CIMAT Centro de Investigación en Matemáticas, A.C.

ÍNDICES DE CLASIFICACIÓN PARA CONDICIONES DIABÉTICAS BASADO EN ANÁLISIS DE DATOS DE LA PRUEBA ORAL DE TOLERANCIA A LA GLUCOSA

T E S I S

Que para obtener el grado de

Doctora en Ciencias

con Orientación en

Ciencias de la Computación

Presenta:

Paola Vargas Bernal

Directores de Tesis:

Dr. Miguel Ángel Moreles Vázquez

Dr. Joaquín Peña Acevedo

Autorización de la versión final

Guanajuato, Gto., 11 de Julio de 2022

Agradecimientos

Quiero agradecer:

- A Dios que es mi fortaleza y me demuestra que en Él todo lo puedo.
- A mis asesores el Dr. Miguel Ángel Moreles Vázquez y al Dr. Joaquín Peña Acevedo por continuar trabajando conmigo en el Doctorado, por sus enseñanzas, su apoyo incondicional y motivicación continua, a su vez por su entrega y dedicación para la realización de este trabajo de investigación.
- A mis sinodales la Dra. Adriana Monroy por recibirme en el Hospital y acompañarme en el proceso de investigación, a el Dr. Johan Van Horebeek, el Dr. Rogelio Ramos y el Dr. Marcos Capistrán por su seguimiento y retroalimentación al trabajo doctoral. Además, por aceptar revisar mi trabajo y tener el espacio para platicar sobre el mismo.
- Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por su financiamiento en la realización de mis estudios y de mi estancia en México.
- Al Centro de Investigación en Matemáticas, A.C. (CIMAT) por seguir siendo mi lugar de aprendizaje y crecimiento formativo a lo largo de mis estudios de posgrado.
- A mis padres Alfonso Vargas y Cecilia Bernal, y a mi hermana Fernanda Vargas por estar presente, animarme a lograr mis sueños aún en medio de las adversidades que se presentaron en el camino, por su sacrificio y paciencia para culminar con éxito mi doctorado.
- Y a mis amigos que conocí durante mi estancia en Guanajuato, en especial a aquellos que se convirtieron en mi familia y en mi red de apoyo.

Índice general

Introducción **7**

1. La Prueba Oral de Tolerancia a la Glucosa **13**
1.1. Protocolo clínico 13
1.2. Condición diabética según el criterio de la ADA 14
1.3. Perfiles de glucosa durante la OGTT 16
1.4. Población bajo estudio 17

2. Clasificación de condición diabética con la concentración máxima de glucosa y la tasa media de eliminación de glucosa, una prueba de concepto **19**
2.1. Modelo de Ackerman de la cinética del sistema glucosa-insulina 19
2.2. Estimación bayesiana de parámetros dados los datos de la OGTT 21
2.2.1. Perfiles de Glucosa 23
2.3. Clasificación con Máquinas de Soporte Vectorial 24
2.3.1. Clasificación SVM con un conjunto de entrenamiento 27
2.3.2. Población de prueba 27
2.4. Clasificación de pacientes disglicémicos 28
2.5. Conclusiones de la prueba de concepto 28

3. Metodología para la clasificación de condición diabética e índice de identificabilidad **31**
3.1. Distribuciones a priori 31
3.2. Estimador puntual 33
3.3. Validez del modelo de Ackerman 34
3.4. Clasificación SVM lineal 35
3.5. Índice de identificabilidad práctica 36
3.5.1. Identificabilidad 36

3.5.2. Construcción del índice de identificabilidad 40

4. Aplicación a la población bajo estudio 43
4.1. Estimación de parámetros y criterio de validez 43
4.2. Clasificación SVM 44
4.3. Índice de identificabilidad 47

5. Exploración de distintos métodos y propuestas en algunas etapas de la metodología 51
5.1. Estimación de parámetros 51
5.2. Distribuciones a priori 55
5.3. Estimadores puntuales 57
5.4. Criterio de validez del modelo de Ackerman 58
5.5. Resultados con las propuestas de distribuciones a priori y estimadores . . . 62
5.5.1. Criterio de validez 62
5.5.2. Clasificación lineal 63
5.6. Comentarios 65

6. Clasificación no lineal y regresión logística 67
6.1. Máquinas de soporte vectorial con fronteras de decisión no lineales 67
6.1.1. Resultados de SVM no lineal 69
6.2. Regresión logística 73
6.2.1. Resultados de regresión logística 74
6.3. Índice de identificabilidad en los distintos clasificadores 78

7. Conclusiones y trabajo futuro 83

Introducción

La diabetes mellitus es una enfermedad que se presenta cuando el páncreas no secreta la insulina necesaria o la insulina que se secreta no es usada de forma eficaz. Más del 95 % de personas que padecen diabetes presentan la tipo 2, en ésta el organismo usa de forma ineficaz la insulina y se ve relacionada con el exceso de peso y la inactividad física. Los síntomas se parecen a los de diabetes tipo 1, aunque son menos intensos, es por esto que puede llegar a ser diagnosticada años después de presentarse los primeros síntomas cuando ya empiezan a surgir complicaciones [30].

Esta enfermedad ha tenido un gran aumento en el número de personas que la padecen pasando de 108 millones de personas en 1980 a 422 millones de personas en 2014, aumentando con mayor rapidez en países de ingresos bajos y medios que en países de ingresos altos [31]. La diabetes es una de las principales causas de morbilidad, provoca miles de muertes y genera un gasto de miles de millones de dólares [25, 24]. En 2019, se le atribuyen 1.5 millones de muertes a esta enfermedad, siendo la novena causa más importante de muerte [30]. Para el año 2010, el gasto sanitario mundial se aproxima en 376000 millones de dólares y se pronostica que para el año 2030 aumente a 490.000 millones de dólares [25].

México ocupo el sexto puesto a nivel mundial en la prevalencia de la diabetes en 2015 [37] y es uno de los países con más afectados en la región de América [5]. El 7 % de la población mexicana tenía diabetes en 2006, aumentando a un 10.4 % para 2016 [24], y se proyecta que para 2050 la prevalencia de esta enfermedad en adultos puede estar entre el 13.7 - 22.5 %, afectando de 15 a 25 millones de personas [25]. México también presenta una de las tasas más altas de mortalidad a causa de la diabetes [24]. Se estima que el promedio del gasto anual para un paciente con diabetes en México esta entre 700 y 3200 dólares [25].

En México, una alta cantidad de adultos con diabetes no tienen un diagnóstico previo y el control glucémico de pacientes con diabetes es bajo. Aunque ha aumentado el porcentaje de la diabetes diagnosticada, pasando de 7.3 a 9.5 % de 2006 a 2016 y ha dis-

minuido la diabetes no diagnosticada pasando de 7.1 a 4.1 % en estos mismos años, aún así es necesario fortalecer la detección, diagnóstico oportuno y el control glucémico [5]. En [24] sugieren realizar campañas de cribado, es decir, detectar la enfermedad sin signos ni síntomas.

Debido a la situación actual de la diabetes y a las proyecciones que se hacen de esta enfermedad tanto en México como a nivel mundial es necesaria la prevención y la investigación en las diferentes áreas del conocimiento.

En recientes estudios, [3, 16, 17, 28, 22], se han interesado por la importancia de la prueba oral de tolerancia a la glucosa (OGTT) en el diagnóstico de la diabetes. La OGTT se realiza en la mañana después de tener al menos tres días de dieta sin restricciones y de actividad física usual. El sujeto debe tener un ayuno nocturno de 8 a 14 horas, en las cuales es permitido el consumo de agua. Factores externos que puedan alterar el resultado de la prueba deben ser registrados (p. ej. algún medicamento, inactividad, alguna infección). Se empieza recolectando una muestra de sangre en ayunas, después el sujeto ingiere 75 g de glucosa anhidra en 250-300 ml de agua durante 5 minutos y se recolectan muestras de sangre por las próximas 2 horas en intervalos de 30 minutos.

Con estos datos de concentración de glucosa en la sangre se han propuesto diversas técnicas de diagnóstico. Una de ellas es de acuerdo a los criterios de la Asociación Americana de Diabetes de 2021 [4]. También se han estudiado las curvas de glucosa (perfiles) obtenidas de la OGTT para extraer información que sirva en la detección de la enfermedad o del riesgo futuro de padecerla, usando métodos estadísticos para el ajuste de los perfiles.

El enfoque que se sigue en este trabajo es desde la modelación computacional, en donde se plantean modelos matemáticos que explican la cinética del sistema de glucosa-insulina durante pruebas de tolerancia a la glucosa tanto oral como intravenosa. En [32, 35] hacen una revisión de estos modelos. La modelación del sistema regulador de glucosa es un campo de investigación activo, en donde se agregan procesos más complejos. Como en [9] que extienden el modelo de compartimentos a un modelo de cuerpo entero que explica la regulación de la glucosa hepática y de la insulina pancréatica, o en [7] donde agregan la hormona de glucagón al sistema.

Este tipo de modelos están definidos en términos de ecuaciones diferenciales ordinarias, pueden ser complejos e involucrar varios parámetros desconocidos. La estimación de parámetros es un tema clásico y se han usado varias metodologías para resolver este

problema. En [35] se hace una revisión de algunas técnicas de estimación. Es importante la estimación robusta de los parámetros debido a que éstos se pueden utilizar para dar información extra al paciente, como por ejemplo, con el cálculo del índice de sensiblidad a la insulina que sirve para protocolos clínicos, ver [6].

Una buena alternativa es usar estimación bayesiana para encontrar los parámetros desconocidos, como lo hacen en [7] y [23]. El manejo de los datos con ruido, el poder obtener buenos estimadores puntuales y el obtener información adicional de los parámetros son algunas de las ventajas de este método. En [33] se hace una comparación de las técnicas de estimación de parámetros fisherianas, como la de máxima verosimilitud, y las bayesianas. Se encuentra que las técnicas fisherianas pueden tener problemas con la identificabilidad y que los métodos bayesiano son menos sensibles, en exactitud y precisión, en la estimación del índice de sensibilidad a la insulina en el modelo minimal.

También el uso de técnicas de aprendizaje para la clasificación de distintas enfermedades ha aumentado. Además, se siguen realizando investigaciones con ayuda de estas herramientas para el diagnóstico y predicción de enfermedades que son de gran ayuda en el área médica, como se menciona en [15]. En el artículo se comparan diferentes métodos y algoritmos de clasificación aplicado a datos en los que se quiere detectar diabetes de tipo 2.

El trabajo de tesis desarrollado es en este ámbito, se propone una herramienta de clasificación para determinar el estado diabético de un paciente. Se utilizan datos de la OGTT y el modelo de Ackerman et al [1] de la dinámica glucosa-insulina. El modelo es la ecuación diferencial de un oscilador armónico amortiguado. La clasificación se basa en dos parámetros inferidos del modelo, los cuales son la concentración máxima de glucosa y la tasa media de eliminación de glucosa. La propuesta surge del análisis de datos de la prueba OGTT recogidos en el Hospital General de México.

La herramienta de clasificación se construye usando una técnica básica de la inteligencia artificial, el método de máquinas de soporte vectorial con un kernel lineal, en el plano de los parámetros referidos.

La investigación se describe en dos fases, una prueba de concepto en una población de tamaño reducido, y la extensión a una población altamente heterogénea de tamaño significativo. El éxito de la prueba de concepto, ha sido publicado en [40]. En este trabajo se amplia la exposición, y se profundiza en las metodologías desarrolladas.

Se puede resumir los aspectos principales de la metodología propuesta de la siguiente manera:

1. Propuesta de distribuciones a priori en el método de estimación bayesiana.
2. Método de estimación del MAP basado en el método de kerneles de aproximación de densidades.
3. Propuesta de un criterio de validez del modelo de Ackerman.
4. Uso del clasificador SVM lineal.
5. Propuesta de índice de identificabilidad práctica.

En el primer punto, se muestra evidencia de la pertinencia de la estimación bayesiana y de una exploración de los datos, por lo cual se hace una propuesta de distribuciones a priori más informativas.

En el segundo aspecto, los métodos de estimación puntual son clásicos, y existen rutinas de software que pueden ser utilizadas. Con el fin de desarrollar una herramienta computacional que no dependa de la plataforma, se ha propuesto una estimación basada en análisis de datos de la población completa a partir del muestreo de la distribución a posteriori de los parámetros.

En la tercera propuesta, el criterio de validez del modelo de Ackerman nos indica si los datos de la OGTT de un paciente se pueden explicar a través de este modelo.

En el cuarto ítem es interesante que un kernel lineal es suficiente para la clasificación. Además, el grupo de estudio se divide en datos de entrenamiento y datos de prueba como se hace en las técnicas de aprendizaje.

En el último punto, la identificabilidad es una propiedad relevante en la estimación de parámetros. La propuesta de identificabilidad no es sobre los parámetros estimados, lo cual es el enfoque clásico. La identificabilidad consiste en mostrar que existe un mapeo inyectivo entre los datos y los parámetros [27]. Al procesar los datos de OGTT de un paciente, la metodología propone una clasificación binaria en pacientes sanos y pacientes disglicémicos (diabéticos y pre-diabéticos) con base en los parámetros. La pregunta de interés es si esta clasificación es identificable. La propuesta es un índice de identicabilidad práctica [27, 26] que se construye con la distribución marginal de los parámetros e indica

que tanto se puede confiar en la predicción de la clasificación de un dato.

Es importante notar que esta investigación no se ha realizado sobre un grupo controlado. Los datos de la OGTT utilizados en este trabajo fueron recolectados en el Hospital General de México como parte de un proyecto de investigación. Fueron proporcionados datos del año 2016 hasta el 2019. Los datos suministrados son de 1911 sujetos que no están relacionados entre sí. Algunos sujetos han tenido seguimiento a lo largo de los años y tienen registrada más de una muestra. El rango de edad de las mujeres está entre 17 y 80 años y el índice de masa corporal está entre 15.4 y 56.9 kg/m^2. El rango de edad de los hombres está entre 18 y 79 años y el índice de masa corporal está entre 14.97 y 57.7 kg/m^2.

La estructura del trabajo es la siguiente: en el *capítulo uno* se hace una descripción breve sobre la diabetes, la herramienta que se usa para su diagnóstico, la clasificación de las condiciones diabeticas y los perfiles de la curva de glucosa. Además, se describe el conjunto de datos que nos compartieron para el desarrollo de la investigación. En el *capítulo dos* se presenta la prueba de concepto de la metodología con una pequeña muestra de datos. En el *capítulo tres* se proponen las extensiones de la metodología, como la propuesta de distribuciones a priori, un método para encontrar un estimador puntual y el criterio de validación del modelo de Ackerman. También, se presenta un índice de identificabilidad práctico construido a partir del muestreo de la distribución a posteriori de cada sujeto. En el *capítulo cuatro* se aplica la metodología presentada a la población de estudio y se muestran los resultados obtenidos en cada etapa. En el *capítulo cinco* se muestra las diferentes propuestas que se hacen en algunas etapas de la metodología, la exploración y comparación de diferentes métodos y propuestas para justificar la selección hecha en cada etapa de la metodología a partir de los resultados obtenidos. En el *capítulo seis* se hace una revisión de otros clasificadores para determinar el peso que estos tienen en el resultado. En el *capítulo siete* se presentan las conclusiones finales y el trabajo futuro que se puede realizar.

Capítulo 1

La Prueba Oral de Tolerancia a la Glucosa

La diabetes es una enfermedad crónica en la que se presenta un desorden metabólico caracterizado por el aumento de glucosa en la sangre debido a que el páncreas no produce la insulina necesaria o la insulina que se produce no se utiliza de forma eficiente. Cuando la enfermedad no se controla bien puede causar lesiones graves en el corazón, en los vasos sanguineos, en los ojos, en los riñones y en los nervios, y aumentar el riesgo de muerte prematura. Esta enfermedad ha aumentado progresivamente tanto en el número de personas que la padecen como en su prevalencia en los últimos años. Es un problema de salud pública y una de las enfermedades no transmisibles que los líderes mundiales buscan mitigar [31].

De esta manera, se presenta la prueba oral de tolerancia a la glucosa (OGTT) usada para el diagnóstico de diabetes y prediabetes, los criterios según la Asociación Americana de Diabetes (ADA) para interpretar los resultados de la prueba, y la forma en que se pueden clasificar las curvas de OGTT de acuerdo a la concentración de glucosa en los distintos tiempos.

1.1. Protocolo clínico

Una herramienta para el diagnóstico de diabetes y prediabetes es la prueba oral de tolerancia a la glucosa (OGTT), la cual se realiza en la mañana después de tener al menos tres días de dieta sin restricciones y de actividad física usual. El sujeto debe tener un ayuno nocturno de 8 a 14 horas, en las cuales es permitido el consumo de agua. Factores externos que puedan alterar el resultado de la prueba deben ser registrados (p. ej. algún

medicamento, inactividad, alguna infección).

Se empieza recolectando una muestra de sangre en ayunas, después el sujeto ingiere 75 g de glucosa anhidra en 250-300 ml de agua durante 5 minutos y se recolectan muestras de sangre por las próximas 2 horas en intervalos de 30 minutos.

A menos que se pueda determinar la concentración de glucosa inmediatamente, las muestras de sangre se deben recolectar en un tubo que contiene fluoruro de sodio (6 mg por ml de sangre) y centrifugarse inmediatamente para separar el plasma. El plasma se debe congelar hasta que se vaya a estimar la concentración de glucosa [42]. La glucosa en plasma se encuentra con el método de la glucosa oxidasa.

Con la información que se obtiene de esta prueba en el tiempo 0 y 120 se diagnostica diabetes y prediabetes. A continuación se presenta los criterios que se usan para esta clasificación por parte de la Asociación Americana de Diabetes.

1.2. Condición diabética según el criterio de la ADA

La Diabetes Mellitus tipo 2 (**T2DM**), que se debe a una disminución progresiva de la adecuada secreción de insulina por las células β, se diagnóstica según los criterios de la Asociación Americana de Diabetes de 2021 [4] cuando:

- Glucosa plasmática en ayuno (FPG) $\geq$ 126 mg/dl (7 mmol/l), o
- Glucosa plasmática a las 2h durante la OGTT $\geq$ 200 mg/dl (11.1 mmol/l), o
- HbA1c $\geq$ 6.5 % (48 mmol/mol)

La hemoglobina glicosilada (HbA1c) muestra el promedio de glucosa plasmática de las últimas ocho a doce semanas. No requiere que la persona esté en ayunas o alguna otra preparación especial, por lo que se puede realizar en cualquier momento del día [43]. Esta prueba debe realizarse usando un método certificado por la National Glycohemoglobin Standarization Program (NGSP) y estandarizado por el ensayo de referencia Diabetes Control and Complications Trial (DCCT) [4].

En los sujetos en donde no hay concordancia entre los valores de HbA1c y los de glucosa, los resultados de FPG y glucosa plasmática a las 2h durante la OGTT son más precisos

para su diagnóstico.

La "prediabetes" es la condición que tienen aquellas personas con niveles de glucosa mayores a los valores normales, pero su valor no alcanza a cumplir los criterios para considerarse diabetes. Los criterios que definen prediabetes según la Asociación Americana de Diabetes [4] son:

- **IFG:** glucosa plasmática en ayuno entre 100 mg/dl (5.6 mmol/l) a 125 mg/dl (6.9 mmol/l), o
- **IGT:** glucosa plasmática a las 2h durante la OGTT entre 140 mg/dl (7.8 mmol/l) a 199 mg/dl (11 mmol/l), o
- HbA1c entre 5.7 - 6.4 % (39 - 47 mmol/mol)

donde IFG es alteración de glucosa en ayuno e IGT es intolerancia a los carbohidratos.

En la tabla 1.1 se unifican los criterios y las etiquetas que debe presentar un paciente para considerarse sano, diabético o pre-diabético.

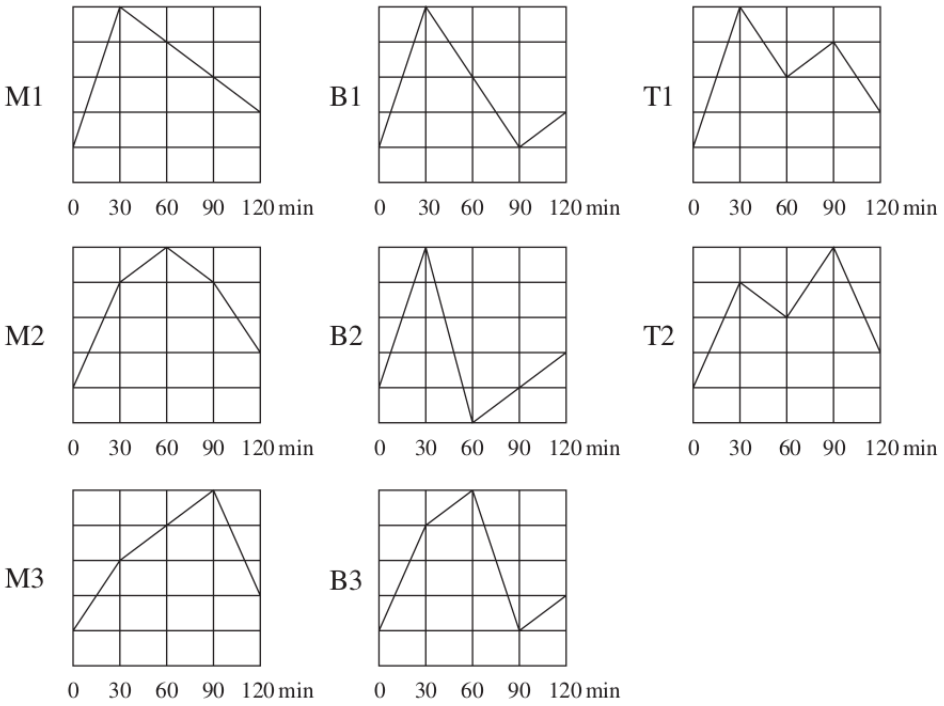
Etiqueta	Condición		Plasma venoso (mg/dl)
NGT	Sano/Normoglicémico	Glucosa en ayuno Glucosa a las 2h	menor a 100 y menor a 140
DMT2	Diabetes Mellitus	Glucosa en ayunas Glucosa a las 2h	mayor o igual a 126 o mayor o igual a 200
IFG	Alteración de glucosa en ayuno	Glucosa en ayunas Glucosa a las 2h	mayor o igual a 100 y menor a 126 y menor a 140
IGT	Intolerancia a los carbohidratos	Glucosa en ayunas Glucosa a las 2h	menor a 100 y mayor o igual a 140 y menor a 200

Tabla 1.1: Valores de diagnóstico de diabetes mellitus y prediabetes según la ADA

También, con las mediciones de la OGTT se obtiene la curva de glucosa, las cuales se clasifican en tres grupos según la forma que presentan y una condición entre las mediciones en el tiempo 90 y 120.

