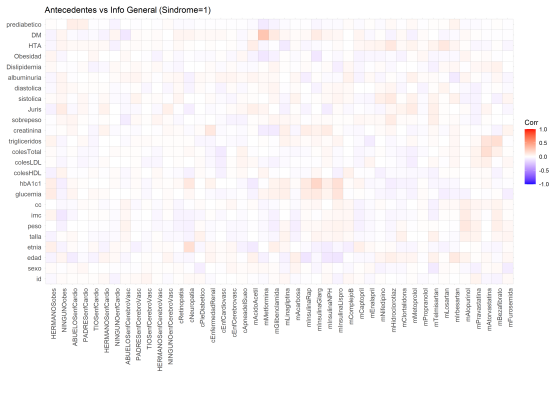
Figura 7.1: Matrices de correlación para pacientes con Dislipidemia



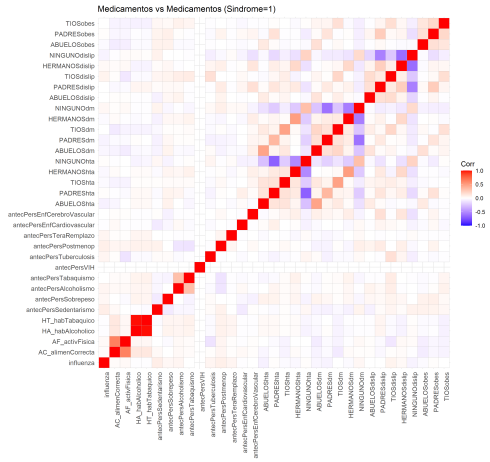
(a) Info General vs Info General



(b) Medicamentos vs Info General

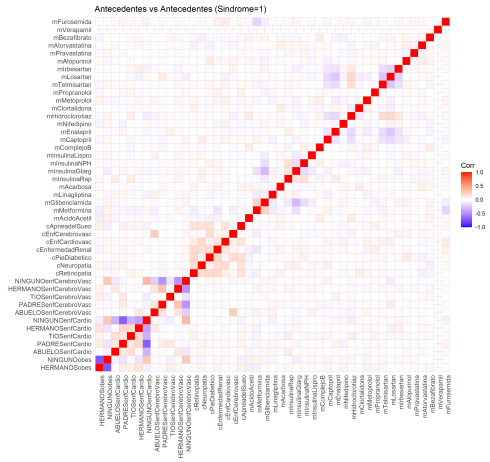


(c) Antecedentes vs Info General



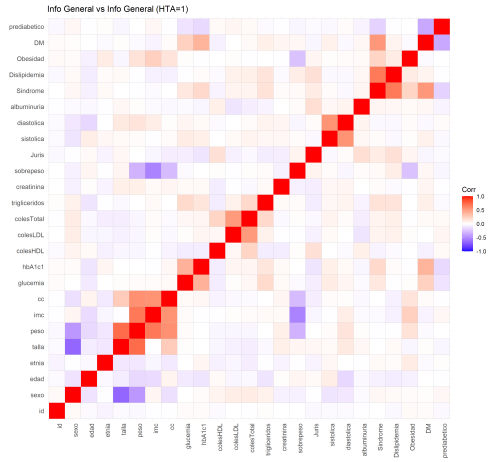
(d) Medicamentos vs Medicamentos

(e) Medicamentos vs Antecedentes

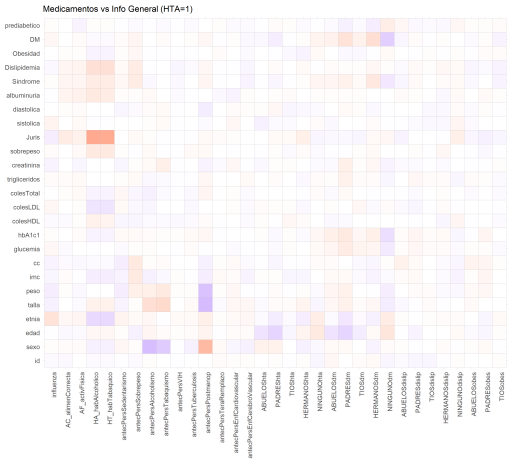


(f) Antecedentes vs Antecedentes

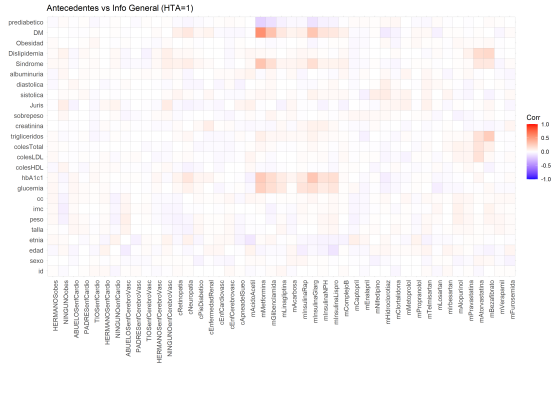
Figura 7.2: Matrices de correlación para pacientes con Síndrome metabólico