1.3. Perfiles de glucosa durante la OGTT

Las curvas de glucosa durante la OGTT se pueden clasificar según la forma en monofásicas (M), bifásicas (B) y trifásicas (T). Estos tipos de curva se dividen en ocho subclases como se muestra en la figura 1.1.



Tomada: Kanauchi, M., Kimura, K., Kanauchi, K., & Saito, Y. (2005).

Figura 1.1: Clasificación de las curvas de glucosa según su forma durante la OGTT.

Las curvas monofásicas se caracterizan por un aumento de la glucosa plasmática después de la carga oral de glucosa hasta alcanzar el valor máximo a los 30 minutos (M1), o a los 60 minutos (M2), o a los 90 minutos (M3) y comienzan a disminuir hasta el minuto 120.

Las curvas bifásicas se caracterizan por un aumento de la glucosa plasmática después de la carga oral de glucosa hasta alcanzar un valor máximo a los 30 minutos o a los 60 minutos, luego empiezan a disminuir hasta llegar al punto más bajo, a los 60 minutos o

a los 90 minutos y comienzan a aumentar hasta el minuto 120. Se puede tener el pico a los 30 minutos y el punto más bajo a los 90 minutos (B1), o el pico a los 30 minutos y el punto más bajo a los 60 minutos (B2), o el pico a los 60 minutos y el punto más bajo a los 90 minutos (B3).

Las curvas trifásicas se caracterizan por una disminución de glucosa plásmatica entre los 30 minutos y los 60 minutos, luego un aumento entre los 60 minutos y los 90 minutos, y por último una disminución entre los 90 minutos y los 120 minutos. Estas curvas tendrán dos picos, se puede tener que el pico más alto es a los 30 minutos (T1), o que el pico más alto es a los 90 minutos (T2). [39, 21]

Además de las características ya mencionadas, también se necesita que el valor absoluto de la diferencia entre la concentración de glucosa a los 90 minutos con la concentración de glucosa a los 120 minutos, debe ser mayor a 4.5 mg/dl (0.25 mmol/l). [39]

Si las curvas no cumplen ninguna de las condiciones mencionadas se consideran curvas "Sin clasificar".

1.4. Población bajo estudio

La ciencia de datos está creciendo muy rápido. Actualmente se cuenta con grandes cantidades de datos y múltiples herramientas computacionales eficaces que sirven para extraer información de los datos y responder preguntas a cerca ellos [12]. En [8], Donoho propone una "Greater Data Science", la cual consta de las siguientes seis actividades:

- Exploración y preparación de datos: esta actividad es necesaria e importante antes del análisis. Sirve para encontrar propopiedades de los datos, características inesperadas y anomalías en los datos.
- Representación y transformación de datos: el uso de base de datos modernas. Además, en algunas ocasiones se debe aplicar una transformación adecuada a los datos de forma que se reestructuren y sean más reveladores para su estudio.
- Cómputo con datos: Uso de lenguajes de computación para el análisis y procesamiento de datos, también uso de ambientes múltiples y de cómputo de alto rendimiento.
- Visualización y presentación de datos: desarrollar visualizaciones para presentar los resultados de manera estática y/o dinámica.

- Modelación de datos: Se usa el modelado generativo, en el cual se propone un modelo que podría generar los datos, y la modelización predictiva que usa el aprendizaje automático.
- Ciencia de análisis de datos.

En el presente trabajo se tienen en cuenta estas seis divisiones para el análisis de nuestros datos. A continuación se hace una breve presentación de los datos con los que se va a trabajar:

Los datos de la OGTT utilizados fueron recolectados en el Hospital General de México como parte de un proyecto de investigación. Para este estudio, fueron proporcionados datos del año 2016 hasta el 2019. El estudio fue aprobado por el Comité de Ética del Hospital General de México, y se obtuvo un consentimiento por escrito por parte de los pacientes. El rango de edad de las mujeres está entre 17 y 80 años y el índice de masa corporal está entre 15.4 y 56.9 kg/m^2, y el rango de edad de los hombres está entre 18 y 79 años y el índice de masa corporal está entre 14.97 y 57.7 kg/m^2. Los datos que se utilizan para este estudio no corresponden a un grupo controlado.

La información antropométrica y clínica está disponible, pero nuestra metodología se basa únicamente en los datos de la OGTT. Los datos sumistrados son de **1911** sujetos que no están relacionados e informaron que no tomaban medicamentos que pudieran afectar su tolerancia a la glucosa, su sensibilidad a la insulina, o su secresión de insulina. Algunos sujetos han tenido un seguimiento a lo largo de los años y por esto tienen registrada más de una muestra.

Los voluntarios se practicaron la prueba OGTT después de ocho horas de ayuno nocturno. Ellos ingirieron una solución de dextrosa de 75 g (Dextrosol, Hycel, Mexico). La clasificación según la ADA de la primera muestra de los 1911 sujetos es de la siguiente manera: 147 IFG, 319 IGT, 156 IFG-IGT, 205 T2DM y 1084 NGT.

Capítulo 2

Clasificación de condición diabética con la concentración máxima de glucosa y la tasa media de eliminación de glucosa, una prueba de concepto

En este capítulo se presenta una prueba de diagnóstico, con un pequeño conjunto de datos, de nuestra metodología, que consiste en seleccionar un modelo matemático que explica la interacción de glucosa-insulina, estimar los parámetros del modelo a partir de los datos de la OGTT y construir un clasificador con los parámetros que presentan una relación con el criterio de la ADA. Por ende, se van a mostrar las distintas etapas de este método, las herramientas que se usaron en cada paso, y los resultados obtenidos con esta población. Este capítulo está basado en Vargas et al [40].

2.1. Modelo de Ackerman de la cinética del sistema glucosa-insulina

Se han propuesto distintos modelos mátematicos que describen la cinética del sistema de glucosa-insulina durante la OGTT [32, 35], de los cuales se puede extraer información que sirve para diagnóstico de la enfermedad. A continuación se presenta un modelo básico que se utilizará para encontrar los parámetros desconocidos a partir de los datos de OGTT.

Sea $G(t)$ la concentración de glucosa plasmática en la sangre y $H(t)$ la concentración neta de hormonas que influyen en los niveles de glucosa en la sangre en el tiempo t. Para las condiciones de la OGTT, la insulina se considera predominante y $H(t)$ es esencialmente su concentración.

Después del ayuno, las concentraciones del paciente tienden a estabilizarse en sus valores basales G_0 y H_0. Por lo que se estudian las pequeñas desviaciones

$$g(t) = G(t) - G_0, \quad h(t) = H(t) - H_0.$$

Un modelo simple derivado en Ackerman et al [1], es

$$\begin{aligned} \dot{g} &= -m_1 g - m_2 h + J \\ \dot{h} &= -m_3 h + m_4 g \end{aligned}$$

donde m_1, m_2, m_3, m_4, son constantes positivas.

El significado de los parámetros es el siguiente:

- m_1: tasa de eliminación de la glucosa independiente de la insulina,
- m_2: tasa de eliminación de la glucosa que depende de la insulina,
- m_3: tasa de eliminación de la insulina independiente de la glucosa,
- m_4: tasa de liberación de la insulina debido a la glucosa.

Después de algún tiempo la carga de glucosa $J(t)$ es absorbida en el sistema y $J(t) \equiv 0$. Se puede eliminar h del sistema y obtener la siguiente ecuación diferencial de segundo orden,

$$\ddot{g} + 2\alpha\dot{g} + \omega_0^2 g = 0 \tag{2.1}$$

donde

$$\alpha = \frac{1}{2}(m_1 + m_3),$$

y

$$\omega_0 = \sqrt{m_1 m_3 + m_2 m_4},$$

es la frecuencia natural del sistema.

Una interpretación del sistema de glucosa-insulina es la de un oscilador armónico amortiguado. Por consiguiente, se supone que

$$\alpha^2 - \omega_0^2 < 0.$$

Por lo tanto, la solución del modelo de Ackerman [1] después de que la carga de glucosa ha sido absorbida, es la solución general de (2.1), que está dada por

$$g(t) = Ae^{-\alpha t}\cos(\omega t - \delta), \tag{2.2}$$

donde

$$\omega = \sqrt{\omega_0^2 - \alpha^2}.$$

A la gráfica de la función $G(t)$, se le llama la curva de OGTT (perfil).

2.2. Estimación bayesiana de parámetros dados los datos de la OGTT

Una vez escogido el modelo con el que se va a trabajar, se pasa a resolver un problema inverso para encontrar los parámetros desconocidos a partir de la información de la OGTT de cada paciente. Esto se plantea como un problema de optimización. Enseguida se muestra el desarrollo del método que se usa para resolver este problema y ejemplos del ajuste de algunas curvas de OGTT.

Los pacientes son numerados de $j = 1$ a $j = 80$. Se define $g^j(t) = G^j(t) - G^j(0)$ para $t = 0, 30, 60, 90, 120$.

El problema de interés es: Dado los datos de la OGTT del paciente j: g_0^j, g_{30}^j, g_{60}^j, g_{90}^j, g_{120}^j, estimar los parámetros

$$\mathbf{u}_j = (A_j, \alpha_j, \omega_j, \delta_j)^t, \tag{2.3}$$

dado que

$$g^j(t) = A_j e^{-\alpha_j t}\cos(\omega_j t - \delta_j). \tag{2.4}$$

Para resolver este problema de optimización se selecciona un enfoque bayesiano. Se calculan los estimadores puntuales máximo a posteriori (MAP) y media condicional (CM). Por su diseño, el enfoque bayesiano mide la robustez y la fiabilidad de los estimadores, como se ve en [20, 38]. Se darán más detalles a continuación.

En la estimación bayesiana todas las variables son aleatorias, por lo que se considera el modelo

$$\mathbf{y} = \mathcal{G}(\mathbf{u}) + \eta$$

donde $\mathbf{u}$ es el parámetro a estimar, $\mathbf{y}$ son los datos, η es el ruido y $\mathcal{G}$ es el operador de observación.

Una característica importante de la estimación bayesiana es proponer una función de densidad de probabilidad a *priori*, π_0, para el parámetro $\mathbf{u}$. Esta a priori engloba todo lo que se conoce acerca de $\mathbf{u}$. En esencia, se trata de un problema de modelización.

Se supone que el ruido es conocido con densidad ρ e independiente de $\mathbf{u}$. En consecuencia, la densidad condicional $\pi^{\mathbf{y}}(\mathbf{u}) \equiv \pi(\mathbf{u}|\mathbf{y})$, conocida como la a *posteriori*, está dada por la fórmula de Bayes como

$$\pi^{\mathbf{y}}(\mathbf{u}) \propto \rho(\mathbf{y} - \mathcal{G}(\mathbf{u}))\pi_0(\mathbf{u}).$$

El objetivo de la estimación bayesiana es determinar la a posteriori. A partir de ésta, se pueden obtener estimadores puntuales. Específicamente, la media condicional y el estimador máximo a posteriori.

El estimador CM esta dado por la integral

$$\mathbf{u}_{CM} = \int \mathbf{u}\pi^{\mathbf{y}}(\mathbf{u})d\mathbf{u},$$

mientras que para el estiamdor MAP hay que resolver un problema de optimización,

$$\mathbf{u}_{MAP} = \operatorname{argmax} \rho(\mathbf{y} - \mathcal{G}(\mathbf{u}))\pi_0(\mathbf{u}).$$

En nuestro problema, el ruido se toma gaussiano con media cero y desviación estándar γ. De la fórmula de Bayes se tiene,

$$\pi^{y}(\mathbf{u}) \propto \exp\left(-\frac{1}{\gamma^2}\left|y - \mathcal{G}(\mathbf{u})\right|^2\right)\pi_0(\mathbf{u}).$$

La a priori puede influir artificialmente en la determinación de la a posteriori, lo que no es deseable, ya que puede sesgar el resultado. Por esa razón, se prefiere que las a priori no sean informativas en la prueba de concepto, es decir, densidades uniformes. Posteriormente, con lo observado en esta prueba se busca la estructura subyacente de las a priori.

Las densidades uniformes que se utilizan son las siguientes:

- $A \sim U[0.5\, g_m, 2.5\, g_M + 150]$
- $\alpha \sim U[0, 0.1]$
- $\omega \sim U[0, 0.15]$
- $\delta \sim U[-2\pi, 2\pi]$

donde g_m, g_M son respectivamente el mínimo y el máximo de los valores absolutos de la diferencia de los datos de la concentración de glucosa en los tiempos $t = 30, 60, 90, 120$ menos G_0.

También, se permite grandes errores de observación con el modelo gaussiano con desviación estandar $\gamma = 5$, el cual es un valor típico en la literatura.

En la práctica, se construye una muestra de la a posteriori. Se utiliza el módulo de python `emcee`, un muestreador grupal con invariancia afin de Markov Chain Monte Carlo (MCMC) [10]. Luego, para cada parámetro del paciente, se determina una distribución a posteriori y se obtiene su estimador MAP y CM.

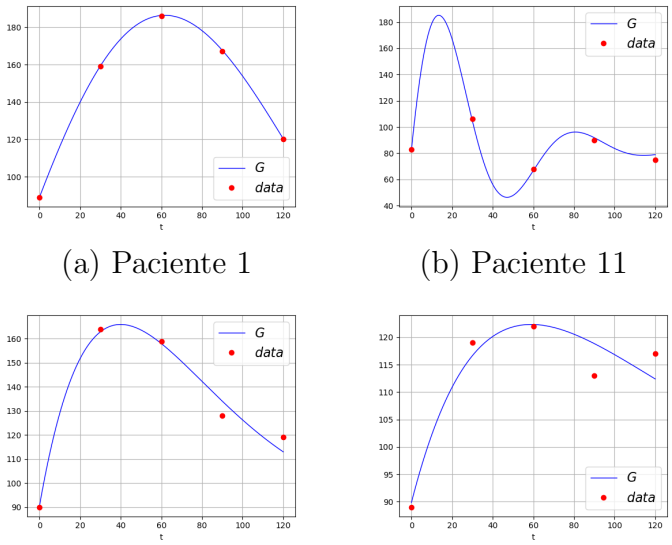
Una ventaja del enfoque bayesiano es que la incertidumbre de las estimaciones puntuales se cuantifica fácilmente por medio de las densidades marginales de la a posteriori. Por ejemplo, dada la densidad a posteriori $\pi^y(\mathbf{u})$, la marginal para α se encuentra integrando sobre las otras variables,

$$\pi^y(\alpha) = \int\int\int \pi^y(A, \alpha, \omega, \delta) dA d\omega d\delta.$$

2.2.1. Perfiles de Glucosa

Para cada paciente, los parámetros (2.3) se estiman para construir las funciones (2.4). Al graficar $g^j(t)$ junto con los datos de los pacientes se puede observar el ajuste del modelo.

Por ejemplo, a continuación se grafican las curvas de OGTT para algunos pacientes con las diferentes condiciones que menciona la ADA. Los perfiles de los pacientes sanos se muestran en la figura 2.1, mientras que los perfiles de los pacientes disglicémicos (diabéticos y pre-diabéticos) se muestran en la figura 2.2.



(a) Paciente 1 (b) Paciente 11

(c) Paciente 27 (d) Paciente 41

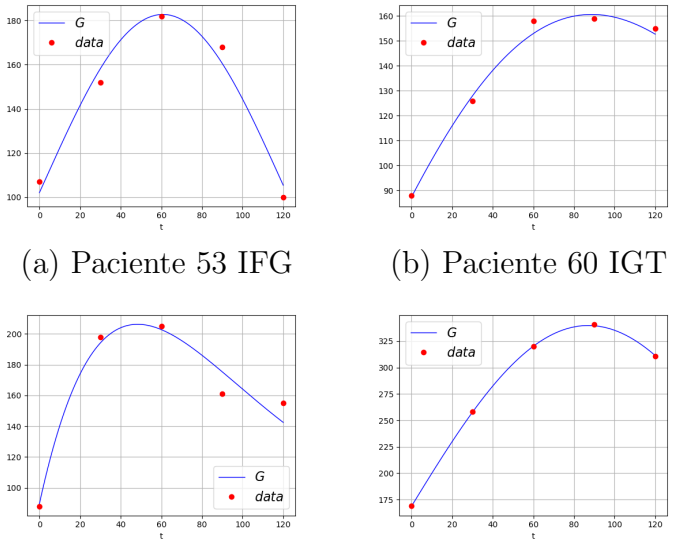
Figura 2.1: Curvas de OGTT de algunos pacientes sanos

Las curvas de OGTT en las figuras muestran que nuestra metodología de estimación de parámetros se ajusta a los datos con precisión.

2.3. Clasificación con Máquinas de Soporte Vectorial

Del proceso de estimación, se explora entre los distintos parámetros si existe alguna relación entre ellos. Para la clasificación, se consideran los parámetros:

- A, *concentración máxima de glucosa*,
- α, *promedio de la tasa de eliminación de glucosa*.



(a) Paciente 53 IFG (b) Paciente 60 IGT

(c) Paciente 70 IGT (d) Paciente 79 T2DM

Figura 2.2: Curvas de OGTT de algunos pacientes disglicémicos

Para el proceso de clasificación se determina un conjunto de entrenamiento. Los sujetos se dividen en dos clases, sanos (+1) y disglicémicos (-1). Para cada paciente j, los parámetros A_j y α_j se seleccionan para la clasificación binaria en el plano A-α.

La clasificación se hace usando el método de *máquina de soporte vectorial (SVM)* [19]. Con este método se busca construir un hiperplano separador con máximo margen entre los conjuntos de datos de cada clase, cuando estos son separable. El margen es la mínima distancia perpendicular de las observaciones al hiperplano. En la figura 2.3 se muestra un ejemplo.

Como se puede tener conjuntos de datos no separables se introducen variables de holgura, de modo que con el hiperplano separador se tenga mayor robustez en la observación individual y mejor clasificación en la mayoría de las observaciones.

El hiperplano se encuentra resolviendo el siguiente problema de optimización con restricciones

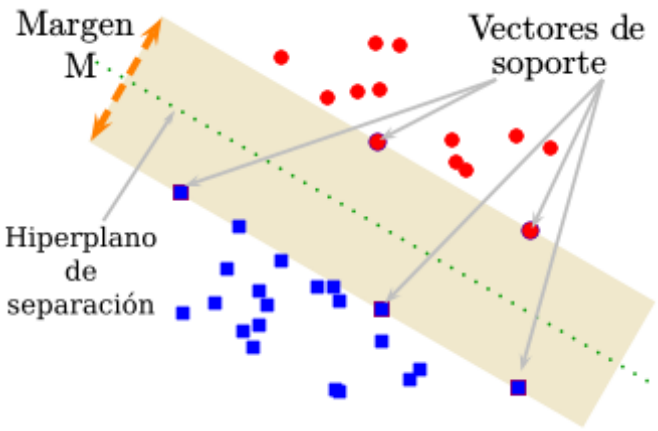


Figura 2.3: Hiperplano de máximo margen

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n} \quad M \tag{2.5}$$

$$\text{sujeto a} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \tag{2.6}$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \varepsilon_i) \tag{2.7}$$

$$\varepsilon_i \geq 0, \quad \sum_{i=1}^{n} \varepsilon_i \leq C, \tag{2.8}$$

donde y_i es la etiqueta de la clase del dato i, $x_{i1}, x_{i2}, \ldots, x_{ip}$ son las características del dato i, C es un parámetro de sintonización no negativo que determina la severidad con la que se permite a los datos violar el margen o el hiperplano, M es la anchura del margen (se busca que este valor sea tan grande como sea posible) y $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son la variable de holgura que permite que cada dato individualmente este en el lado incorrecto del margen o del hiperplano.

La selección del hiperplano depende únicamente de los datos que caen sobre el margen o las que lo violan y se conocen como los vectores de soporte, ver figura 2.3. Los datos que caen en el lado correcto del margen no afectan al clasificador de soporte vectorial.

Para clasificar una nueva observación, x_*, se determina el signo de la función

$$f(x^*) = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \cdots + \beta_p x_p^*, \tag{2.9}$$

para ver a que lado del hiperplano se encuentra.

2.3.1. Clasificación SVM con un conjunto de entrenamiento

Los datos seleccionados son de 80 mujeres sin parentesco. El rango de edades está entre los 18 a los 45 años. La muestra corresponde a los años 2016 y 2017. Las voluntarias se practicaron la prueba OGTT, y se clasifican según la ADA de la siguiente manera: 5 IFG, 14 IGT, 7 IFG-IGT, 3 T2DM y 51 NGT.

En la figura 2.4, se muestra la clasificación binaria en el plano A-α. La clasificación es exitosa para el 85 % de pacientes. Notese que un separador lineal es suficiente.