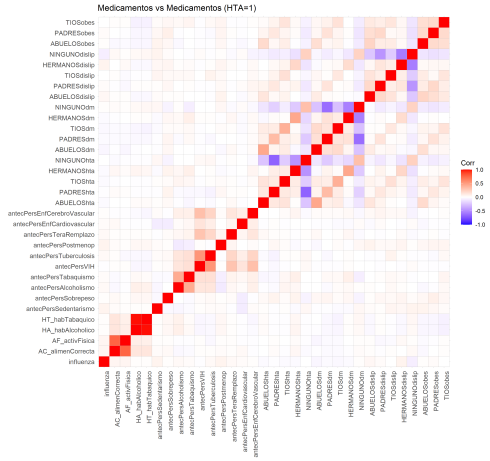
(a) Info General vs Info General



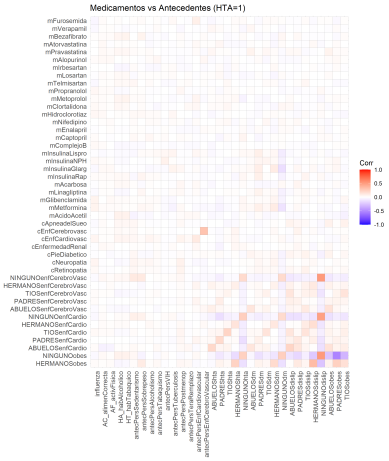
(b) Medicamentos vs Info General



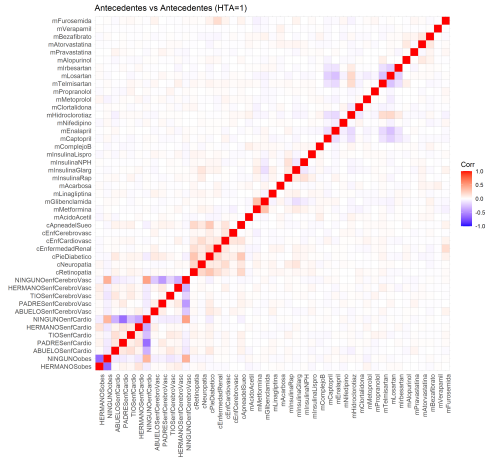
(c) Antecedentes vs Info General



(d) Medicamentos vs Medicamentos

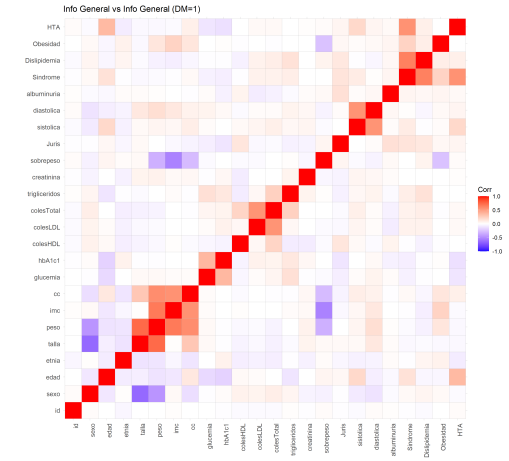


(e) Medicamentos vs Antecedentes

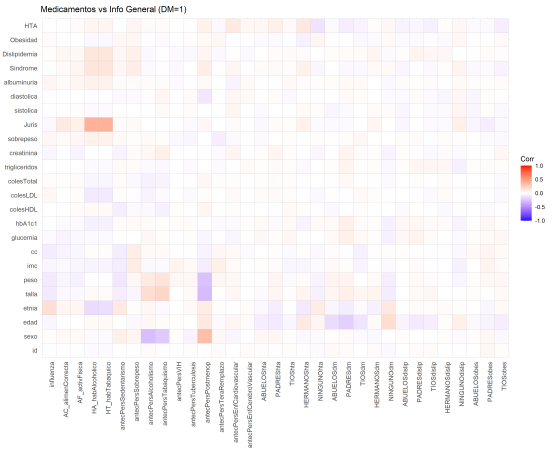


(f) Antecedentes vs Antecedentes

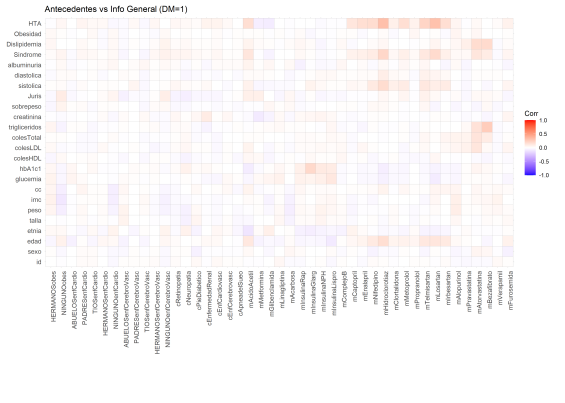
Figura 7.3: Matrices de correlación para pacientes con HTA



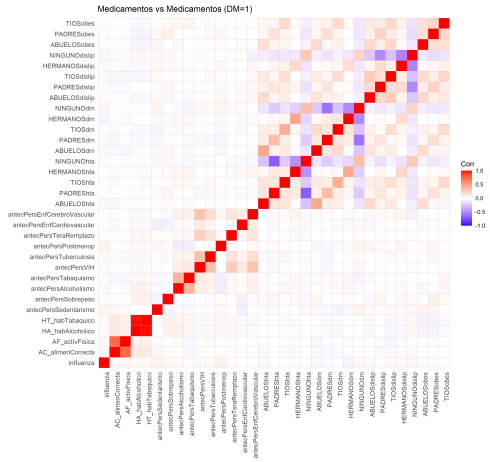
(a) Info General vs Info General



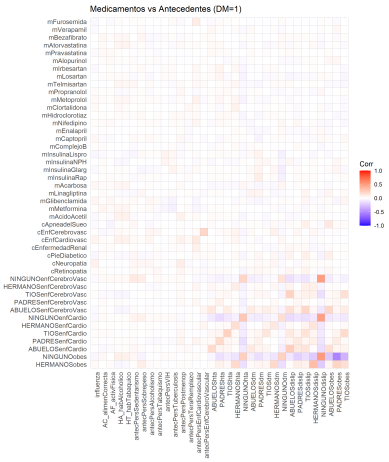
(b) Medicamentos vs Info General



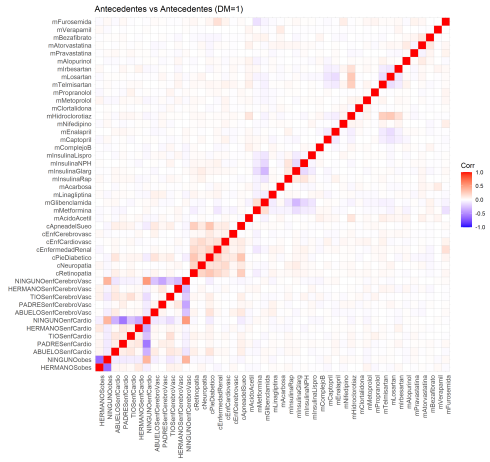
(c) Antecedentes vs Info General



(d) Medicamentos vs Medicamentos

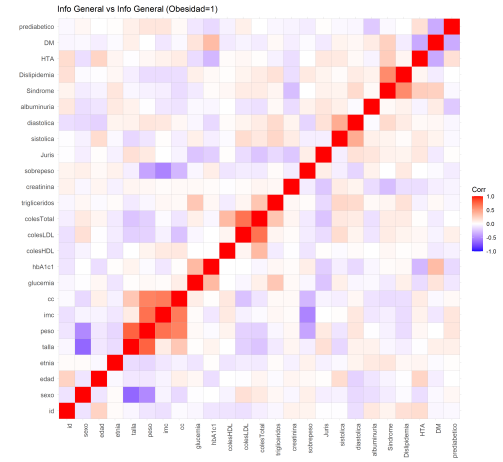


(e) Medicamentos vs Antecedentes

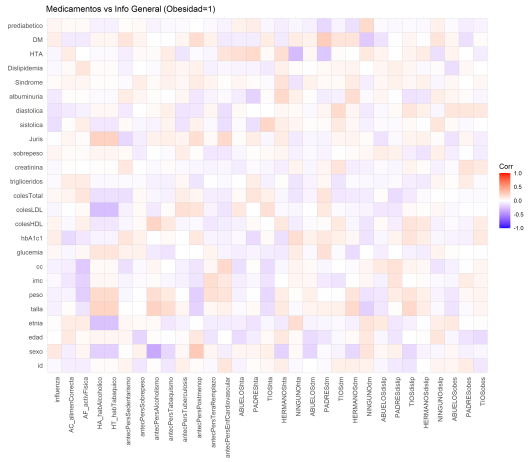


(f) Antecedentes vs Antecedentes

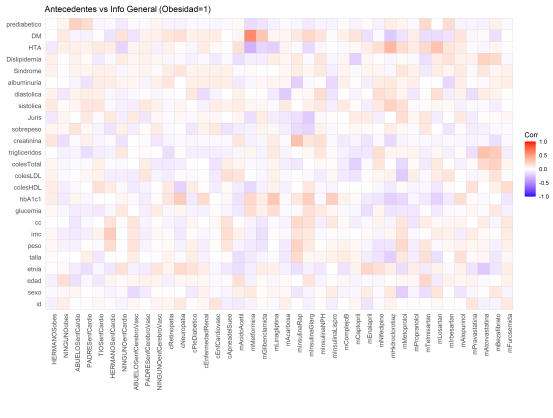
Figura 7.4: Matrices de correlación para pacientes con Diabetes