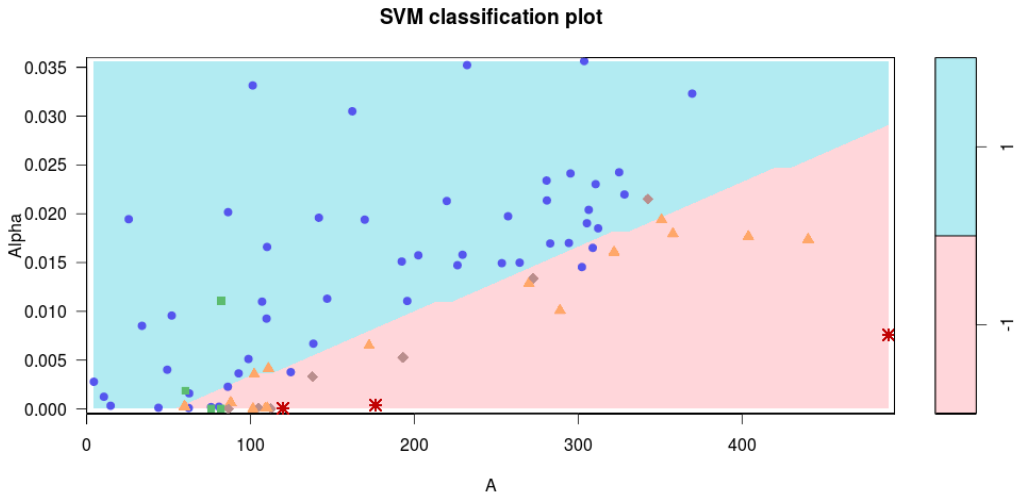


Figura 2.4: NGT (•), IFG (□), IGT (△), IFG-IGT (⋄), T2DM (∗).

2.3.2. Población de prueba

Una vez construida la línea de separación, se elige el siguiente conjunto de prueba independiente, de un tamaño comparable, para evaluar el rendimiento del clasificador.

Una dificultad que se adiciona es seleccionar una población mixta.

- 24 Hombres: 15 NGT, 3 IGT, 2 IFG-IGT, 4 T2DM,
- 33 Mujeres: 17 NGT, 1 IFG, 11 IGT, 2 IFG-IGT, 2 T2DM.

La clasificación es exitosa para el 91 % de hombres y el 87 % de mujeres. Ver figura 2.5.

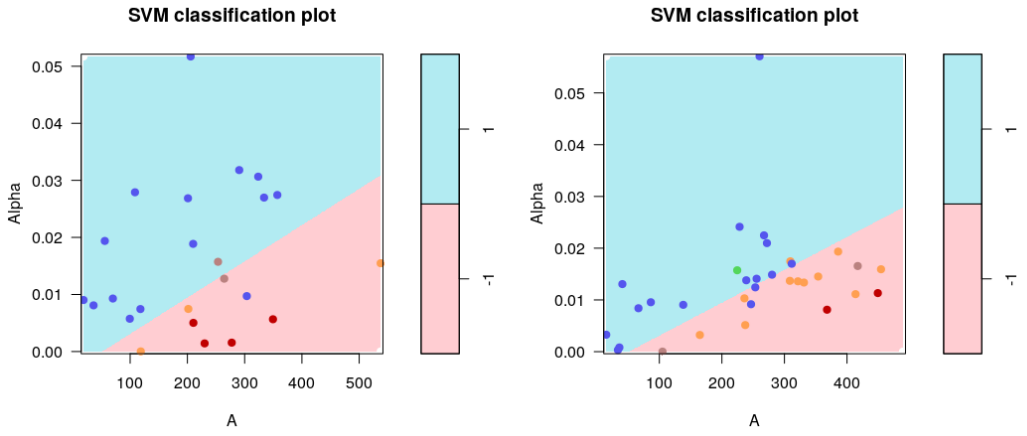


Figura 2.5: Conjunto de prueba: Hombres (izquierda), Mujeres (derecha).

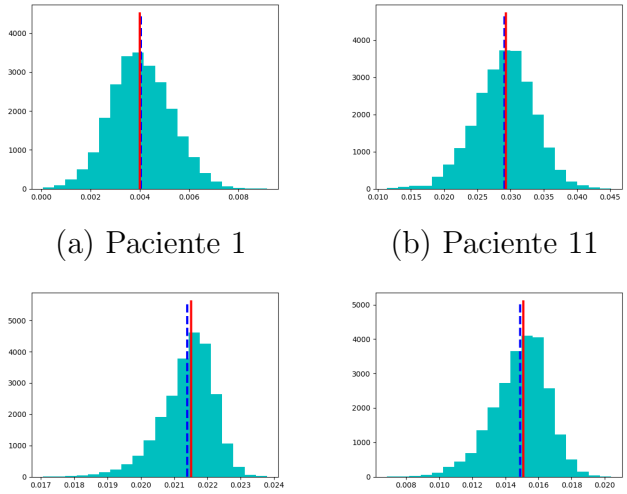
2.4. Clasificación de pacientes disglicémicos

Hay dos posibles errores de clasificación. Clasificar a un sujeto sano como disglicémicos, y a un sujeto disglicémicos como sano. Este último es el peor de los casos. Como se muestra en las figuras 2.4 y 2.5, los pacientes con condiciones de IGT, IFG-IGT y T2DM están la mayoría por debajo de la línea de separación. Por consiguiente, los pacientes disglicémicos en una gran cantidad de casos fueron clasificados correctamente.

2.5. Conclusiones de la prueba de concepto

Se menciona en [2] que el parámetro α es muy sensible a los errores en la concentración de glucosa. Su uso no ha sido recomendado para un criterio de diagnóstico. Sin embargo,

basados en los experimentos, el enfoque bayesiano sugiere lo contrario. Aunque se consideran grandes errores de observación con el modelo gaussiano, las densidades marginales obtenidas de la a posteriori para el parámetro α, son unimodales y los estimadores puntuales no son ambiguos. Estas características se ilustran en las figuras 2.6 y 2.7, donde se presenta una muestra de sujetos sanos y disglicémicos.



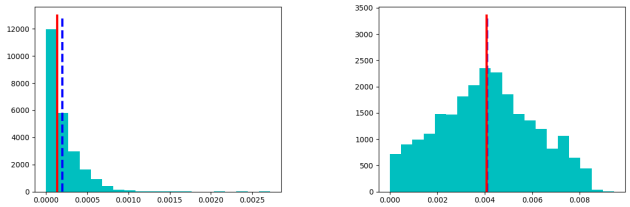
(a) Paciente 1 (b) Paciente 11

(c) Paciente 27 (d) Paciente 41

Figura 2.6: Densidades a posteriori marginales del parámetro α para pacientes sanos. MAP la línea sólida. CM la línea punteada.

Al examinar los resultados de la clasificación SVM, figura 2.4, se pueden plantear dos aplicaciones clínicas. La primera es que la concentración máxima de glucosa, A, y el promedio de la tasa de eliminación de glucosa, α, pueden ser considerados como índices del paciente con el potencial de ser utilizados para la clasificación de las diferentes condiciones que menciona la ADA. La segunda puede ser la detección temprana de pacientes propensos a padecer la enfermedad, debido a que se observa que existe una aparente transición en el *sentido de las agujas del reloj* desde los pacientes sanos hasta los pacientes con Diabetes Mellitus tipo 2.

Por otra parte, el tamaño de la muestra es adecuado para la aplicación y las conclusiones del método de máquina de soporte vectorial. También, el conjunto de entrenamiento



(a) Paciente 53 IFG (b) Paciente 60 IGT

(c) Paciente 70 IGT (d) Paciente 79 T2DM

Figura 2.7: Densidades a posteriori marginales del parámetro α para pacientes disglicémicos. MAP la línea sólida. CM la línea punteada.

fue seleccionado porque contenía todas las condiciones de diabetes y prediabetes que menciona la ADA y el tamaño apropiado para la clasificación. Para evitar que el género fuera un factor, se eligió una población completamente femenina en el conjunto de entrenamiento. Aunque, al ver los resultados del conjunto de prueba, se desprende que la variable género no juega ningún papel. Nuestros resultados se obtienen del uso directo de los datos de OGTT.

Capítulo 3

Metodología para la clasificación de condición diabética e índice de identificabilidad

Como la prueba de concepto realizada en el capítulo anterior fue exitosa, se procesa la base de datos completa descrita en la sección 1.4. A partir de lo observado con este nuevo conjunto de datos se proponen extensiones a la metodología. En particular, se proponen a prioris más informativas en la estimación de parámetros, el método de estimación puntual, y la manera de filtrar los datos que no tienen un buen ajuste.

Una pregunta natural en estimación de parámetros es su identificabilidad, esto es, si existe un mapeo inyectivo de parámetros a los datos, ver [27]. En nuestro estudio, el interés no solo es la estimación robusta de los parámetros, sino sobre la clasificación de la condición diabética. En consecuencia, se propone un índice de identificabilidad práctica sobre la clasificación de un sujeto, basado en la distribución marginal de los parámetros estimados.

3.1. Distribuciones a priori

Se parte de la solución general del modelo de Ackerman,

$$G(t) = G_0 + Ae^{-\alpha t}\cos(\omega t - \delta). \tag{3.1}$$

Se estiman los parámetros del nuevo conjunto de datos que tiene un mayor tamaño,

con las distribuciones a priori como uniformes y se selecciona el estimador MAP de cada uno de los datos. El conjunto de datos se divide en cuatro grupos, hombres con mediciones de glucosa e insulina, hombres con mediciones solo de glucosa, mujeres con mediciones de glucosa e insulina y mujeres con mediciones solo de glucosa. Después se generan histogramas de cada parámetro con el resultado del estimador seleccionado en cada uno de los cuatro grupos. El resultado obtenido se muestra en la figura 3.1.

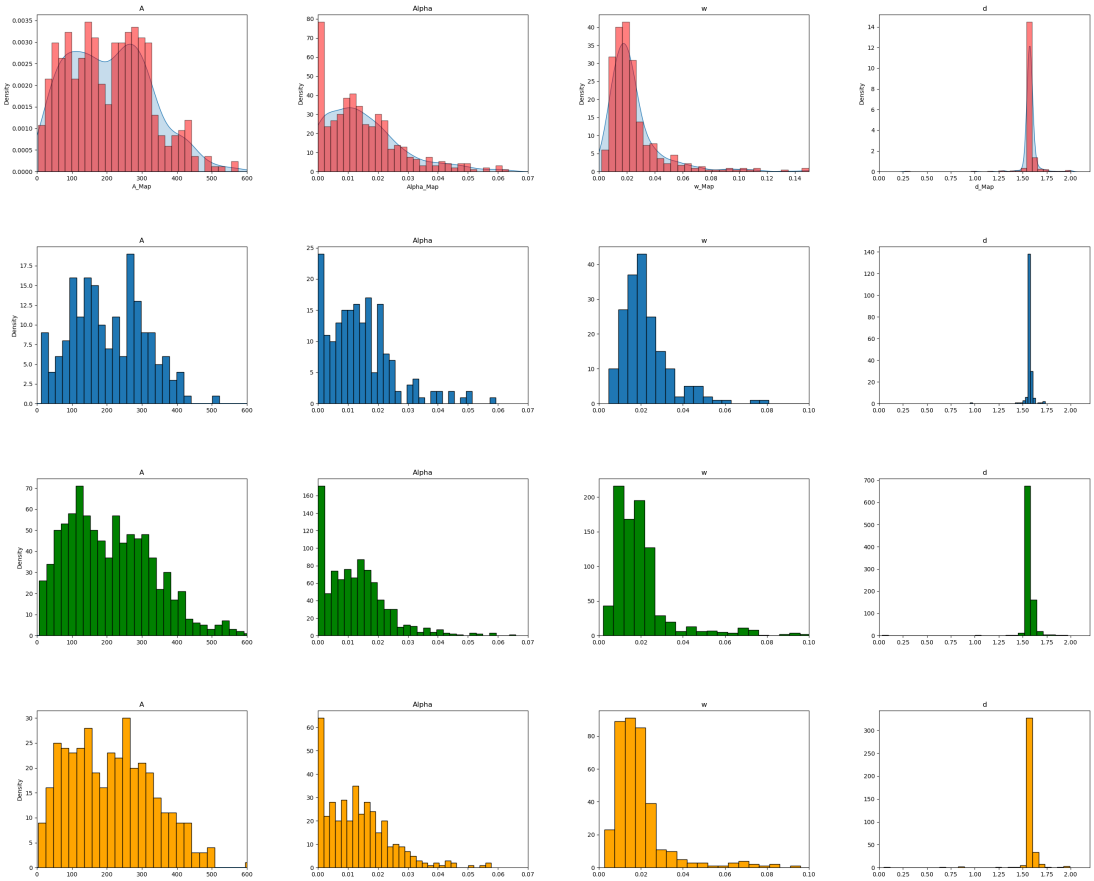


Figura 3.1: Histograma de los valores del MAP de cada parámetro estimado del grupo de hombres con glucosa e insulina (primero), hombres con solo glucosa (segundo), mujeres con glucosa e insulina (tercero) y mujeres con solo glucosa (cuarto)

Con lo observado en la figura 3.1 se proponen distribuciones a priori más informativas para cada una de los parámetros que se van a estimar y nos permite fijar un valor para el parámetro δ. Además, se ve que los parámetros tienen un comportamiento similar en los cuatro grupos, por lo que nos reitera que la variable del género no aporta información extra.

Como en los histogramas del parámetro δ se observa que la distribución de este parámetro sería una normal centrada en $\frac{\pi}{2}$ con una varianza muy pequeña. Es por esto, que se toma el valor de δ como $\frac{\pi}{2}$.

A partir de lo que se observa en estos histogramas se proponen distribuciones a priori más informativas para los tres parámetros que se van a estimar. Las densidades para estos parámetros son normales truncadas definidas de la siguiente manera:

- $A \sim N[200,150]$ truncada en (10,600)
- $\alpha \sim N[0.015,0.02]$ truncada en (0,0.08)
- $\omega \sim N[0.04,0.03]$ truncada en (0, 0.15).

Esta distribución a priori se llamará distribución TTT.

Después de estimar los parámetros con estimación bayesiana se obtiene una muestra de la distribución a posteriori de cada parámetro. A partir de ésta, se seleccionan estimadores puntuales como se menciona en el capítulo anterior. Se va a presentar el estimador que se usará para el desarrollo de los siguientes pasos de la metodología presentada.

3.2. Estimador puntual

Para la construción del clasificador se necesita un A y un α asociado a cada dato. Esta información se obtiene tomando un estimador puntual con la información obtenida en la estimación bayesiana. En el capítulo anterior se considero el estimador MAP, el cual se encuentra resolviendo un problema de optimización con la información de la cadena. También, dependiendo el paquete o la implementación del método MCMC que se use estas plantean alguna forma de encontrarlo y el estimador puede variar.

Es por esto que se propone un método para encontrar un estimador que solo depende de la información de la cadena y el cual se puede replicar con los distintos paquetes que se usen para muestrear. El método se describe a continuación:

1. Con la información de la cadena se construye el histograma de la distribución conjunta de los tres parámetros.

2. Usando el método de kernels, se aproxima al histograma una función de densidad, $h(x)$. Esta función representa la probabilidad conjunta.

3. Se evalua cada uno de los puntos de la cadena en la función h.

4. El punto de la cadena que obtiene el mayor valor en la evaluación de la función h será el estimador puntual que se va a utilizar. A este estimador se le llama MAP_cal.

Más adelante se verá que con la propuesta del estimador del item 4. en una gran cantidad de datos se obtiene un buen ajuste. Además, se observa un comportamiento interesante de este estimador en el plano A-α. Es por esto, que el MAP_cal es una buena alternativa para usar como estimador puntual en nuestra metodología.

Una vez se estiman los parámetros y se selecciona el estimador puntual que ajusta el modelo a cada dato, se construye el clasificador. En la prueba de concepto el clasificador se obtiene con todos los datos procesados, pero se observó que algunos datos no logran tener un buen ajuste con el modelo de Ackerman, por esta razón se plantea un criterio de validez para filtrar a los pacientes que tienen un buen ajuste o un buen comportamiento de la curva.

3.3. Validez del modelo de Ackerman

En esta sección se proponen criterios para determinar cuando los datos de la prueba OGTT de un paciente, son satisfactoriamente descritos por el modelo de Ackerman.

El modelo es físicamente análogo a un oscilador armónico con amortiguamiento. En esta aplicación, el periodo debe ser restringido para tener sentido biofísico. En término de los parámetros, se prescribe una cota superior de la frecuencia angular estimada (ω).

También se considera la clasificación de la forma de la curva para poder identificar si las tomas de glucosa se pueden clasificar dentro de un grupo.

Para la bondad de ajuste de los datos, se considera el error absoluto entre las mediciones de cada paciente y el valor de la glucosa con los parámetros del estimador seleccionado en

el modelo de Ackerman. Esto es

$$\text{Error}_{abs} = \frac{1}{5} \sum_{i \in \{0,30,60,90,120\}} |G_i - G_i^{Est}|,$$

donde G_i es la toma de glucosa en el tiempo i y G_i^{Est} es el valor de glucosa en el tiempo i usando los valores del estimador seleccionado en el modelo de Ackerman.

Se recuerda de [39], que el valor absoluto de la diferencia entre la concentración de glucosa a los 90 minutos con la concentración de glucosa a los 120 minutos, debe ser mayor a 4.5 mg/dl (0.25 mmol/l).

En consecuencia, se dice que el modelo de Ackerman es valido para describir los datos OGTT de un paciente, si $\omega < 0.09$ y se cumple alguna de las siguientes condiciones:

1. La curva se clasifica según su forma y $|G_{90} - G_{120}| > 4.5$ mg/dl; $\text{Error}_{abs} < \epsilon_1$.
2. La curva se clasifica según su forma y $|G_{90} - G_{120}| < 4.5$ mg/dl; $\text{Error}_{abs} < \epsilon_2$.
3. La curva se considera "Sin clasificar"; $\text{Error}_{abs} < \epsilon_3$.

donde $\epsilon_1 > \epsilon_2 > \epsilon_3$. Nótese que se pide un mejor ajuste de los datos en base a no satisfacerse condiciones biomédicas. En un capítulo posterior se muestra como se selecciona el valor de estos ϵ.

El valor $\omega < 0.09$ limita el número de oscilaciones del perfil de glucosa y corresponde a un periodo menor a 70 minutos aproximadamente.

Después de seleccionar los datos que cumplen el criterio de Ackerman, se construye el clasificador con la información de los parámetros A y α de cada dato. Para este paso del proceso se continua trabajando con el método de máquinas de soporte vectorial.

3.4. Clasificación SVM lineal

En la prueba de concepto se veía un comportamiento con respecto a las manecillas del reloj en el plano A-α de pacientes sanos, enfermos y con alguna alteración, y esto se ve más evidente cuando aumenta el número de datos. Es por esto, que seguir proponiendo el

uso de un separador lineal es natural.

Se exploró el método SVM con otro kerneles y con otros métodos de clasificación, y se compararon los resultados con el método de SVM lineal, lo cual se muestra en un capítulo posterior. Como el SVM lineal da buenos resultados se continua trabajando con este método.

3.5. Índice de identificabilidad práctica

La clasificación se va a complementar con un índice de identificabilidad práctica que nos da un porcentaje de que tan bien estaría clasificado un sujeto. Se muestra como se contruye este índice con las muestras obtenida de la distribución posterior de los parámetros de interés para la clasificación, A y α.

3.5.1. Identificabilidad

Supongamos que tenemos un sistema de ecuaciones diferenciales

$$\begin{aligned} \dot{x} &= f(x, p, u) \\ y &= h(x, p, u), \end{aligned}$$

donde x son las variables dependientes, y son las medidas, u los parámetros conocidos y p los parámetros desconocidos.

El sistema se va a decir que es **identificable** si los parámetros desconocidos p, pueden ser únicamente determinados por los parámetros conocidos u y las medidas y que se tienen del modelo. Es decir, si para cualquier parámetro admisible u y cualesquiera dos parámetros desconocidos p_1 y p_2 en el espacio de parámetros

$$y(u, p_1) = y(u, p_2) \Leftrightarrow p_1 = p_2. \tag{3.2}$$

En otro caso, el sistema se dice **inidentificable**.

Como este análisis depende de las características de la estructura del sistema, se le denominará **identificabilidad estructural**. Este enfoque supone que la estructura del modelo es precisa y no hay error de medición. Como estas suposiciones son poco realistas es necesario introducir una identificabilidad práctica que incluya aleatoriedad. Esto

se puede proponer por ejemplo mediante simulaciones de Monte Carlo o un análisis de identificabilidad basado en la sensibilidad [26, 27].

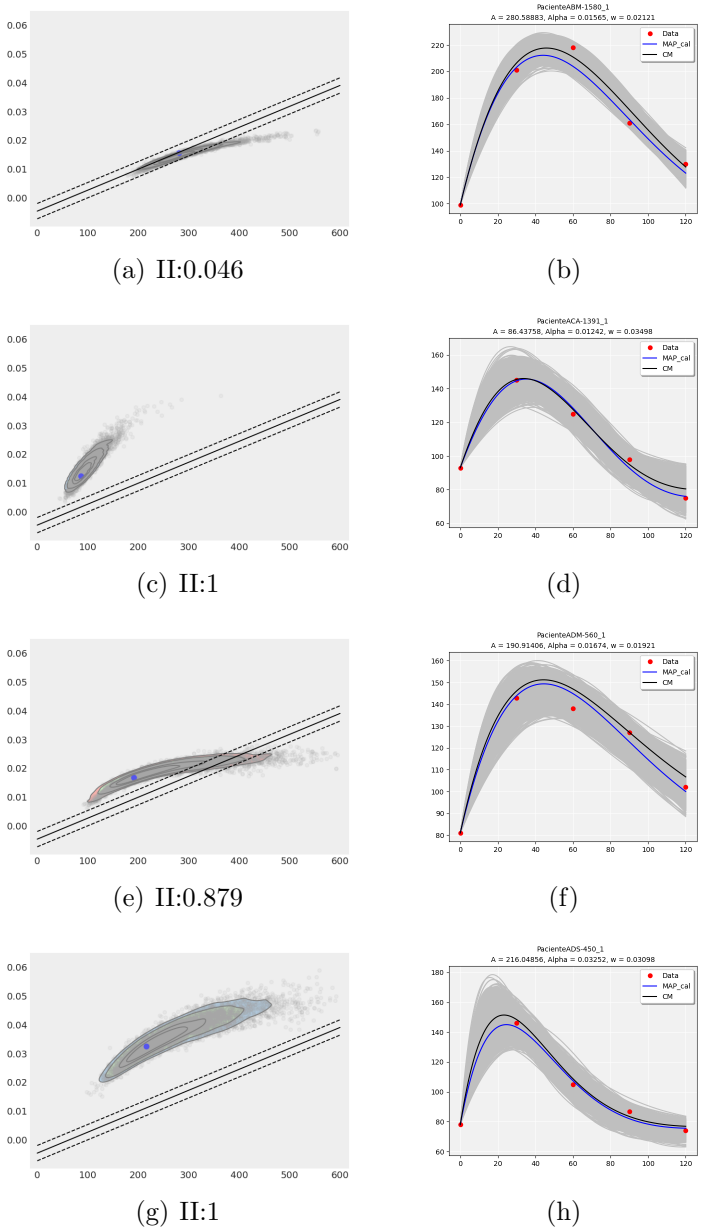
La inidentificabilidad teórica debe implicar la inidentificabilidad práctica. Se debe hacer tanto el análisis de identificabilidad estructural como el práctico para asegurar la fiabilidad de la estimación de los parámetros. Si los parámetros estimados tienen intervalos de confianza finitos, encontrados a partir de la matriz de información de Fisher (FIM), se puede decir que estos son estructural y prácticamente identificables [41].

Como en la construcción del clasificador los parámetros que se usan son A y α, nos interesa la identificabilidad de estos. Si se observa la distribución marginal de estos parámetros se ve que esta es unimodal, esbelta y en algunos casos la dispersión es alta, ver figura 3.2 y 3.3, pero para nuestro problema nos interesa el comportamiento de esta distribución en el plano A-α con el clasificador, debido a que se busca que la clasificación de los sujetos no sea variante, es decir, que si se tomará otro estimador se pueda obtener el mismo resultado de clasificación.

En las figuras de la izquierda de 3.2 y 3.3 se ven en gris las distintas parejas de A y α que se obtienen como resultado de la estimación bayesiana de un paciente, que corresponde con la zona gris de la gráfica de ajuste que se muestra a la mano derecha, y los contornos con probabilidad de 30 %, 50 %, 80 % y 90 % de la densidad posterior entre A y α. Hay sujetos con contornos que tienen baja dispersión y otros donde la dispersión es alta, pero al observar el comportamiento de estos contornos una gran parte de puntos tienen el mismo resultado de clasificación que el estimador MAP, por lo que la clasificación es consistente. También, estos contornos se dispersan en una dirección en donde el cambio de las condiciones diabéticas con respecto a las manecillas del reloj se mantendría si se cambia el estimador.

En la figura 3.2 se ven algunos ejemplos de pacientes sanos. En (a) se observa un paciente mal clasificado, en (c) el representante está cerca de la línea separadora pero los puntos más dispersos se alejan de la línea de división, en (e) una pequeña parte de puntos pasan la línea de separación y en (g) el representante esta más alejado de la línea separadora, los puntos más disperso van en la dirección de la línea divisoria y aún así los puntos de la cadena quedan en la zona correcta.

En la figura 3.3 se ven algunos ejemplos de pacientes disglicémicos. En (a) todos los puntos de la cadena quedan bien clasificados y los puntos más dispersos van en dirección



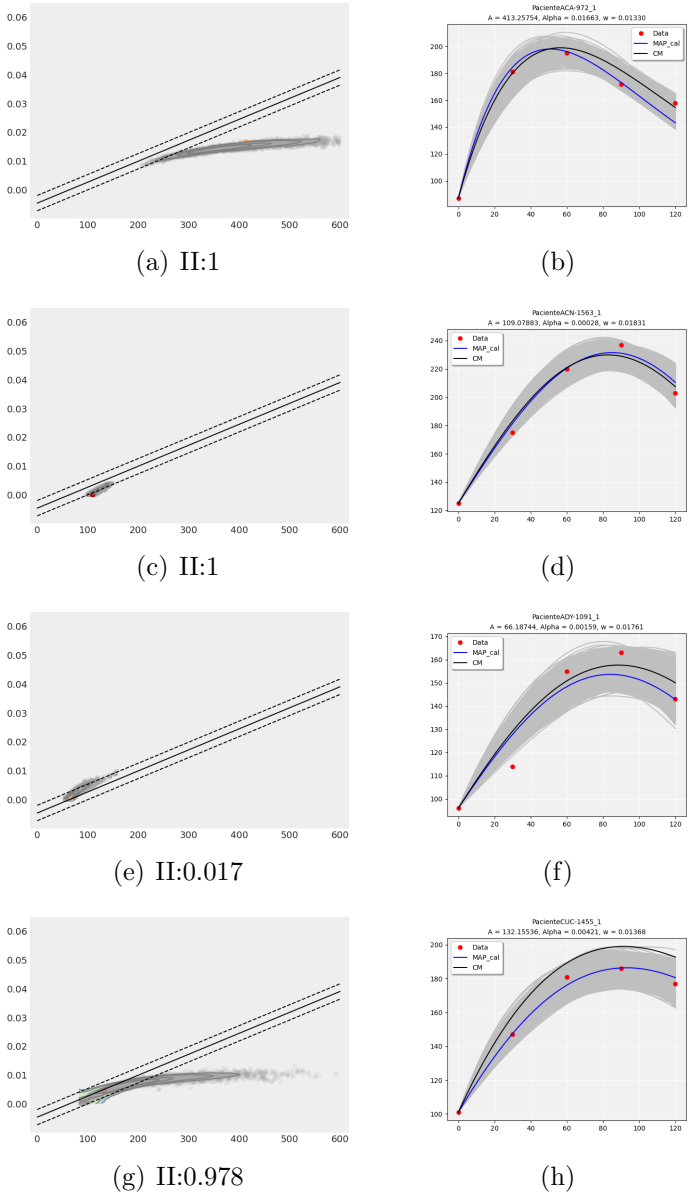
(a) II:0.046 (b)

(c) II:1 (d)

(e) II:0.879 (f)

(g) II:1 (h)

Figura 3.2: Índice de identificabilidad de pacientes sanos con su respectiva gráfica de ajuste.



(a) II:1 (b)

(c) II:1 (d)

(e) II:0.017 (f)

(g) II:0.978 (h)

Figura 3.3: Índice de identificabilidad de pacientes disglicémicos con su respectiva gráfica de ajuste.

contraria a la línea de división, en (c) el representante esta cerca de la línea separadora pero la dirección del contorno es paralela a la línea de separación, en (e) se observa un paciente mal clasificado y en (g) una pequeña parte de puntos pasan la línea de separación.

3.5.2. Construcción del índice de identificabilidad

El clasificador se construye con el estimador puntual MAP_cal que se obtiene con la información de la cadena de Markov, sin embargo, como una ventaja de los métodos bayesianos es que se puede cuantificar la incertidumbre, entonces, se utilizan los distintos valores de A y α para validarlos con el clasificador y de esta manera para cada paciente se tiene un valor que indica que tan fiable es la clasificación. A este valor se le conocerá como índice de identificabilidad (II).

Para la construcción del índice de identificabilidad se siguen los siguientes pasos:

1. Seleccionar el individuo al que se le va a calcular el índice. Sea x_i el sujeto con el que se va a trabajar y y_{x_i} su etiqueta según la clasificación de la ADA.
2. Se toma la información de A y α que queda en la cadena del sujeto x_i después de descartar los p pasos iniciales. Sean $(A_{p+1}, \alpha_{p+1}), (A_{p+2}, \alpha_{p+2}), \ldots, (A_n, \alpha_n)$ dichos puntos.
3. Se evalúa cada (A_j, α_j), con $j = p+1, p+2, \ldots, n$, en el clasificador que se construye con el estimador puntual MAP_cal, usando la etiqueta y_{x_i}.
4. Se obtiene la cantidad de puntos que quedaron bien clasificados. Este valor será m_{x_i}.
5. Se normaliza m_{x_i} para que el índice tome valores entre 0 y 1. Luego,

$$\text{II} = \frac{m_{x_i}}{n - (p+1)}$$

En las figuras 3.2 y 3.3 se muestra algunos ejemplos del resultado del índice de identificabilidad en pacientes sanos y disglicémicos.

Este índice es nombrado de esta manera debido a que en la teoría clásica de problemas inversos se estudia la identificabilidad de los parámetros, observando si el mapeo de datos a parámetros es inyectivo. En este caso, el mapeo de interés es de los datos de la OGTT a la distribución conjunta a posteriori de los parámetros de clasificación (A, α). Además,

el índice de identicabilidad se relaciona con la incertidumbre que se tiene en el proceso de clasificación del sujeto, y es por esto que este valor puede dar información complementaria de pacientes sanos que pueden estar en riesgo.

Capítulo 4

Aplicación a la población bajo estudio

La metodología propuesta en el capítulo 3 se pone en práctica con la población bajo estudio, que se describe en la sección 1.4. En este capítulo se muestran los resultados obtenidos con esta población en cada una de las etapas de la metodología. También, se revisan los métodos que se escogen y los paquetes de python que se usan para el desarrollo de los métodos.

Los métodos que se presentan se usan para hacer la revisión de la metodología. En capítulos posteriores se muestra la exploración y revisión de las distintas alternativas que se tenían en las diferentes etapas del proceso.

4.1. Estimación de parámetros y criterio de validez

Las pruebas de esta nueva población se realizan con la solución analítica del modelo de Ackerman

$$G(t) = G_0 + Ae^{-\alpha t}\,\mathrm{sen}(\omega t).$$

La estimación de los parámetros desconocidos se hace con el método de estimación bayesiana que se presenta en la sección 2.2 con la distribución a priori TTT que se describió en la sección 3.1. Se utiliza el paquete pymc3 [34] para el muestreo de MCMC y la exploración se hace con tres cadenas.

Se selecciona la primera muestra de todos los sujetos de la base de datos. Después de procesarlos, se aplica el criterio de validez del modelo de Ackerman con el estimador MAP_cal. El número de sujetos que pasan este criterio es 1302, lo cual corresponde al 68.12 % del conjunto de datos total. La cantidad de datos que cumplen el criterio en cada uno de los cuatro grupos en los que se dividió la población es la siguiente:

- Grupo de hombres con muestras de glucosa e insulina: 296 lo que corresponde al 67.27 %.
- Grupo de hombre con solo muestras de glucosa: 133 lo que corresponde al 70 %.
- Grupo de mujeres con muestras de glucosa e insulina: 620 lo que corresponde al 65.5 %.
- Grupo de mujeres con solo muestras de glucosa: 253 lo que corresponde al 69.27 %.

Como los porcentajes de los cuatro grupos son similares, se trabaja con toda la población sin dividirla por género.

La cantidad de datos en cada una de las condiciones de diabetes y prediabetes es: 102 IFG, 186 IGT, 106 IFG-IGT, 129 T2DM y 687 NGT.

4.2. Clasificación SVM

Se toman los parámetros A y α como índices de clasificación, como se presenta en la sección 2.3. Para el proceso de clasificación los datos son escalados debido a la diferencia de magnitudes entre ambos parámetros y se utiliza la implementación de la librería sklearn. Se evalúa el desempeño del clasificador lineal construido en la sección 2.3 con los datos de las 80 mujeres usando el conjunto de datos que pasaron el criterio de validez del modelo de Ackerman. En la figura 4.1, se muestra el resultado del clasificador, el cual tiene un éxito del 83.56 %. Si no se considera la información de los pacientes con IFG se tiene un éxito del 89.09 %.

Como se tienen más datos, en la figura 4.1 se puede observar cómo se forman tres regiones dadas por pacientes sanos y con IFG, con IGT o IFG-IGT y pacientes con diabetes. La mezcla entre pacientes sanos y con IFG se debe a que el comportamiento de la curva de los pacientes con IFG a las dos horas es como la de un paciente sano.

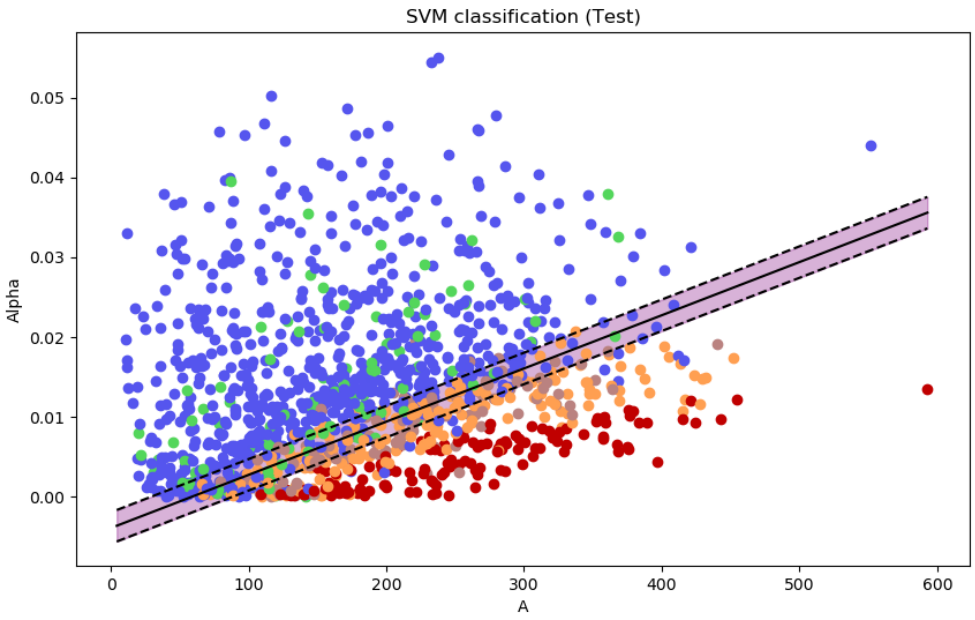


Figura 4.1: NGT(●), IFG(●), IGT(●), IFG-IGT(●), T2DM(●)

Por lo tanto, con los parámetros de clasificación A y α se puede ver un patrón interesante en el cambio de estado en dirección de las manecillas del reloj.

Por otro lado, la línea de separación se obtuvo de una población pequeña y al probarla en una más grande se ve que esta línea se sitúa entre las personas sanas y las disglicémicas. Esto muestra que el clasificador es robusto, ya que da buenos resultados con las distintas poblaciones que se han tomado.

Al haber obtenido buenos resultados con esta línea de separación, se dividen los 1911 sujetos en un conjunto de entrenamiento, conformado por el 60 % y un conjunto de prueba, conformado por el 40 %. El total de datos en cada grupo es de 1146 en el de entrenamiento y 765 en el de prueba.

Después de aplicar el criterio de validez del modelo de Ackerman, el conjunto de entremaniento quedó con 794 sujetos y el de prueba con 508 sujetos.

Para construir el clasificador se debe seleccionar un valor para la variable de sintonización C. Para esto se toman diferentes valores de C y se evalua la precisión del clasificador, se toma el valor de C con el que se obtiene mejor precisión. Se observa que valores cercanos al C seleccionado también dan buenos resultados en la exactitud del clasificador, por lo tanto este parámetro no es muy sensible a pequeños cambios.

Se construyó el clasificador sin considerar los pacientes con IFG, tomando $C = 3{,}536$ y se evalúa su rendimiento. La cantidad de sujetos con IFG en el conjunto de entrenamiento es 61 y en el de prueba es 49. En la figura 4.2 se muestra el resultado, en donde el éxito del clasificador en el conjunto de entremaniento es del 88.81 % y en el de prueba del 90.84 %.

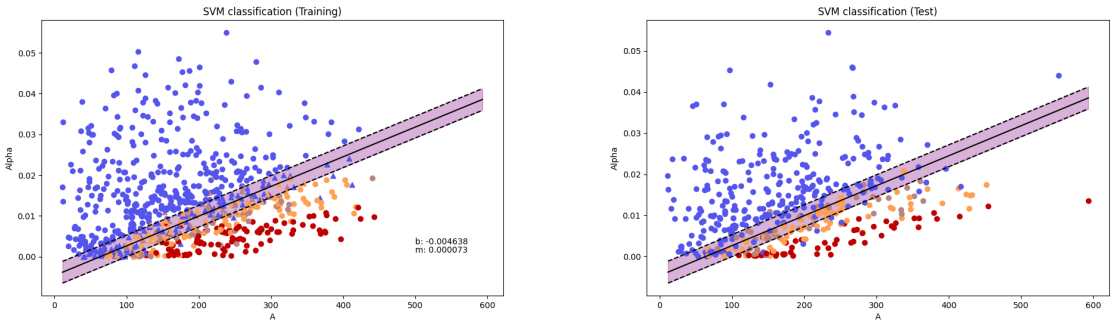


Figura 4.2: Clasificación lineal con el conjunto de entrenamiento (izquierda) y con el conjunto de prueba (derecha) usando los resultados de la distribución TTT y el estimador MAP_cal.

La matriz de confusión del conjunto de datos que se muestran en la figura 4.2 es:

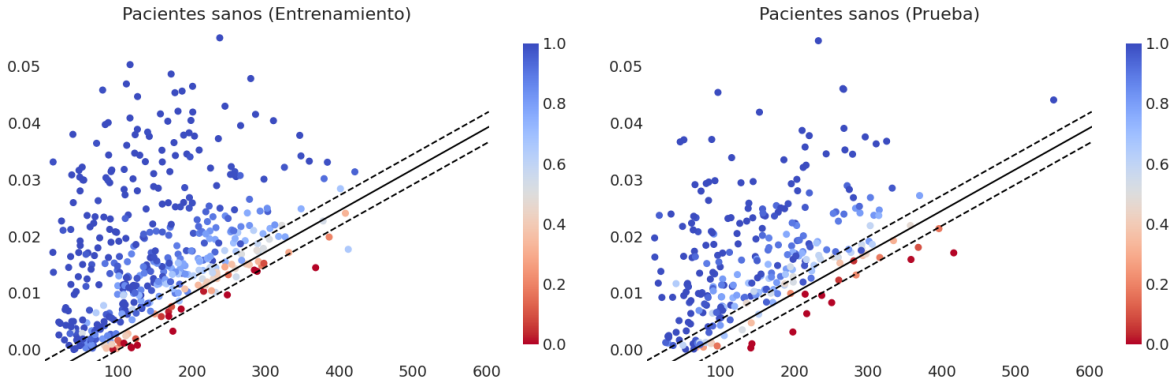
- Conjunto de entrenamiento: 404 sanos bien clasificados, 247 disglicémicos bien clasificados, 42 sanos mal clasificados y 40 disglicémicos mal clasificados,
- Conjunto de prueba: 266 sanos bien clasificados, 151 disglicémicos bien clasificados, 27 sanos mal clasificados y 15 disglicémicos mal clasificados.

La cantidad de datos disglicémicos en ambos grupos es menor, por lo que nuevamente se obtiene que una gran cantidad de datos disglicémicos se clasifican de forma correcta.

4.3. Índice de identificabilidad

A partir de la información de la cadena de cada sujeto, se evalúa el número de puntos que quedan bien clasificados y se le asocia un índice de identificabilidad. La cantidad de puntos de la cadena que se usan para calcular el índice es de 6000, ya que se generan tres cadenas de 5000 pasos por cada dato y se descartan los primeros 3000 pasos.

En la figura 4.3 se presenta el valor de este índice de pacientes sanos en los conjuntos de entrenamiento y prueba, y en las figura 4.4 el de pacientes disglicémicos en los conjuntos de entrenamiento y prueba. El clasificador que se muestra en estas figuras es el construido con los 733 sujetos del conjunto de entrenamiento y presentado en la figura 4.2.

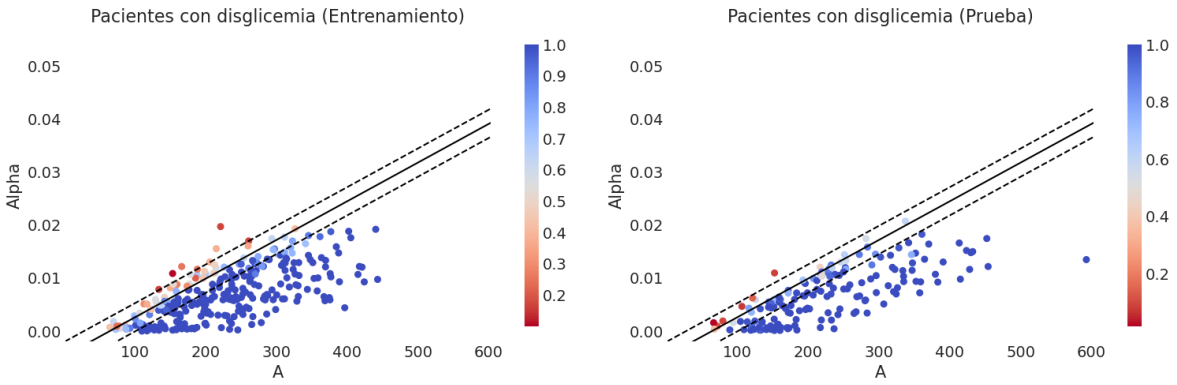


(a) (b)

Figura 4.3: Índice de identificabilidad de sujetos sanos

En la tabla 4.1 se muestra la cantidad de datos en los distintos valores que puede tomar el índice de identificabilidad, con su respectivo porcentaje tanto en el conjunto de entrenamiento como en el de prueba, sin considerar los datos de IFG. En las figuras y en la tabla se observa que un gran número de datos tienen un índice de identificabilidad entre 0.8 y 1. Por ejemplo, en el conjunto de entrenamiento el 74.47 % están en este intervalo y en el conjunto de prueba el 76.24 %.