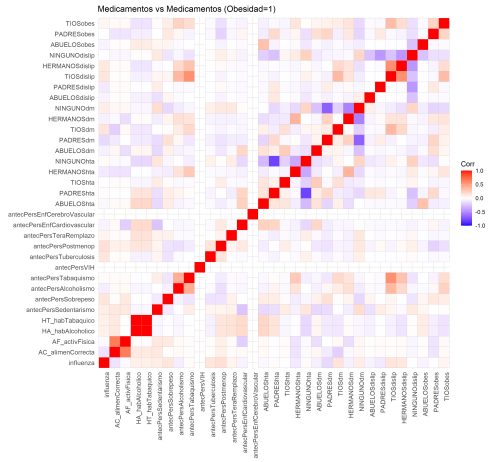
(a) Info General vs Info General



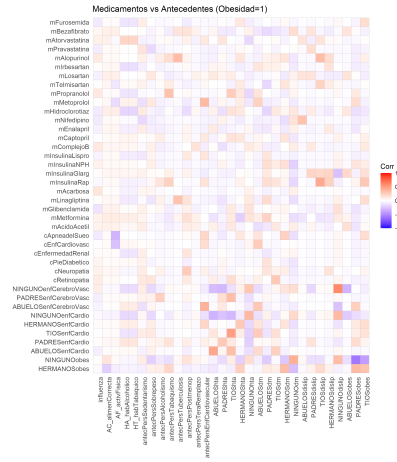
(b) Medicamentos vs Info General



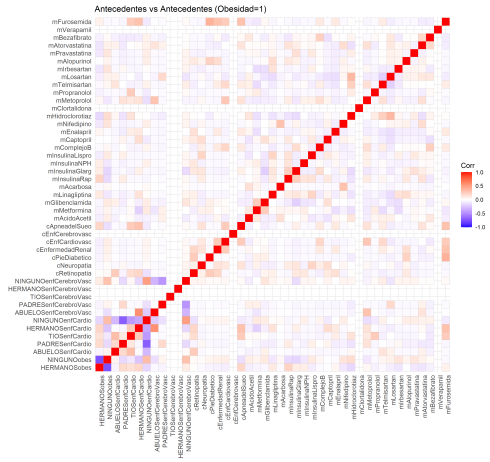
(c) Antecedentes vs Info General



(d) Medicamentos vs Medicamentos



(e) Medicamentos vs Antecedentes



(f) Antecedentes vs Antecedentes

Figura 7.5: Matrices de correlación para pacientes con Obesidad

Bibliografía

1. World Health Organization. *Obesidad y sobrepeso* World Health Organization. https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight (2025).
2. De las Personas Adultas Mayores, I. N. *En México, 80 % de las muertes de todas las edades corresponde a enfermedades no transmisibles* gob.mx. https://www.gob.mx/inapam/articulos/en-mexico-80-de-las-muertes-de-todas-las-edades-corresponde-a-enfermedades-no-transmisibles (2025).
3. De la Salud, O. P. *Enfermedades no transmisibles* OPS/OMS. https://www.paho.org/es/temas/enfermedades-no-transmisibles (2025).
4. Adab, P., Haroon, S., O'Hara, M. E. & Jordan, R. E. Comorbidities and COVID-19. *The BMJ* **377** (2022).
5. Djaharuddin, I. *et al.* Comorbidities and mortality in COVID-19 patients. *Gaceta sanitaria* **35,** S530-S532 (2021).
6. Russell, C. D., Lone, N. I. & Baillie, J. K. Comorbidities, Multimorbidity and COVID-19. *Nature Medicine* **29,** 334-343 (2023).
7. Rufín-Gómez, L. Á., Martínez-Morejón, A., Rufín-Bergado, A. M. & Méndez-Martínez, J. Síndrome metabólico, un factor de riesgo en pacientes de COVID-19. *Redalyc.* https://www.redalyc.org/journal/3782/378277400012/html/ (2025) (2022).
8. Wang, H. H., Lee, D. K., Liu, M., Portincasa, P. & Wang, D. Q.-H. Novel Insights into the Pathogenesis and Management of the Metabolic Syndrome. Inglés. *Pediatric Gastroenterology, Hepatology & Nutrition* **23,** 189 (2020).
9. Tablero de Control de Enfermedades Crónicas. *Tablero de Control de Enfermedades Crónicas* https://www.tablerocronicassic-sinba.com/TableroSIC/SIC-Detalle?entidad=-1&jurisdiccion=-1&municipio=-1&unidadsalud=-1 (2025).
10. Instituto Nacional de Estadística y Geografía (INEGI). *Estadísticas de Defunciones Registradas (EDR) de enero a junio de 2023 (preliminar)* Comunicado de prensa 26/24 (INEGI, Aguascalientes, México, Enero de 2024). https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2024/EDR/EDR2023_En-Jn.pdf.
11. National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). *¿Qué es la diabetes?* National Institutes of Health. Accedido en 2024. Diciembre de 2016. https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/que-es.
12. Fang, M., Wang, D., Coresh, J. & Selvin, E. Undiagnosed Diabetes in US Adults: Prevalence and Trends. *Diabetes Care* **45,** 1994-2002 (2022).