Se presenta en la tabla 4.2 la cantidad de datos que estuvieron bien y mal clasificados en los distintos intervalos de valores que puede tomar el índice de identificabilidad, en el



(a) (b)

Figura 4.4: Índice de identificabilidad de sujetos disglicémicos

Intervalo del índice	Sujetos (Entrenamiento)			Sujetos (Prueba)		
	Cantidad	%	Acumulado	Cantidad	%	Acumulado
1	282	38.47	38.47	157	34.2	34.2
0.9 - 0.99	205	27.97	66.44	149	32.46	66.66
0.8 - 0.89	59	8.05	74.49	44	9.59	76.25
0.7 - 0.79	50	6.82	81.31	28	6.1	82.35
0.6 - 0.69	36	4.91	86.22	19	4.14	86.49
0.5 - 0.59	24	3.27	89.49	22	4.79	91.28
0.4 - 0.49	19	2.59	92.08	9	1.96	93.24
0.3 - 0.39	20	2.73	94.81	6	1.31	94.55
0.2 - 0.29	11	1.5	96.31	4	0.87	95.42
0.1 - 0.19	10	1.36	97.67	6	1.31	96.73
0.0 - 0.09	17	2.32	99.99	15	3.27	100

Tabla 4.1: Número de datos en los distintos intervalos del índice de identificabilidad en el conjunto de entrenamiento y prueba.

conjunto de entrenamiento y en el de prueba. Estos valores son consistente, ya que los datos que quedaron bien clasificados tienen un índice alto y los datos que quedaron mal clasificados tienen un índice bajo. Se obseva que solo dos datos quedaron mal clasificados y tuvieron un índice mayor a 0.8, por lo que se podría hacer seguimiento de estos casos y

observar el comportamiento de su curva de glucosa.

Intervalo del índice	Sanos(Entrenamiento)		Sanos(Prueba)		Disglicémicos(Entrenamiento)		Disglicémicos(Prueba)	
	Bien	Mal	Bien	Mal	Bien	Mal	Bien	Mal
1	101	0	57	0	181	0	100	0
0.9 - 0.99	169	0	115	0	36	0	33	1
0.8 - 0.89	48	1	38	0	10	0	6	0
0.7 - 0.79	36	0	21	0	10	4	6	1
0.6 - 0.69	24	2	13	2	5	5	1	3
0.5 - 0.59	12	3	15	2	2	7	3	2
0.4 - 0.49	7	4	4	2	1	7	2	1
0.3 - 0.39	5	7	2	3	1	7	0	1
0.2 - 0.29	2	2	1	3	1	6	0	0
0.1 - 0.19	0	6	0	4	0	4	0	2
0.0 - 0.09	0	17	0	11	0	0	0	4

Tabla 4.2: Número de datos que quedaron bien y mal clasificados en los respectivos intervalos del índice de identificabilidad en el conjunto de entrenamiento y prueba.

Por otro lado, la distribución de este índice en el plano A-α corrobora la identificabilidad de estos parámetros frente a la clasificación, ya que no se ve una mezcla del valor de este índice en el plano sino que se observa que este valor va disminuyendo a medida que se acerca a la línea separadora.

En las figuras 4.3 y 4.4 se observa que muchos pacientes que están cerca de la línea separadora tienen un valor de índice alto, por lo que tomar como referencia la distancia del representate del sujeto al separador no es un buen referente para seleccionar pacientes propensos a tener diabetes. Debido a que el índice de identificabilidad se construye con la información de cada sujeto, este nos sirve como un indicador para el seguimiento de un paciente. El umbral sería seleccionado por el médico.

Capítulo 5

Exploración de distintos métodos y propuestas en algunas etapas de la metodología

En este capítulo se comparan los resultados que se obtienen con distintos métodos en la estimación de parámetros para justificar la elección del enfoque seleccionado, se hacen propuestas de distribuciones a priori más informativas y de algunas estimadores puntuales, y se pone a prueba su desempeño tanto en el criterio de validez como en la etapa de clasificación. También, se muestra como se seleccionan las cotas en el criterio de validez.

5.1. Estimación de parámetros

Una vez seleccionado el modelo que explica la interacción de glucosa-insulina, se estiman los parámetros de interés de este modelo. Para resolver este problema existen distintos métodos que se puede utilizar. Se exploran dos opciones: una basada en aplicar métodos deterministas de optimización clásica y otra basada en un enfoque bayesiano.

Los métodos deterministas de optimización son más baratos computacionalmente comparados con los métodos bayesianos para la estimación de parámetros pero se debe hacer una revisión del comportamiento de estos métodos para el problema que se quiere resolver.

Para el enfoque bayesiano se tomaron las distribuciones a priori como uniforme, pero se

hacen algunos cambios en la elección de los intervalos descritos en el capítulo 2. Además, se toma el parámetro $\delta = \frac{\pi}{2}$, como se explica en el capítulo 3. Las densidades uniformes que se utilizan son:

- $A \sim U[0, 700]$
- $\alpha \sim U[0, 0.1]$
- $\omega \sim U[0, 0.1]$.

En los métodos de optmización se evaluaron dos algoritmos, el primero es el algoritmo de Broyden-Fletcher-Goldfarb-Shanno (BFGS) sin restricciones [29], el cual resuelve el problema

$$\min_{x^j} \quad F(x^j) = \frac{1}{2}||G^j(t) - y_t^j||^2 \tag{5.1}$$

donde $x^j = (A^j, \alpha^j, \omega^j)$ son los parámetros del sujeto j, y_t^j es la muestra de la OGTT en el tiempo t y $G^j(t) = y_0^j + A^j e^{-\alpha^j t} \operatorname{sen}(\omega^j t)$ con $t = 0, 30, 60, 90, 120$.

El segundo es el método de Newton truncado (TNC) con restricciones de caja [29], el cual resuelve el problema

$$\min_{x^j} \quad F(x^j) = \frac{1}{2}||G^j(t) - y_t^j||^2 \tag{5.2}$$

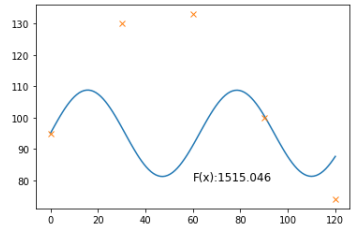
$$\text{sujeto a} \quad 0 < A < 700 \tag{5.3}$$

$$0 < \alpha < 0.1 \tag{5.4}$$

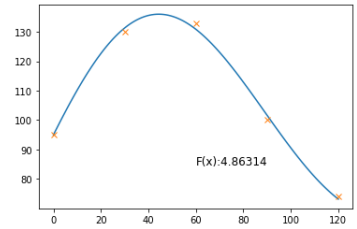
$$0 < \omega < 0.1. \tag{5.5}$$

La solución en los métodos de optimización depende del punto inicial que se tome. Por ejemplo, utilizando el método TNC, se ve en la figura 5.1 que se pueden obtener distintas soluciones dependiendo de donde se inicialice el algortimo. Los resultados obtenidos en la gráfica 5.1(a) fueron iniciando en (70.0, 0.0004, 0.09) y los de la gráfica 5.1(b) iniciando en (30.0, 0.0004, 0.02). Por lo tanto, para los algoritmos de optimización se debe seleccionar un buen punto inicial para no caer en extremos locales y esta búsqueda puede ser exhaustiva.

Después de explorar los dos métodos de optimización, se observa que con ambos se encuentran sujetos con un buen ajuste y con una solución similar. En la figura 5.2 se

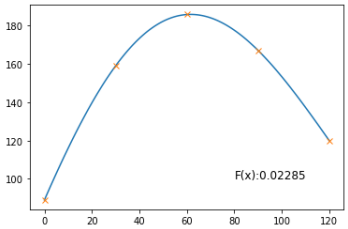


(a) A=138.07, α=0.00005, ω=0.1

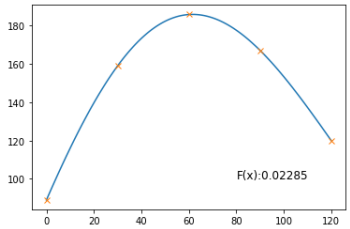


(b) A=49.03, α=0.0038, ω=0.0327

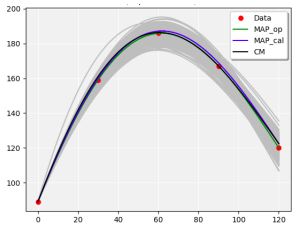
Figura 5.1: Solución del método de optimización TNC tomando distintos puntos iniciales.



(a) **BFGS:** A=124.43, α=0.0038, ω=0.0228



(b) **TNC:** A=124.43, α=0.0038, ω=0.0228



(c) A=128.36, α=0.004, ω=0.0225

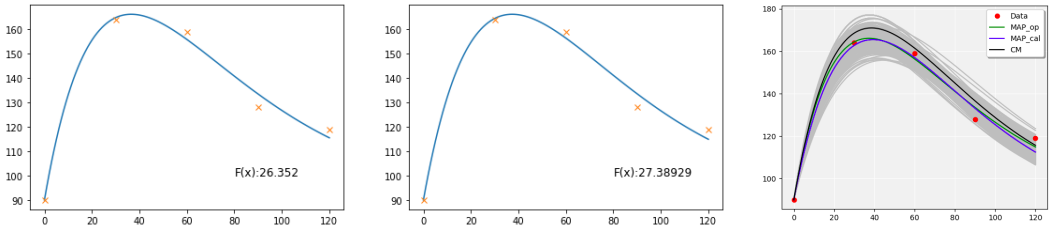
Figura 5.2: Sujeto en el que los métodos de optimización y de bayes dan una solución similar.

ve un ejemplo en donde se muestra esto. En los casos de optimización se inician ambos algoritmos en el punto (70.0, 0.0004, 0.02).

También, se pueden encontrar sujetos en los cuales se ven similares las gráficas de la curva, pero los parámetros que se encuentran varían mucho. En la figura 5.3 se muestra un ejemplo en donde ocurre lo mencionado. En los casos de optimización se inician los algoritmos en el punto (70.0, 0.0004, 0.02). En la gráfica 5.3(b), el parámetro A toma el valor límite y en la gráfica 5.3(a) como con el método BFGS no se restringe el valor de A, éste toma un valor muy grande.

El valor de la función objetivo (5.2) en la figura 5.3(b) es mayor que la función objetivo (5.1) en la figura 5.3(a). Se puede pensar que como el método TNC tiene restricciones y

el parámetro A alcanza el valor límite de esta, sin la restricción el valor de A continuaría aumentando para que la función objetivo pueda alcanzar un menor valor pero el valor que toma A es físicamente irreal.



(a) **BFGS:** A=16553.17, α=0.0273, ω=0.000342 (b) **TNC:** A=700, α=0.0259, ω=0.0078 (c) A=345.62, α=0.0226, ω=0.0142

Figura 5.3: Sujeto con curvas similares pero valores en los parámetros diferentes.

Esta situación ocurrió con varios datos y como la metodología propuesta depende del valor de los parámetros, se optó por utilizar las técnicas bayesianas para estimar los parámetros del modelo, dado que no se tuvo este tipo de problemas, aunque el costo computacional fuera más alto. El método bayesiano tarda 20 veces más que los métodos de optimización.

En la figura 5.4 se muestran los histogramas de las distribuciones marginales de los parámetros del sujeto de la figura 5.3. Esta es otra ventaja de la estimación bayesiana, que se puede obtener esta información a través de la muestra de la distribución a posteriori que genera el algoritmo MCMC. Este método nos da mayor información de los parámetros que se estan estimando, a diferencia de los algoritmos de optimización que solo encuentran una solución puntual.

Otras razones por las que se selecciona un enfoque bayesiano son las siguientes:

- En el método bayesiano se pueden encontrar un intervalo con las distintas muestras de la distribución y cuantificar la incertidumbre de la solución.
- Los algoritmos de optimización encuentra un valor máximo, mientras que el bayesiano encuentra el valor que se alcanza con más frecuencia, lo cual lo hace menos dependiente del ruido. [11]

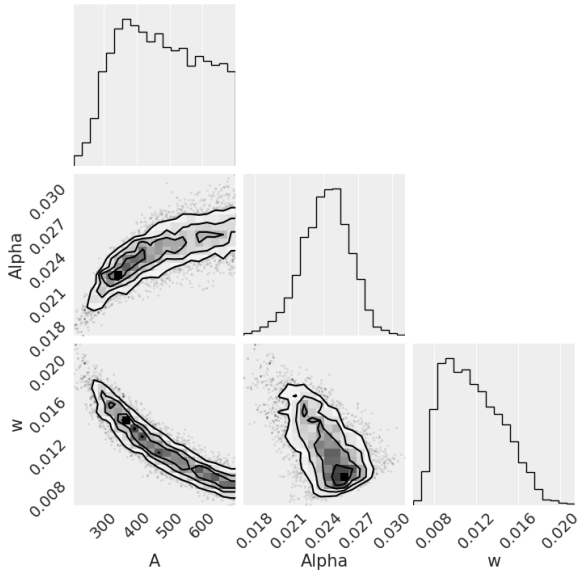


Figura 5.4: Distribuciones marginales para cada parámetro del sujeto de la figura 5.3

5.2. Distribuciones a priori

Después de seleccionar un enfoque bayesiano para la estimación de parámetros, se debe tomar las distribuciones a priori de los parámetros de acuerdo a la información conocida de estos parámetros. En este caso se proceso la base de datos con distribuciones a priori uniformes siguiendo lo propuesto en el capítulo 2.

Con el estimador MAP obtenido en este caso se generaron los histogramas de la figura 3.1 del capítulo 3. En la partición que se hizo de los datos en cuatro grupos, el comportamiento de los histogramas en cada parámetro es similar. En la figura 5.5 se muestran los histogramas del grupo de hombres con mediciones de glucosa e insulina, sobre estos se graficó la función de densidad obtenida con el método de kernels.

En estos histogramas se puede ver que el comportamiento de los parámetros no es de una distribución uniforme. Es por esto que se proponen dos opciones para tomar las distribuciones a priori, debido a que al ver el histograma del parámetro ω se piensa en dos formas de seleccionar su distribución. Las dos propuestas que se hacen son las siguientes:

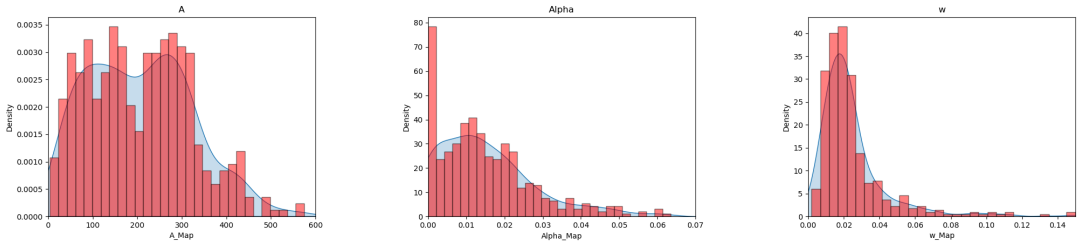
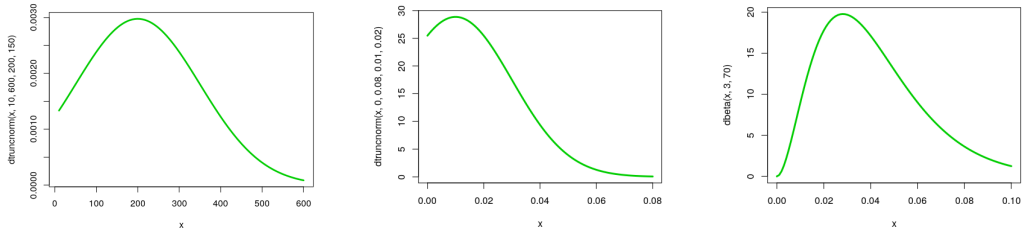


Figura 5.5: Histograma del MAP de los parámetros estimados del grupo de hombres con glucosa e insulina con la función de densidad usando el método de kernels

Distribución TTB

Para el parámetro A se toma una normal truncada, para el parámetro α se toma una normal truncada y para el parámetro ω se toma una beta.

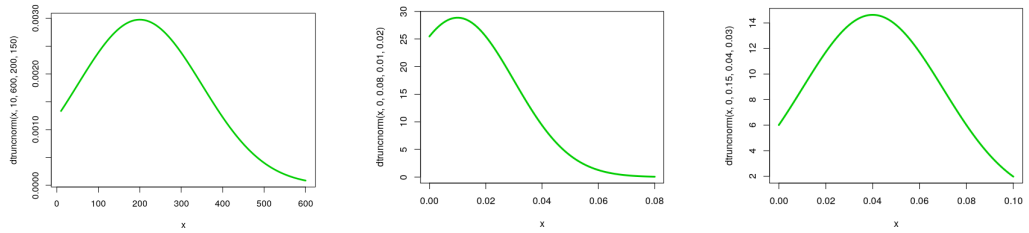


(a) Parámetro A (b) Parámetro α (c) Parámetro ω

Figura 5.6: $A \sim N[200,150]$ truncada en (10,600), $\alpha \sim N[0.015,0.02]$ truncada en (0,0.08) y $\omega \sim$ B[3,70].

Distribución TTT

Para el parámetro A se toma una normal truncada, para el parámetro α se toma una normal truncada y para el parámetro ω también se toma una normal truncada.



(a) Parámetro A (b) Parámetro α (c) Parámetro ω

Figura 5.7: $A \sim N[200,150]$ truncada en (10,600), $\alpha \sim N[0.015,0.02]$ truncada en (0,0.08) y $\omega \sim N[0.04,0.03]$ truncada en (0, 0.15).

Los parámetros de los sujetos de la base de datos se estimaron con ambas propuestas para ver cuál era la opción más conveniente para tomar como distribución a priori del parámetro ω. Para esto se tiene en cuenta la cantidad de datos que pasan el criterio de validez y el éxito del clasificador como se mostrará más adelante.

5.3. Estimadores puntuales

Como la construcción del clasificador se hace con un representante de cada sujeto, se debe tomar un estimador puntual a partir de la muestra de la distribución a posteriori que genera el algoritmo de MCMC. En el capítulo 2 se presentan dos estimadores, los cuales son el máximo a posteriori (MAP) y la media condicional (CM).

Una dificultad del estimador MAP es que para encontrarlo se debe resolver un problema de optimización. Algunos paquetes que implementan el algoritmo MCMC, tienen alguna función o algún método para encontrar este estimador pero puede haber algunos problemas en el cálculo. Es por esto que se propone el estimador MAP_cal el cual se puede encontrar con la información de la cadena. En el capítulo 3 se describe como se construye.

Para probar el desempeño de este estimador se compararon los resultados de éste con otros dos estimadores tanto en el momento de seleccionar curvas con un buen ajuste como en la contrucción del clasificador. Por lo tanto, los tres estimadores que se ponen a prueba son:

- La media condicional (CM) que se encuentra con la información de la cadena.
- El estimador MAP_op que se encuentra utilizando la función find_MAP() del paquete pymc3. Esta función resuelve un problema de optimización, utilizando la información de las distribuciones a priori y el modelo de verosimilitud.
- El estimador MAP_cal que se encuentra ajustando una función de densidad con el método de kernels al histograma generado de la distribución conjunta a posteriori de los tres parámetros estimados.

5.4. Criterio de validez del modelo de Ackerman

En la prueba de concepto se construye el clasificador usando la información estimada de los 80 datos, pero después de estimar los parámetros se encuentran curvas que no tienen buen ajuste o que el modelo no puede explicar la información del sujeto, y estos datos pueden producir ruido al momento de generar el clasificador. Es por esto que antes de la etapa de clasificación se propone un criterio de validez del modelo de Ackerman. De esta manera, el clasificador se construye con la información de sujetos que tienen curvas con un buen comportamiento.

Para la construcción de este criterio de validez se consideran tres características:

- Los perfiles de glucosa durante la OGTT mencionados en la sección 1.3.
- El error absoluto entre las mediciones del sujeto G_i y el valor de la glucosa en el modelo de Ackerman con los parámetros del estimador seleccionado G_i^{Est}.

$$\text{Error}_{abs} = \frac{1}{5} \sum_{i \in \{0,30,60,90,120\}} |G_i - G_i^{Est}|$$

- La frecuencia angular de la curva de ajuste (ω).

En donde se utiliza la primera característica para determinar que tan riguroso se debe tomar el valor para acotar el error y la última para evitar curvas con un gran número de oscilaciones.

El modelo de Ackerman describe los datos OGTT de un paciente, si $\omega < 0.09$ y se cumple alguna de las siguientes condiciones:

1. La curva cumple alguna condición para clasificarse según su forma y $|G_{90} - G_{120}| > 4.5$ mg/dl; $\text{Error}_{abs} < 7.5$.

2. La curva cumple alguna condición para clasificarse según su forma y $|G_{90} - G_{120}| < 4.5$ mg/dl; $\text{Error}_{abs} < 5$.

3. La curva se considera "Sin clasificar"; $\text{Error}_{abs} < 4.5$.

En los perfiles de glucosa durante la OGTT se observa que las curvas pueden tener a lo más dos oscilaciones durante los primeros 120 minutos. De forma experimental se encuentra que tomando a $\omega < 0.09$ se lograba esto. Es por esto que el valor de la frecuencia angular es acotado con este valor. Se toma en cuenta esta característica para evitar que se obtengan curvas con errores pequeños en el ajuste debido a que tienen muchas oscilaciones, como se muestra en la figura 5.8(a).