13. Badran, M., Morsy, R., Soliman, H. & Elnimr, T. Assessment of trace elements levels in patients with type 2 diabetes using multivariate statistical analysis. *Journal of Trace Elements in Medicine and Biology* **33,** 114-119 (2016).

14. Campos-Nonato, I. *et al.* Prevalencia, tratamiento y control de la hipertensión arterial en adultos mexicanos: resultados de la Ensanut 2022. *Salud pública de México* **65,** s169-s180 (2023).

15. Kishore, J., Gupta, N., Kohli, C. & Kumar, N. Prevalence of Hypertension and Determination of its Risk Factors in Rural Delhi. *International Journal of Hypertension* **2016,** 1-6 (2016).

16. Pavía-López, A. A. *et al.* Guía de práctica clínica mexicana para el diagnóstico y tratamiento de las dislipidemias y enfermedad cardiovascular aterosclerótica. *Archivos de cardiología de México* **92,** 1-62 (2022).

17. Xi, Y. *et al.* Prevalence of Dyslipidemia and Associated Risk Factors among Adults Aged $\geq$ 35 Years in Northern China: a Cross-Sectional Study. *BMC Public Health* **20,** 1-9 (2020).

18. GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry* **9,** 137-150 (2022).

19. Tang, R., Li, Q., Zhang, J. *et al.* Effects of BMI, dyslipidemia, and their interaction on hypertension: a cross-sectional study. *BMC Cardiovascular Disorders* **22,** 1-10 (2022).

20. Cao, R. Y., Zheng, H., Redfearn, D. & Yang, J. FNDC5: A Novel Player in Metabolism and Metabolic Syndrome. *Biochimie* **158,** 111-116 (2019).

21. Ahima, R. S. en *Metabolic Syndrome: a Comprehensive Textbook* 3-14 (Springer, 2024).

22. Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. *Multivariate Data Analysis* 8th (Cengage Learning, 2019).

23. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis* (John Wiley & Sons, 2003).

24. Jobson, J. D. *Applied Multivariate Data Analysis: Volume II: Categorical and Multivariate Methods* (Springer, 2012).

25. Montgomery, D. C., Peck, E. A. & Vining, G. G. *Introduction to Linear Regression Analysis* 6th (John Wiley & Sons, 2019).

26. Pearson, K. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A* **187,** 253-318 (1896).

27. Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Human Genetics* **7,** 179-188. `https://doi.org/10.1111/j.1469-1809.1936.tb02137.x` (1936).

28. McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition* (John Wiley & Sons, 2004).

29. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd (Springer, 2009).

30. Rao, C. R. *Linear Statistical Inference and Its Applications* 2nd (John Wiley & Sons, 2009).

31. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* 4th (Springer, 2002).

32. Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* 3rd (John Wiley & Sons, 2013).

33. Agresti, A. *Foundations of Linear and Generalized Linear Models* (John Wiley & Sons, 2015).