Los perfiles de glucosa indican si las mediciones de glucosa tiene cierta forma, o si se consideran curvas sin clasificar o anómalas, y dentro de las curvas sin clasificar se dividen en las que la diferencia entre la medición 90 y 120 es menor a 4.5 mg/dl y las que no cumplen las características mencionadas en la sección 1.3. Debido a esto, las curvas se dividen en estos tres grupos y la cota del error en cada grupo se maneja de forma independiente, siendo más flexible con las curvas que sí se clasifican por su forma y más exigentes con las curvas sin clasificar. Las curvas sin clasificar no se descartan ya que se encuentran curvas con un error bajo y un buen ajuste, como se observa en la figura 5.9(b) y (c).

Para la selección de la cota del error, se analiza el error absoluto entre las mediciones de glucosa y el valor estimado de glucosa usando los valores de cierto estimador en cada uno de los grupos. El estudio se realizó con las a priori con distribución uniforme, TTB y TTT, y los tres estimadores propuestos. En la tabla 5.1 se muestra el error mínimo, máximo y medio en cada grupo con las distintas combinaciones de distribución y estimador.

Se observa que los valores del error medio del estimador MAP_op y el estimador MAP_cal son similares en las distintas combinaciones que se tienen, mientras que los valores del error con el CM son más grandes en los tres grupos. Esto ocurre debido a que el paquete que se usa genera varias cadenas de exploración, y en algunos sujetos estas cadenas generan una distribución a posteriori multimodal, por lo que el estimador CM es un mal representante en este caso y genera un error alto.

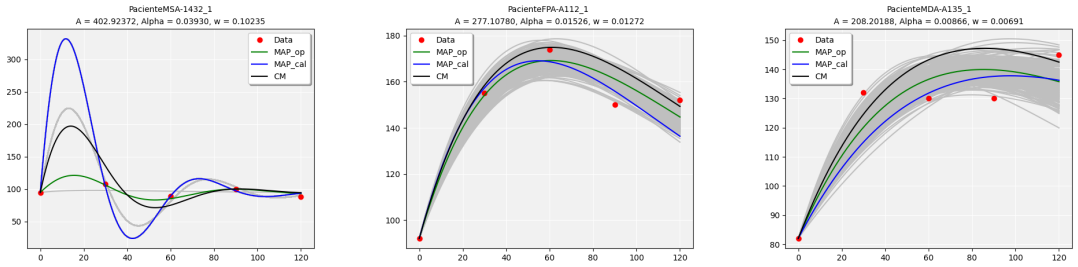
Estimador	Grupo	Error	Uniforme	TTB	TTT
MAP_op	Clasificado	mínimo	6.87e-06	0.109	0.112
		máximo	85.802	308.656	85.832
		medio	5.043	5.4	6.425
	Sin Clasificar porque $\|G_{90} - G_{120}\| < 4.5$	mínimo	0.052	0.24	0.234
		máximo	18.014	17.994	59.864
		medio	4.726	5.005	5.859
	"Sin clasificar"	mínimo	0.312	0.524	0.638
		máximo	127.4	295.456	103.564
		medio	8.711	9.966	11.331
MAP_cal	Clasificado	miñimo	0.307	0.331	0.249
		máximo	85.812	85.585	85.519
		medio	5.757	5.771	5.959
	Sin Clasificar porque $\|G_{90} - G_{120}\| < 4.5$	mínimo	1.355	0.941	1.22
		máximo	18.397	18.972	30.326
		medio	5.394	5.461	5.581
	"Sin clasificar"	mínimo	1.092	0.964	1.422
		máximo	64.793	58.772	97.87
		medio	8.336	8.379	9.904
CM	Clasificado	mínimo	0.742	0.74	0.765
		máximo	102.879	173.536	161.276
		medio	8.099	8.25	8.765
	Sin Clasificar porque $\|G_{90} - G_{120}\| < 4.5$	mínimo	0.244	0.331	0.364
		máximo	59.542	129.379	225.746
		medio	7.913	7.865	8.742
	"Sin clasificar"	mínimo	3.372	2.239	1.92
		máximo	64.27	151.232	224.191
		medio	12.614	15.664	20.723

Tabla 5.1: Estudio del error absoluto en las curvas clasificadas según su forma y las curvas sin clasificar con las distintas distribuciones y los distintos estimadores.

Por esta razón, para elegir las cotas de las errores se tiene en cuenta principalmente los resultados de los estimadores MAP_op y MAP_cal. Luego, para el grupo de curvas que se clasifican según su forma, en donde se es más flexible con el error, se toma un valor por

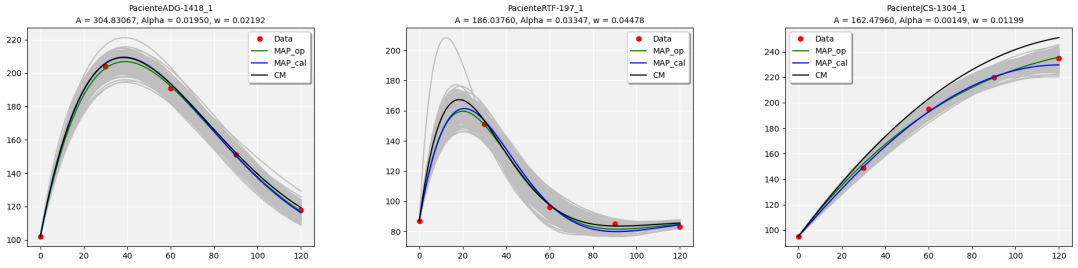
encima de la media y que no alcance el valor medio del estimador CM, por esto la cota es 7.5. Para el grupo de curvas que no se clasifican porque la diferencia de las dos últimas mediciones es menor a 4.5 , se toma como cota de error un valor cercano a la media, por eso la cota del error es 5. Para el grupo "Sin clasificar"los errores son más altos y por esto se toma una cota por debajo de la media y menor a la del segundo grupo, por esta razón la cota del error se selecciona como 4.5.

En la figuras 5.8 se muestran algunas curvas que no cumplen las condiciones mencionadas y en la figura 5.9 algunas curvas que cumplen las condiciones.



(a) Condición 1 (b) Condición 2 (c) Condición 3

Figura 5.8: (a)Curva trifásica, $|G_{90} - G_{120}| = 12$, Error$_{abs}$=2.658 y ω=0.102. (b) Curva bifásica, $|G_{90} - G_{120}| = 2$, Error$_{abs}$=5.834 y ω=0.0127. (c) Error$_{abs}$=7.03 y ω=0.0069.



(a) Condición 1 (b) Condición 2 (c) Condición 3

Figura 5.9: (a) Curva monofásica, $|G_{90} - G_{120}| = 33$, Error$_{abs}$=1.299 y ω=0.021. (b)Curva monofásica, $|G_{90} - G_{120}| = 2$, Error$_{abs}$=2.164 y ω=0.044. (c) Error$_{abs}$=1.674 y ω=0.011.

5.5. Resultados con las propuestas de distribuciones a priori y estimadores

Para probar el desempeño de las distintas propuestas de distribuciones a priori (Uniforme, TTB, TTT) y estimadores puntuales (MAP_op, MAP_cal, CM), se calculan cuántos datos pasan el criterio de validez y el éxito del clasificador. La distribución uniforme que se toma es la definida en la sección 5.1.

Para las pruebas la población se divide en un conjunto de entrenamiento, conformado por el 60 % y un conjunto de prueba, conformado por el 40 %. El total de datos en cada grupo es de 1146 en el de entrenamiento y 765 en el de prueba.

5.5.1. Criterio de validez

Se aplica el criterio de validez del modelo de Ackerman a las distintas combinaciones que se forma entre distribuciones y estimadores. Además, con el estimador MAP_op se tuvo un filtro adicional para la selección de la curva, si el parámetro A tomaba el valor del límite superior tampoco se selecciona el sujeto. Esto ocurrió en varios casos debido a que el estimador MAP_op resuelve un problema de optimización, y como se vió en la sección 5.1 esta era un problemática de este método.

El número de datos en los conjuntos de entrenamiento y de prueba después de aplicar el criterio de validez se muestran en la tabla 5.2. También, se presenta el porcentaje al que corresponde este número dentro del grupo (% grupo), es decir, de entrenamiento o de prueba y el porcentaje al que corresponde dentro del conjunto total de datos (% total).

Los resultados con el estimador MAP_cal en las tres distribuciones son similares. El estimador MAP_op con la distribución uniforme son los que tienen los resultados más bajos y esto se debe a que hay bastantes casos en los que el parámetro A se va al límite superior, esto ocurre en menor medida con la distribución TTT y la TTB. El estimador CM selecciona un número alto con las distribuciones TTB y TTT, pero como se mencionó en la sección anterior al ver el análisis del error de este estimador, cuando al final de la exploración de un sujeto se observan dos cadenas distintas, este estimador no es la mejor opción.

Distribución		MAP_op	MAP_cal	CM	Conjunto
Uniforme	N°	625	807	520	Entrenamiento
	% grupo	54.53	70.41	45.37	
	% total	32.7	42.22	27.21	
	N°	383	497	350	Prueba
	% grupo	50.06	64.96	45.75	
	% total	20.04	26	18.31	
TTB	N°	828	787	736	Entrenamiento
	% grupo	72.25	68.67	64.22	
	% total	43.32	41.18	38.51	
	N°	523	494	467	Prueba
	% grupo	68.36	64.57	61.04	
	% total	27.36	25.85	24.43	
TTT	N°	735	794	704	Entrenamiento
	% grupo	64.13	69.28	61.43	
	% total	38.46	41.54	36.83	
	N°	465	508	449	Prueba
	% grupo	60.78	66.40	58.69	
	% total	24.33	26.58	23.49	

Tabla 5.2: Cantidad de datos seleccionados con cada distribución y cada estimador

5.5.2. Clasificación lineal

Al observar los datos en el plano A-α, se ve un cambio de estado en la dirección de las manecillas del reloj, y por esto un kernel lineal es una muy buena alternativa para explicar nuevos conjuntos de datos, como se muestra en la figura 4.2.

Los resultados que se obtuvieron con los datos que fueron seleccionados, sin considerar a los pacientes con alteración de glucosa en ayuno (IFG), con cada distribución apriori y cada estimador se presentan en la tabla 5.3. En este caso las clasificación se realizó sin balancear los datos.

Como el número de pacientes sanos es mayor que el de pacientes con diabetes o con alguna alteración, se realiza un balanceo de datos y se construye el clasificador. Los resultados con los datos balanceados se presentan en la tabla 5.4. El valor de éxito disminuye

		MAP_op	MAP_cal	CM	
Kernel	Distribución	Precisión(%)	Precisión(%)	Precisión(%)	Conjunto
Lineal	Uniforme	86.91	86.04	88.17	Entrenamiento
		85.5	84.88	86.26	Prueba
	TTB	88.28	89.56	90.08	Entrenamiento
		89.24	89.01	92.10	Prueba
	TTT	89.73	88.81	91	Entrenamiento
		89.97	90.84	91.56	Prueba

Tabla 5.3: Resultados del clasificador lineal con las distintas pruebas que se realizaron entre distribuciones y estimadores con los datos sin balancear

en todos los casos y al observa en la figura 4.2, la línea separadora que se obtiene de los datos sin balancear, se aprecia que esta queda bien ubicada entre los pacientes sanos y los que tienen alguna alteración. Por lo que, para la parte de clasificación se trabaja con los datos originales sin balancear.

		MAP_op	MAP_cal	CM	
Kernel	Distribución	Precisión(%)	Precisión(%)	Precisión(%)	Conjunto
Lineal	Uniforme	85.86	81.07	85.37	Entrenamiento
		85.79	84.88	84.02	Prueba
	TTB	85.28	85.02	86.68	Entrenamiento
		86.7	87.44	85.64	Prueba
	TTT	85.93	86.08	85.89	Entrenamiento
		85.2	87.14	84.36	Prueba

Tabla 5.4: Resultados del clasificador lineal con las distintas pruebas que se realizaron entre distribuciones y estimadores con los datos balanceados

Después de construir el clasificador omitiendo los datos de las personas que tienen alteración de glucosa en ayuno, se observó el éxito del clasificador al considerar toda la población que paso el criterio de validez en cada una de las distribuciones y en cada uno de los estimadores. Los resultados se presentan en la tabla 5.5.

		MAP_op	MAP_cal	CM	
Kernel	Distribución	Precisión(%)	Precisión(%)	Precisión(%)	Conjunto
Lineal	Uniforme	81.28	80.79	80.19	Entrenamiento
		79.37	81.08	80.28	Prueba
	TTB	83.93	85	84.23	Entrenamiento
		83.93	82.79	83.72	Prueba
	TTT	84.76	83.62	83.8	Entrenamiento
		82.58	84.44	84.18	Prueba

Tabla 5.5: Resultado del clasificador lineal construido sin los datos con IFG, al considerar todos los datos.

5.6. Comentarios

En la figura 5.4 se observa la distribución marginal de los parámetros del sujeto cuyo ajuste se muestra en la figura 5.3 cuando se toman uniformes como distribución a priori. En la distribución marginal de A se ve como si se pudiera seguir explorando en valores más grandes, aunque viendo el resultado de los parámetros estimados de este paciente su estimador MAP no esta cerca del límite. Este comportamiento se pudo observar en varios sujetos.

Después de seleccionar otras distribuciones a priori, el comportamiento de las distribuciones marginales de este paciente cambiaron, como se muestra en la figura 5.10. Lo que nos sugiere que si era necesario usar distribuciones a priori más informativas.

También, al seleccionar las distribuciones a priori diferente a uniformes se tiene una ganancia ya que permite encontrar varias alternativas para escoger un estimador que pase el criterio de selección y permita tener una buena cantidad de datos para el proceso de clasificación.

La selección de un clasificador lineal permite una buena generalización para los nuevos datos que se procesen. Como los resultados obtenidos con la propuesta TTB y TTT son similares, estos nos permite escoger cualquiera de las dos opciones como un buen punto de partida y seleccionar el estimador MAP_cal que fue consistente tanto en el número de datos que pasaron el criterio de selección como en el porcentaje de éxito del clasificador.

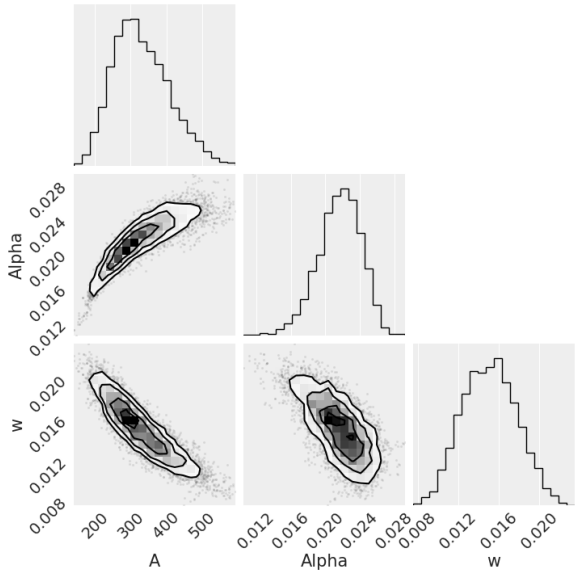


Figura 5.10: Distribuciones marginales para cada parámetro del sujeto de la figura 5.3 tomando como distribución a priori la TTT

Se toma la distribución TTT ya que fijando el estimador MAP_cal es la que tienen un poco más de datos que pasan el criterio de validez y además la que tiene un mejor desempeño en la clasificación.

Capítulo 6

Clasificación no lineal y regresión logística

Se hace una revisión de otras técnicas de clasificación para determinar la influencia que tiene la elección del método de clasificación en los resultados. Estos métodos se comparan con SVM lineal debido a los buenos resultados que se tienen con este clasificador. Los clasificadores que se analizan son máquinas de soporte vectorial con fronteras de decisión no lineales y regresión logística con distintos modelos que generan separadores lineales y no lineales. Este último método se examina ya que tiene un enfoque probabilístico a diferencia de las máquinas de soporte vectorial que tienen un enfoque geométrico.

6.1. Máquinas de soporte vectorial con fronteras de decisión no lineales

Para algunos conjuntos de datos tener un hiperplano como separador de las dos clases puede no ser la mejor frontera de decisión, ya que el comportamiento de los datos por clase sugieren una frontera no lineal como separador. Es por esto que se amplia el espacio de características utilizando funciones de los predictores, a través de kernels [19, 13].

En el nuevo espacio de características, la frontera de decisión es lineal, pero cuando se ve el resultado en el espacio de características original el separador toma una forma no lineal.

El clasificador lineal se puede representar por:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{6.1}$$

donde x es la nueva observación y x_i son los datos del conjunto de entrenamiento.

Se puede reemplazar el producto punto que aparece en la ecuación (6.1) por una generalización del producto punto de la forma

$$K(x, x_i) \tag{6.2}$$

donde $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ es una función *kernel*, es decir, una función que expresa la similaridad de dos observaciones.

Algunas de las elecciones de kernels son:

- Lineal
$$K(x, x_i) = \langle x, x_i \rangle$$
- Polinomial de grado d
$$K(x, x_i) = (1 + \langle x, x_i \rangle)^d$$
- Base radial
$$K(x, x_i) = \exp(-\gamma ||x - x_i||^2).$$

También, se pueden transformar algunos de los predictores. Se resuelve el problema lineal para el conjunto de predictores transformados y sin transformar. Luego, en el espacio original de los predictores se obtendrá una frontera de decisión no lineal.

Por ejemplo, se toma el caso en que se tienen dos observaciones por dato, x_{i1} y x_{i2}. Se transforma la primera observación en $\log(x_{i1})$. Luego, se encuentra el hiperplano que resuelva el problema

$$\begin{aligned} \max_{\beta_0,\beta_1,\beta_2,\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n} \quad & M \\ \text{sujeto a} \quad & \beta_1^2 + \beta_2^2 = 1 \\ & y_i(\beta_0 + \beta_1 \log(x_{i1}) + \beta_2 x_{i2}) \geq M(1 - \varepsilon_i) \\ & \varepsilon_i \geq 0, \quad \sum_{i=1}^{n} \varepsilon_i \leq C. \end{aligned}$$

La frontera de separación en el plano x_1-x_2 se ve como una función logarítmica.

6.1.1. Resultados de SVM no lineal

Se exploran otros tipos de clasificadores, para ver si un separador no lineal divide mejor a los datos de cada clase representados por los puntos (A_i, α_i) con los que se esta trabajando. Para esto, se calculan los porcentajes de éxito de la clasificación y si el clasificador se puede generalizar para un nuevo conjunto de datos.

Se toma el kernel polinomial de grado 3 y el de base radial. También, se transforma la variable A en $\log(A)$ y se calcula el clasificador lineal en el plano $\log(A) - \alpha$. Los clasificadores se construyen con los datos obtenidos de las distintas combinaciones de distribuciones propuestas y de estimadores, sin considerar los datos con alteración de glucosa en ayuno.

Kernel de base radial

En la tabla 6.1 se muestra el éxito del clasificador con un kernel de base radial, en el conjunto de entrenamiento y en el de prueba. Los resultados obtenidos con el kernel radial son buenos. En todos los casos son un poco más altos que con el kernel lineal. Esto puede ocurrir debido a la curva que se forma en la parte en donde existe mayor acumulación de datos.

Una desventaja de este clasificador es que la región de pacientes disglicémicos es una curva cerrada contenida en la región de pacientes sanos, la cual puede generar un sobreajuste en el conjunto de entrenamiento. En la figura 6.1 se ve esta situación, ya que se tiene un paciente enfermo con un valor de A grande, que queda mal clasificado, debido a que en el conjunto de entrenamiento no se tuvieron pacientes disglicémicos con valores de A grandes.

Por esta razón, aunque el valor de éxito es alto, la forma en como divide el espacio este clasificador no es lo más conveniente para este conjunto de datos.

Kernel polinomial de grado 3

En la tabla 6.2 se muestra el éxito del clasificador con un kernel polinomial de grado tres. Con este clasificador se obtienen los valores de éxito más bajos.

En la figura 6.2 se ve que la separación de la región más grande de disglicémicos con la región de sanos, se puede aproximar con una recta. Por lo que, no se tiene ganancia con este kernel. Además esta región está un poco más abajo que la recta que se logra con el

		MAP_op	MAP_cal	CM	
Kernel	Distribución	Precisión(%)	Precisión(%)	Precisión(%)	Conjunto
Radial	Uniforme	89.52	88.59	89.46	Entrenamiento
		89.56	89.77	87.53	Prueba
	TTB	90.1	90.52	91.27	Entrenamiento
		91.13	91.03	92.58	Prueba
	TTT	91.05	89.9	92.09	Entrenamiento
		91.4	91.06	92.55	Prueba

Tabla 6.1: Resultados del clasificador con un kernel radial para las distintas combinaciones de distribuciones y estimadores

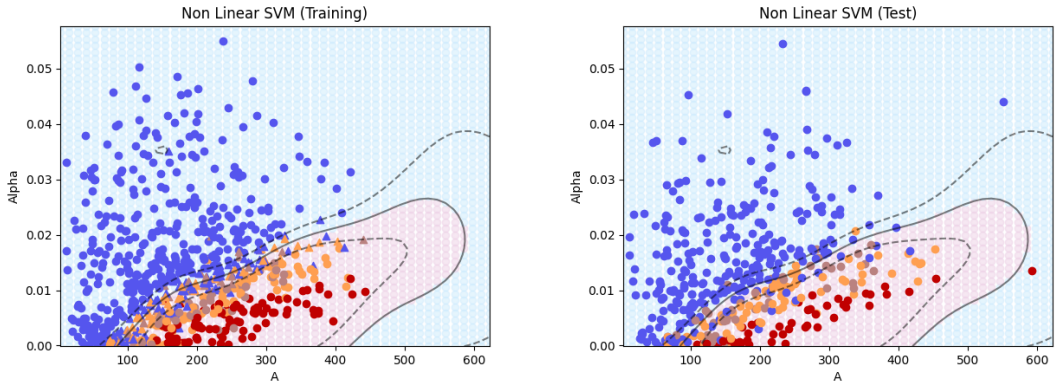


Figura 6.1: Clasificación SVM con kernel radial usando los resultados de la distribución TTT y el estimador MAP_cal. El conjunto de entrenamiento (izquierda) y el conjunto de prueba (derecha).

kernel lineal, por lo que la cantidad de pacientes disglicémicos mal clasificados aumenta y el valor de éxito del clasificador disminuye.

Mas aún, este kernel forma una región de disglicémicos cerca de la esquina superior derecha de la figura, que no se es natural para el comportamiento de estos datos, lo que puede provocar que se clasifique mal a pacientes sanos, debido a que los pacientes sanos pueden tener valores de A y α grandes. En el conjunto de prueba de la figura 6.2 se ve esta situación.

		MAP_op	MAP_cal	CM	
Kernel	Distribución	Precisión(%)	Precisión(%)	Precisión(%)	Conjunto
Polinomial	Uniforme	72.6	83.08	87.52	Entrenamiento
		77.39	83.55	88.17	Prueba
	TTB	86.84	88.59	90.53	Entrenamiento
		87.34	88.11	90.9	Prueba
	TTT	88.12	87.72	91	Entrenamiento
		88.3	87.58	91.31	Prueba

Tabla 6.2: Resultados del clasificador con un kernel polinomial de grado 3 para las distintas combinaciones de distribuciones y estimadores

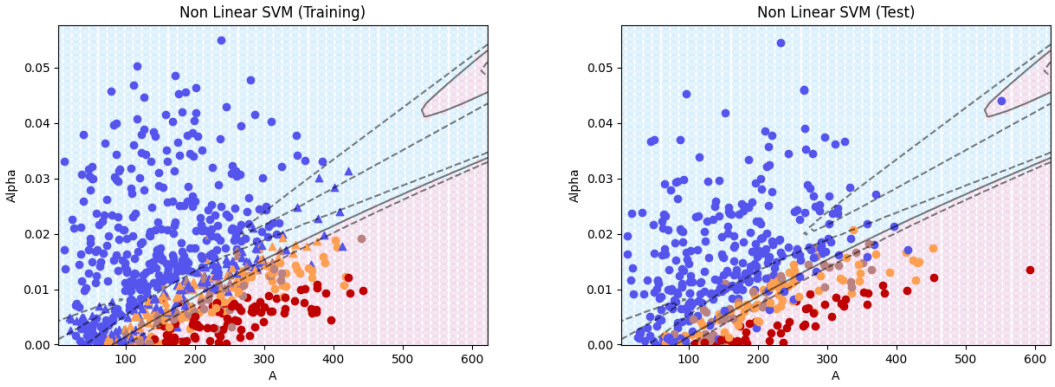


Figura 6.2: Clasificación SVM con kernel polinomial de grado tres usando los resultados de la distribución TTT y el estimador MAP_cal. El conjunto de entrenamiento (izquierda) y el conjunto de prueba (derecha).

Frontera de decisión logarítmico

En la tabla 6.3 se muestra el éxito del clasificador con una frontera de decisión logarítmico. Los resultados obtenidos con la transformación de la variable A en $\log(A)$ son buenos, se obtienen valores similares que cuando se usa el kernel de base radial y en la mayoría de casos más altos que con el clasificador lineal.

Con esta transformación en los datos se pudo obtener una curva, que era una ventaja del kernel de base radial. Además, al igual que en el caso lineal la forma en como se divide

el espacio en las regiones de sanos y de disglicémicos se ve natural para este conjunto de datos, esto se puede deber a la relación logarítmica que existe entre los parámetros A y α.

Por lo tanto, este clasificador como el clasificador lineal son buenas alternativas. En una sección posterior se observa el comportamiento del índice de identificabilidad para este clasificador y de esta manera ver si hay una ganancia extra en alguno de los dos clasificadores.

		MAP_op	MAP_cal	CM	
Kernel	Distribución	Precisión(%)	Precisión(%)	Precisión(%)	Conjunto
Lineal-log	Uniforme	89	89.53	88.81	Entrenamiento
		90.14	90.22	89.13	Prueba
	TTB	89.71	90.1	91.27	Entrenamiento
		91.35	91.7	93.3	Prueba
	TTT	91.2	89.08	91.93	Entrenamiento
		90.93	92.37	92.05	Prueba

Tabla 6.3: Resultados del clasificador lineal construido con $\log A$ y α

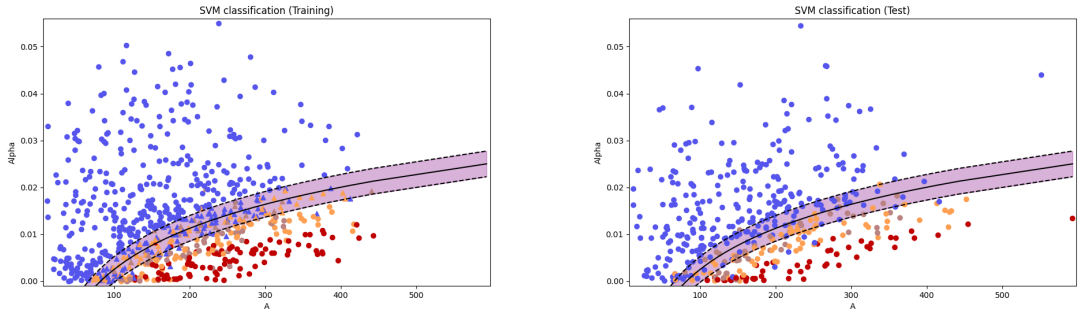


Figura 6.3: Clasificación SVM lineal usando los resultados de la distribución TTT y el estimador MAP_cal, al conjunto de datos de construido con $\log A$ y α. El conjunto de entrenamiento (izquierda) y el conjunto de prueba (derecha).

6.2. Regresión logística

El método de máquinas de soporte vectorial construye el clasificador teniendo en cuenta la distancia de las observaciones al separador, su enfoque es geométrico. Es por esto, que surge la pregunta de si se puede tener alguna ganancia en la clasificación al considerar un método que tenga un enfoque probabilístico.

Una alternativa es usar el método de regresión logística, en el cual se modela la probabilidad de que un dato pertenezca a una clase [19, 13]. En este método se pueden considerar varias clases.

Si se considera que se tienen dos clases, donde la variable Y representa que ocurra o no un evento. Esta toma el valor de 1 si ocurre el suceso y 0 en caso contrario. Se supone que Y depende de las variables $X_1, X_2, \ldots, X_p$, que son las observaciones que se tienen de cada dato.

Se busca una función que modele la probabilidad de que ocurra el suceso dado que el sujeto tiene las siguientes características $X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p$. Para esto, se utiliza la función logística

$$P(Y = 1|x_1, x_2, \ldots, x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} \tag{6.3}$$

Para estimar el intercepto β_0 y los coeficientes $\beta_1, \beta_2, \ldots, \beta_p$, a partir de los datos del conjunto de entrenamiento, se usa el método de máxima verosimilitud. Para esto, se calcula el máximo del logaritmo de la función de verosimilitud

$$L(\beta_0, \beta_1, \ldots, \beta_p; y) = \sum_{i=1}^{n} y_i \log(P_i) + (1 - y_1) \log(1 - P_{i'}) \tag{6.4}$$

donde $y = (y_1, y_2, \ldots, y_n)$.

De forma similiar al caso de kernels en el método SVM, se puede ampliar el espacio de características mediante funciones cuadráticas, cúbicas o polinómicas de orden superior de las observaciones. También, se puede modificar alguna de las observaciones con una función y en el espacio de características se reemplaza esa variable por la variable transformada.

6.2.1. Resultados de regresión logística

Se proponen tres formas de seleccionar el espacio de características:

- A y α
- $\log(A)$ y α
- A, A^2, A^3 y α,

con la primera propuesta se obtiene un separador lineal, con la segunda un separador logarítmico y con la tercera un separador polinomial.

Las propuestas se hacen de esa manera para poder hacer una comparación con los resultados de SVM y observar si hay alguna ganancia al utilizar otro método de clasificación. El espacio de características con funciones polinomiales se hace de esa forma para evitar la región que se forma en la esquina superior derecha de la figura 6.2.

De igual manera se realiza la clasificación con los datos obtenidos de las distintas combinaciones que se tienen entre las propuestas realizadas de distribuciones a priori y de estimadores.

Modelo de A y α

En este caso se va a tener que la probabilidad de que un sujeto sea sano esta dado por

$$P(Y = \text{sano}|A, \alpha) = \frac{\exp(\beta_0 + \beta_1 A + \beta_2 \alpha)}{1 + \exp(\beta_0 + \beta_1 A + \beta_2 \alpha)}. \tag{6.5}$$

En la figura 6.4 se ven tres rectas, las líneas punteadas corresponden a la probabilidad del 30 % y del 70 %, y la línea lisa corresponde a la probabilidad del 50 %. Las líneas punteadas y las del margen en SVM cambian un poco, mientras que la línea lisa se ve muy similar a la recta que se encuentra con el método de SVM. Esto se corrobora cuando se ven los resultados de éxito que se presentan en la tabla 6.4, en la cual se encuentra que en varios casos se obtiene el mismo resultado que con SVM o la diferencia es mínima.

Desviación residual: 373.58 con 730 grados de libertad.

	MAP_op	MAP_cal	CM	
Distribución	Precisión(%)	Precisión(%)	Precisión(%)	Conjunto
Uniforme	86.21	84.69	88.6	Entrenamiento
	85.79	85.55	86.9	Prueba
TTB	88.15	89.69	90.82	Entrenamiento
	88.81	89.01	92.1	Prueba
TTT	90.02	88.4	90.69	Entrenamiento
	89.73	90.84	91.56	Prueba

Tabla 6.4: Resultados de regresión logística en el caso lineal con las distintas pruebas que se realizaron entre distribuciones y estimadores

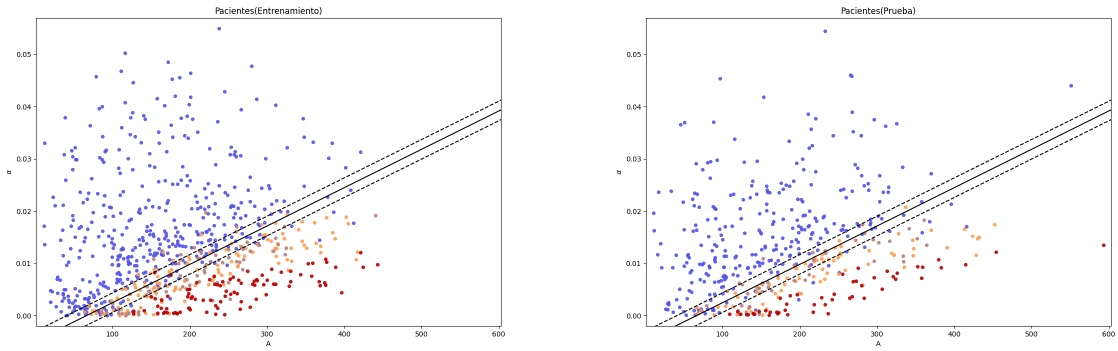


Figura 6.4: Clasificación lineal con regresión logística usando los resultados de la distribución TTT y el estimador MAP_cal. El conjunto de entrenamiento (izquierda) y el conjunto de prueba (derecha).

Modelo de $\log(A)$ **y** α

En este caso se va a tener que la probabilidad de que un sujeto sea sano esta dado por

$$P(Y = \text{sano}|A, \alpha) = \frac{\exp(\beta_0 + \beta_1 \log(A) + \beta_2 \alpha)}{1 + \exp(\beta_0 + \beta_1 \log(A) + \beta_2 \alpha)}. \tag{6.6}$$

Como con SVM se ven buenos resultados con este modelo, se explora el desempeño que tendría este clasificador. En el tabla 6.5, se muestran los valores de éxito que se ob-

tuvieron. Nuevamente, los resultados en ambos casos son muy similares. En la figura 6.5, se muestran las curvas que corresponden a la probabilidad del 30 %, 50 % y 70 % con los datos obtenidos con las distribución TTT y el estimador MAP_cal.

Desviación residual: 346.23 con 730 grados de libertad.

	MAP_op	MAP_cal	CM	
Distribución	Accuary(%)	Accuary(%)	Accuary(%)	Conjunto
Uniforme	89	88.59	89.46	Entrenamiento
	89.85	90	89.77	Prueba
TTB	89.84	89.69	91.71	Entrenamiento
	91.13	91.03	93.06	Prueba
TTT	91.34	89.35	92.09	Entrenamiento
	91.88	92.37	91.81	Prueba

Tabla 6.5: Resultados de regresión logística con el modelo $\beta_0 + \beta_1 \log(A) + \beta_2\alpha$ con las distintas pruebas que se realizaron entre distribuciones y estimadores

Modelo de A, A^2, A^3 y α

En este caso se va a tener que la probabilidad de que un sujeto sea sano esta dado por

$$P(Y = \text{sano}|A, \alpha) = \frac{\exp(\beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 A^3 \beta_4 \alpha)}{1 + \exp(\beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 A^3 \beta_4 \alpha)}. \tag{6.7}$$

En la tabla 6.6 se muestra el éxito del clasificador cuando se considera en el espacio de características funciones polinomiales de A. En este caso, la frontera de decisión se ve como un polinomio pero se obtiene algo distinto que con SVM, ya que en SVM se toman funciones polinomiales de α. En este caso, se ve una ganancia en el clasificador. Los resultados son similares a los obtenidos con el separador con forma logarítmica.

En la figura 6.6 se ve que en la parte donde hay mayor acumulación de los datos la curva que se genera se asemeja a la curva que se obtiene con el modelo anterior, pero después del punto de inflexión de la curva no se ve que esta ajuste con el comportamiento que se ha visto de los datos para valores grandes de A. Es por esto que no consideramos

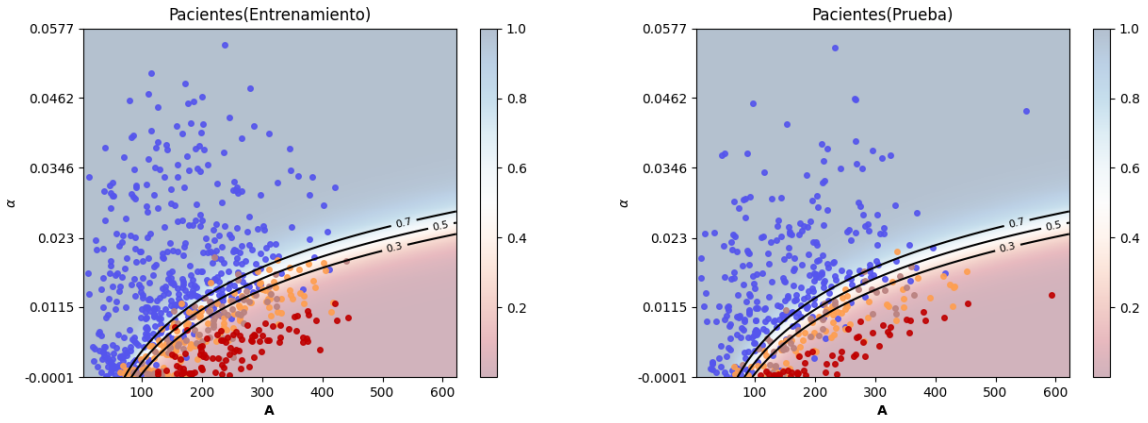


Figura 6.5: Clasificación con regresión logística con el modelo $\beta_0 + \beta_1 \log(A) + \beta_2 \alpha$ usando los resultados de la distribución TTT y el estimador MAP_cal. El conjunto de entrenamiento (izquierda) y el conjunto de prueba (derecha).

este clasificador, aunque tenga buenos resultados de éxito, ya que es difícil generalizar para nuevos conjuntos de datos.

Desviación residual: 345.18 con 728 grados de libertad.

	MAP_op	MAP_cal	CM	
Distribución	Accuary(%)	Accuary(%)	Accuary(%)	Conjunto
Uniforme	88.65	88.72	88.6	Entrenamiento
	90.43	90	89.77	Prueba
TTB	90.36	89.97	91.12	Entrenamiento
	91.35	90.35	93.54	Prueba
TTT	91.05	89.35	91.78	Entrenamiento
	92.12	92.37	92.8	Prueba

Tabla 6.6: Resultados de regresión logística con el modelo $\beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 A^3 + \beta_4 \alpha$ con las distintas pruebas que se realizaron entre distribuciones y estimadores

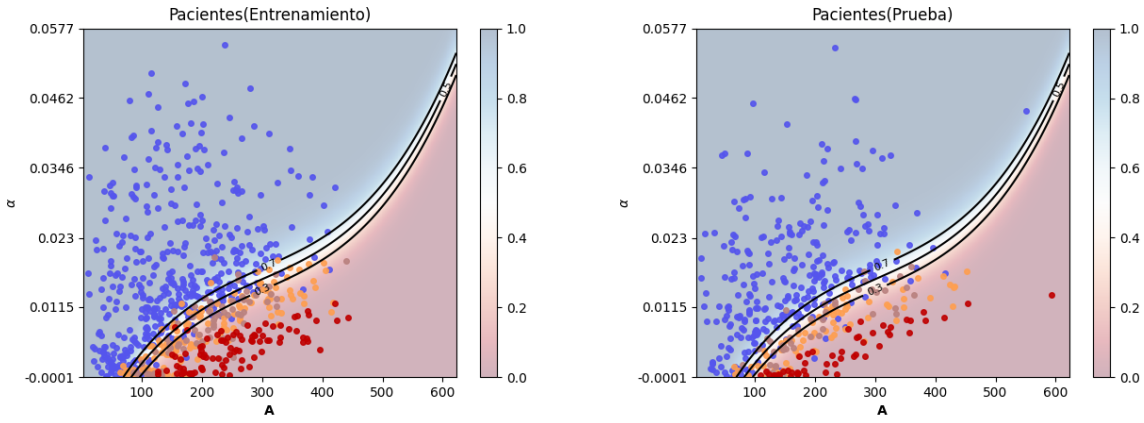


Figura 6.6: Clasificación con regresión logística con el modelo $\beta_0+\beta_1 A+\beta_2 A^2+\beta_3 A^3+\beta_4 \alpha$ usando los resultados de la distribución TTT y el estimador MAP_cal. El conjunto de entrenamiento (izquierda) y el conjunto de prueba (derecha).

La desviación residual indica que tan bien el modelo predice la respuesta con los predictores seleccionados. Valores altos de este índice indican que el modelo no es tan preciso y con valores pequeños el modelo puede tener sobreajuste. Aunque no existe un umbral que indique el mejor valor de este índice, se puede comparar con la desviación nula, el cual indica que tan bien el modelo predice la respuesta cuando solo se considera la intercesión.

En este caso, la desviación nula fue de 981.39 con 732 grados de libertad. Comparando este valor con la desviación residual obtenida en los tres modelos se observa que este valor disminuye. Lo que muestra que los predictores que se agregan ayudan a tener modelos que explican mejor los datos pero este valor sigue siendo un poco alto en los tres casos.

6.3. Índice de identificabilidad en los distintos clasificadores

Una de las preguntas que surgieron era si se podia construir una zona de pacientes susceptibles. Aunque las máquinas de soporte vectorial construyen un margen, esta opción por ser geométrica no sugería ser una buena alternativa. Por esto, se genera el clasificador usando regresión logística pensando que las probabilidades que predicen pertenecer a

una clase pueda servir como umbral para encontrar pacientes susceptibles, pero también se forma una región parecida a la del margen de las máquinas de soporte vectorial. Por tal razón, en este trabajo, se propone usar la información de la cadena generada por el método MCMC que se tiene de cada sujeto junto al clasificador y así generar un índice por paciente que sirva para determinar la susceptibilidad del paciente a padecer la enfermedad.

El valor que toma el índice de identificabilidad de cada individuo está completamente relacionado con la selección del clasificador, ya que con éste se evalua la cantidad de datos dentro de la cadena que quedan bien clasificados. Aunque este valor puede variar según el clasificador seleccionado, se hizo una revisión para determinar si el valor del índice es estable para los distintos clasificadores.

Por lo visto anteriomente, los clasificadores que tienen un buen valor de éxito y además no tienen un sobreajuste de los datos, son el lineal y el logarítmico con cualquiera de los métodos. Es por esto que solo en estos casos se analizó el comportamiento del índice de identificabilidad.

En la figura 6.7 se muestran el valor del índice de identificabilidad tomando el clasificador lineal que se obtiene con regresión logística, de igual forma se colocan las rectas con probabilidad de 30 % y 70 %. Los resultados se ven similares a los presentados en el caso lineal con SVM, y esto se debe como ya se mencionó a que la línea que corresponde a la probabilidad 50 % no varía mucho con respecto al separador lineal de SVM.

Como la ventaja que tiene regresión logística es que su construcción tiene un enfoque probabilista pero estas probabilidades no se toman para identificar susceptibilidad y el índice de identificabilidad no varía con los dos métodos de clasificación, entonces, el proceso de clasificación se puede hacer con cualquiera de estos métodos.

El clasificador con frontera de decisión logarítmica con los dos métodos de clasificación dio un resultado similiar. Es por esto que solo se van a mostrar los resultados del índice de identificabilidad con el clasificador obtenido con máquinas de soporte vectorial, que se pueden apreciar en la figura 6.8.

En ambos casos se encuentran datos que están cerca de la frontera de decisión que tienen un índice de identificabilidad altos. En la parte en donde más se acumulan los datos, se ven similares los valores del índice tanto con el clasificador lineal como el logarítmico, pero si hay una pequeña zona en donde se encuentran algunos datos que pueden cambiar

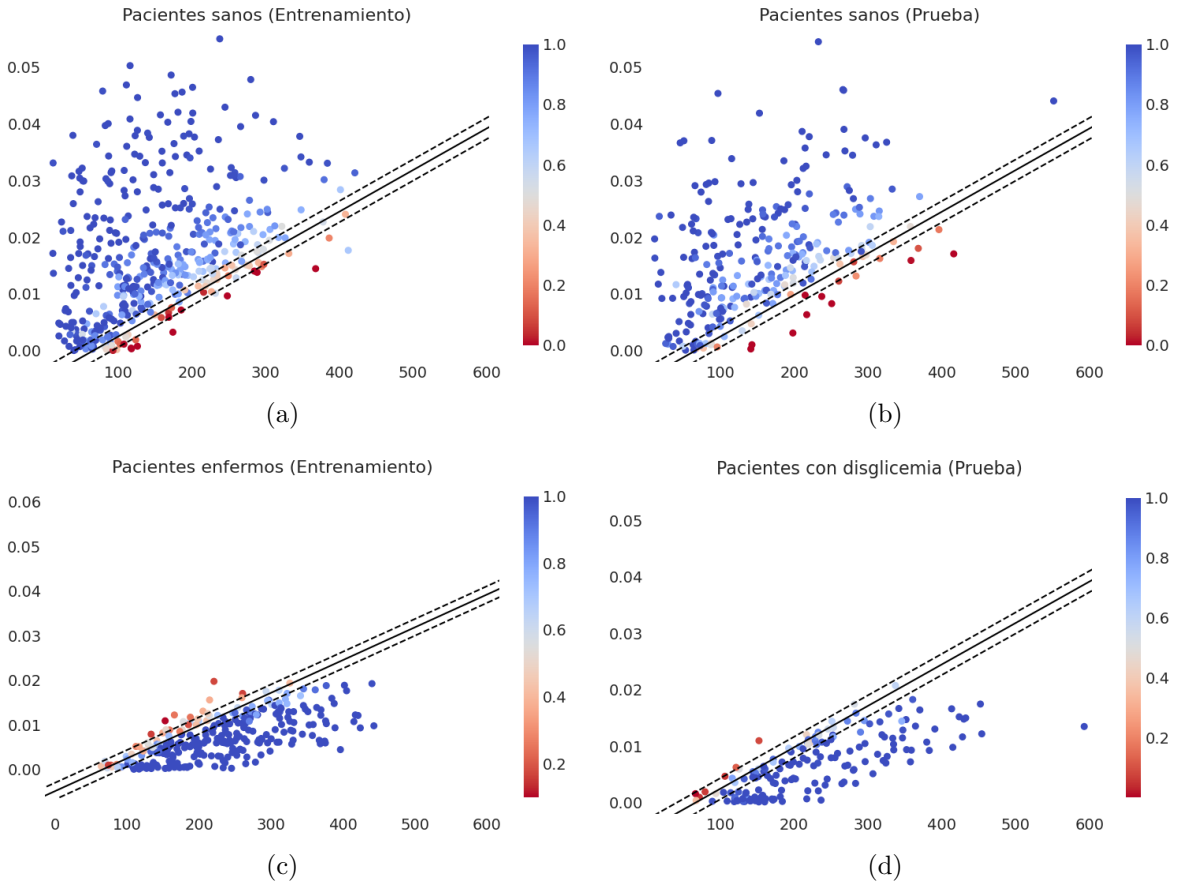


Figura 6.7: Índice de identificabilidad de sujetos sanos (arriba) y disglicémicos (abajo) con el clasificador lineal generado con regresión logísitica.

su índice de identificabilidad de un clasificador al otro.

En la tabla 6.7 se presentan como se comporta el índice en el conjunto de entrenamiento y en el de prueba en los distintos rangos de valor que puede tomar el índice con el clasificador de frontera de decisión logarítmica. En esta tabla se observa que el porcentaje de datos que tienen un índice entre 0.8 y 1, aumenta en un 6 % en el conjunto de entrenamiento y un 5.23 % en el conjunto de prueba con respecto a lo reportado con el clasificador lineal. Este cambio en la tabla se debe a que la construcción del índice se

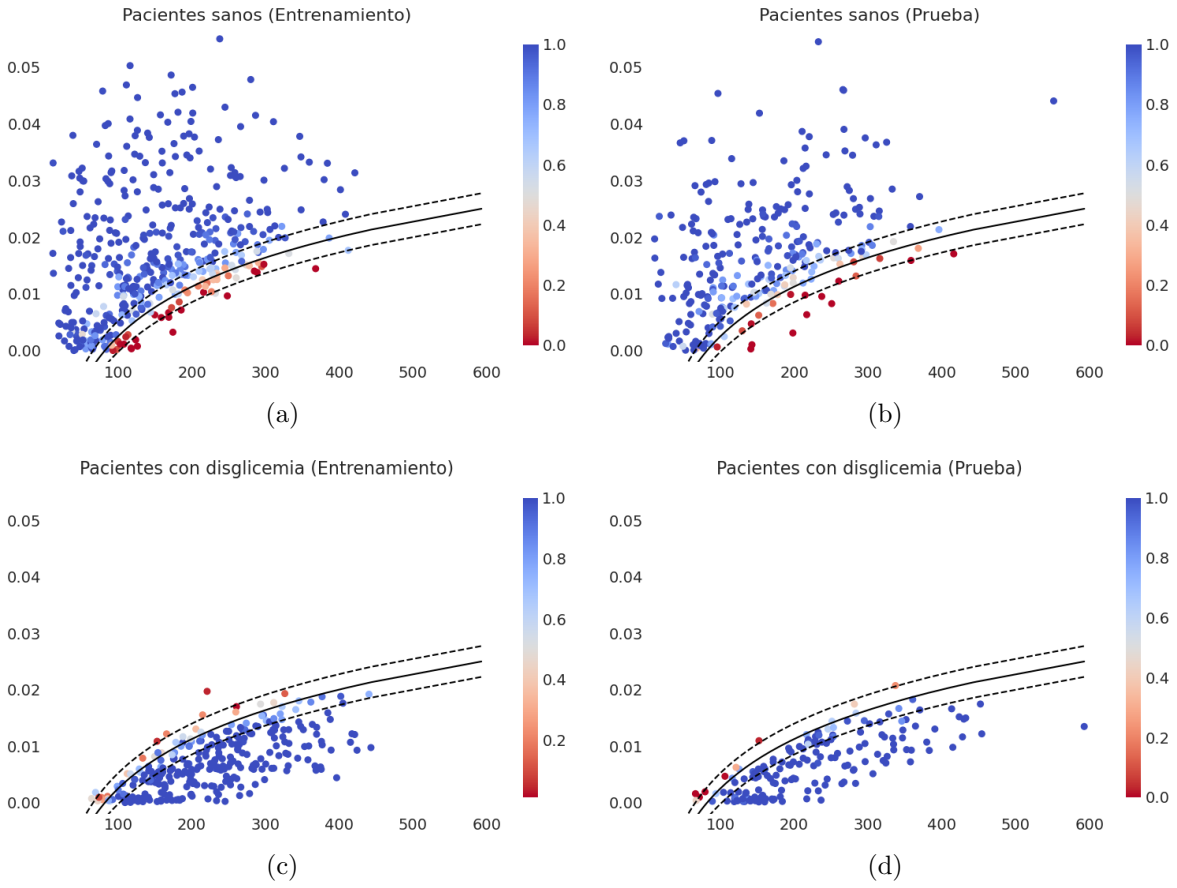


Figura 6.8: Índice de identificabilidad de sujetos sanos (arriba) y disglicémicos (abajo) con una frontera de decisión logarítmica usando SVM.

relaciona con la elección del clasificador, pero al comparar las gráficas el comportamiento del índice es similar en ambos casos. En las zonas cercanas al separador es donde más se encuentran valores del índice de identificabilidad bajos. Por lo tanto, la forma en que se construye este índice es estable y no va a variar mucho entre los diferentes clasificadores.

Es por esto que se puede elegir entre los distintos métodos de clasificación con buen desempeño, ya que el resultado es similar y en el tipo de clasificador que se usa se busca que el ajuste se vea natural al comportamiento de los datos y que se pueda generalizar para

Intervalo del índice	Sujetos (Entrenamiento)			Sujetos (Prueba)		
	Cantidad	%	Acumulado	Cantidad	%	Acumulado
1	388	52.93	52.93	228	49.67	49.67
0.9 - 0.99	163	22.24	75.17	119	25.93	75.6
0.8 - 0.89	39	5.32	80.49	27	5.88	81.48
0.7 - 0.79	32	4.37	84.86	18	3.92	85.4
0.6 - 0.69	25	3.41	88.27	20	4.36	89.76
0.5 - 0.59	21	2.86	91.13	7	1.53	91.29
0.4 - 0.49	8	1.09	92.22	9	1.96	93.25
0.3 - 0.39	13	1.77	93.99	4	0.87	94.12
0.2 - 0.29	7	0.95	94.94	3	0.65	94.77
0.1 - 0.19	8	1.09	96.03	3	0.65	95.42
0.0 - 0.09	29	3.96	99.99	21	4.58	100

Tabla 6.7: Número de datos en los distintos intervalos del índice de identificabilidad cuando se usa una frontera de decisión logarítmica

nueva información, y tanto el clasificador lineal como el de frontera de decisión logarítmica cumplen estas condiciones.

Capítulo 7

Conclusiones y trabajo futuro

En este trabajo se propone una herramienta de clasificación para determinar el estado diabético de un paciente. La clasificación es en términos de la concentración máxima de glucosa A y la tasa media de eliminación de glucosa α. La aplicación de la herramienta para construir un clasificador a la población bajo estudio muestra que estos parámetros son una elección apropiada, ya que se observa una relación con el estado diabético del paciente. Además, se genera el índice de identificabilidad que es un indicador que podría servir para saber cuando un sujeto está propenso a la enfermedad.

El método bayesiano para la estimación de parámetros es una buena alternativa. Una de sus ventajas es que genera una muestra de la distribución a posteriori de los parámetros, cuya información se usa para construir el índice de identificabilidad.

Después de dividir la base de datos en conjunto de hombres y mujeres y procesar la información con el método planteado en la prueba de concepto, se recalca que la variable correspondiente al género no aporta información adicional. El parámetro δ, que representa el ángulo de fase, se puede fijar como $\frac{\pi}{2}$. También, se puede tomar distribuciones a priori más informativas, como lo son normales truncadas para cada uno de los tres parámetros que se estiman por sujeto.

El estimador MAP es una buena opción como estimador puntual para construir el clasificador. El método que se propone para encontrar este estimador se hace de forma que solo dependa de la muestra de la distribución a posteriori de los parámetros estimados y de esta manera se puede replicar con cualquier librería o módulo que se use para resolver el problema de estimación.

Otro aporte en la metodología es el criterio de validez que nos ayuda a determinar los datos que pueden ser explicados con el modelo de Ackerman basado en los perfiles de la curva y el ajuste obtenido con cada sujeto. Un buen porcentaje de datos del problema de glucosa-insulina se pueden explicar con el modelo de Ackerman y con este filtro nos aseguramos que la construcción del clasificador se realiza con la información de sujetos que tienen un buen ajuste.

El kernel lineal da buenos resultados. Se revisan otros clasificadores para ver la influencia de éstos en la clasificación, pero con algunos de estos clasificadores se puede llegar al sobreajuste en el conjunto de entrenamiento. El clasificador con frontera de separación logarítmica es una buena opción, su desempeño es similar al clasificador lineal. Con ambos clasificadores se observa que el resultado se puede generalizar para nuevos conjuntos de datos y el comportamiento de la frontera de separación se ajusta a la distribución que se ve de los datos. Como los resultados con SVM y regresión logística con el clasificador lineal y el logístico no varían mucho, se puede elegir cualquiera de estos métodos en esta etapa de la metodología.

Con el índice de identificabilidad se pudo estudiar la identificabilidad práctica de los parámetros de interés. Además, permite dar un indicador al sujeto a partir de la información recolectada del mismo, y de esta manera la decisión de bien o mal clasificador no va a depender de que tan buena es la selección del estimador puntual. Con este índice se observó que hay un gran número de sujetos con una distribución conjunta con mucha dispersión pero dentro de la zona de la etiqueta que les corresponde y por lo tanto el valor de su índice es alto.

Se obtuvo un grupo de datos que no pasaron el criterio de validez del modelo de Ackerman. Es por esto, que para una investigación futura se puedan consideran modelos más complejos de glucosa-insulina [32, 27]. A partir de los nuevo parámetros estimados, se buscaría determinar la concentración máxima de glucosa y la tasa media de eliminación de glucosa, para evaluar si en otros modelos estos índices tienen una relación con la condición diabética.

El modelo de glucosa-insulina utilizado [1] se basa solamente en los datos de la OGTT. Se deberían considerar otro tipo de modelos si se quisiera trabajar con la información antropométrica y clínica, que es de gran utilidad para una evaluación de riesgo. Estos resultados se pueden complementar con los que se obtiene con la metodología propuesta.

Por último, se puede trabajar con pacientes que se realizan un seguimiento. En el conjunto de datos se tiene la información de algunos pacientes que se realizaron la prueba de dos a cuatro veces. Esta información podría ayudar a seguir estudiando el comportamiento que se viene observando con los índices de clasificación para la condición diabética. Por ejemplo, en la figura 7.1 se observan los parámetros $A - \alpha$ de un sujeto en tres pruebas OGTT que se tomaron en distintos años. Se puede apreciar la evolución de los índices desde la primera prueba en donde el sujeto era pre-diabético a la última en donde padece de diabetes. En la figura 7.2 se muestran los perfiles de las tres pruebas del paciente y la distibución conjunta de los parámetros con el clasificador. Se observa que estos datos tienen un buen ajuste y pasan el criterio de validez del modelo de Ackerman.

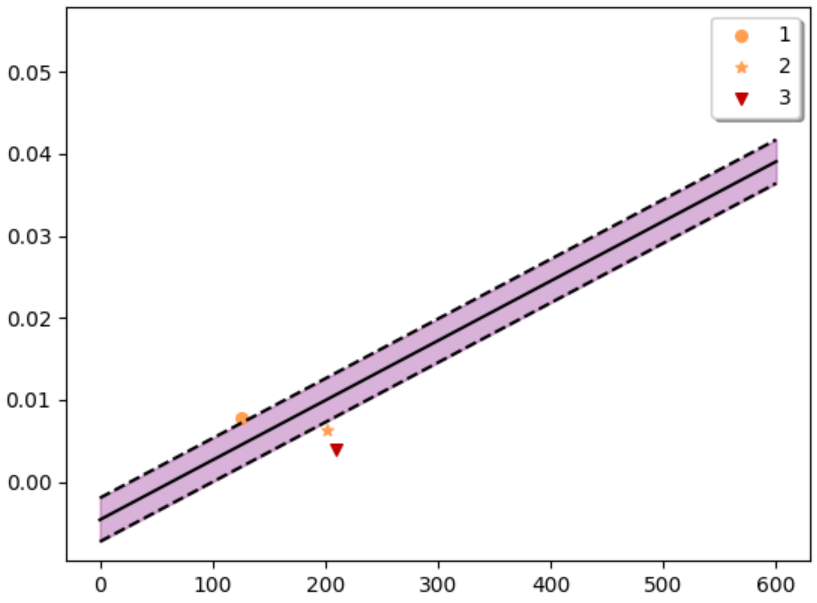
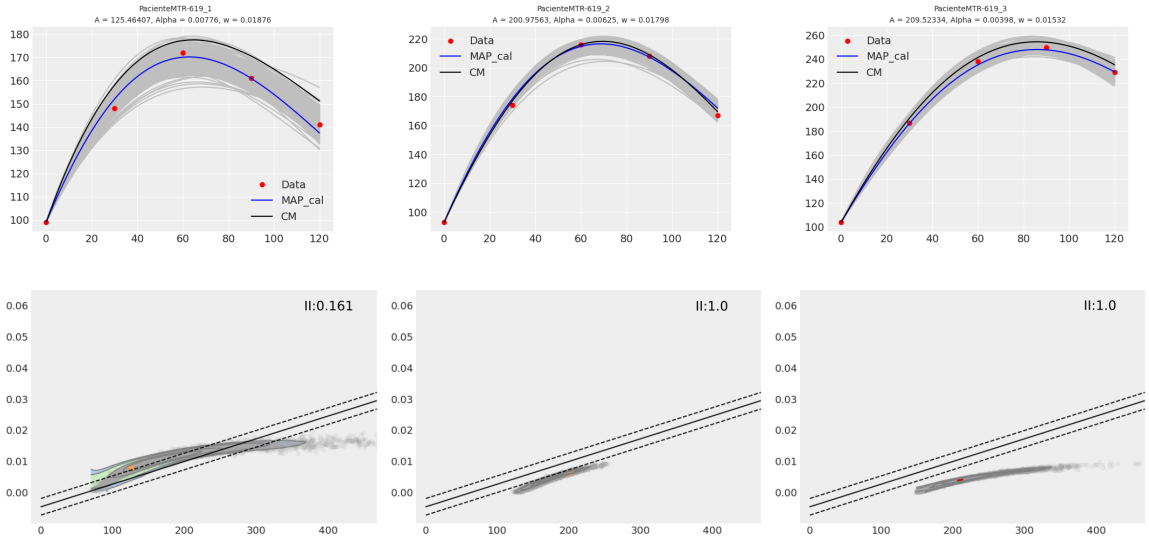


Figura 7.1: Parámetros de un sujeto de seguimiento, el cual se realiza tres OGTT en distintos años.

También, los datos de seguimiento se puede usar para estudiar cual de los dos tipos de frontera separadora que dan buenos resultados puede ser más acorde al comportamiento



(d) Prueba 1 (e) Prueba 2 (f) Prueba 3

Figura 7.2: Curvas de ajuste e índice de identificabilidad de paciente de seguimiento.

de los datos.

Bibliografía

[1] Ackerman E., Rosevear JW., McGuckin W. A mathematical model of the glucose-tolerance test. Phys. Med. BIol. 1964; 2,9: 203-213

[2] Ackerman E., Gatewook L., Rosevear J., Molnar GI. Blood glucose regulation and diabetes. In: Heinmets, F. (ed) Concepts and Models of Biomathematics: 1969. 131-156.

[3] Alyass A., Almgren P., Akerlund M., Dushoff J., Isomaa B., Nilsson P., Tuomi T., Lyssenko V., Groop L., Meyre D. Modelling of OGTT curve identifies 1 h plasma glucose level as a strong predictor of incident type 2 diabetes: results from two prospective cohorts; Diabetologia; 2015. 1,58: 87-97

[4] American Diabetes Association. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes–2021; Diabetes Care. 2021. Suppl 1, 44:S15-S33. https://doi.org/10.2337/dc21-S002

[5] Basto-Abreu, A., López-Olmedo, N., Rojas-Martínez, R., Aguilar-Salinas, C. A., De la Cruz-Góngora, V., Rivera-Dommarco, J., ... & Barrientos-Gutiérrez, T. Prevalence of diabetes and glycemic control in Mexico: national results from 2018 and 2020. salud pública de méxico. 2021. 63(6), 725-733.

[6] Caumo A., Bergman RN., Cobelli C. Insulin sensitivity from meal tolerance tests in normal subjects: a minimal model index; J. Clin. Endocrinol. Metab. 2000. 85 4396.

[7] Christen, J. A., Capistrán, M., Monroy, A., Alavez, S., Vargas, S. Q., Flores-Arguedas, H. A. & Kuschinski, N. A Diabetes minimal model for Oral Glucose Tolerance Tests. arXiv preprint arXiv:1601.04753. 2016.

[8] Donoho, D. 50 years of data science. Journal of Computational and Graphical Statistics. 2017. 26(4), 745-766.

[9] Erlandsen M., Martinussen C., Gravholt CH. Integrated model of insulin and glucose kinetics describing both hepatic glucose and pancreatic insulin regulation. Computer methods and programs in biomedicine. 2018. 156, 121-131.

[10] Foreman-Mackey D., Hogg D., Lang D., Goodman J. (Emcee: The MCMC Hammer. Publications of the Astronomical Society of the Pacific 125, no. 925: 2013. 306-12. doi:10.1086/670067.

[11] Gaucherel, C., Campillo, F., Misson, L., Guiot, J., & Boreux, J. J. Parameterization of a process-based tree-growth model: comparison of optimization, MCMC and particle filtering algorithms. Environmental Modelling & Software. 2008. 23(10-11), 1280-1288.

[12] Hardin, J. S., & Horton, N. J. Ensuring that mathematics is relevant in a world of data science. Notices of the AMS. 2017. 64(9), 986-990.

[13] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. The elements of statistical learning: data mining, inference, and prediction. New York: springer. 2009. Vol. 2, pp. 1-758.

[14] Hernandez-Aguirre A., Mendez-Davila HD., Moreles-Vazquez MA. What kernel size separates my data?. Proceedings of the Fifth Mexican International Conference in Computer Science, 2004. ENC 2004. IEEE, 2004.

[15] Heydari, M., Teimouri, M., Heshmati, Z. et al. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes Dev Ctries 36, 167?173. 2016. https://doi.org/10.1007/s13410-015-0374-4

[16] Hulman A., Vistisen D., Glümer C., Bergman M., Witte D, Færch K. Glucose patterns during an oral glucose tolerance test and associations with future diabetes, cardiovascular disease and all-cause mortality rate; Diabetologia. 2018. 1,61: 101-107

[17] Ismail HM., Xu P., Libman IM., Becker DJ., Marks JB., Skyler JS., Palmer JP., Sosenko J. Type 1 Diabetes TrialNet Study Group The shape of the glucose concentration curve during an oral glucose tolerance test predicts risk for type 1 diabetes; Diabetologia. 2018. 1, 61: 84-92

[18] Jagannathan R., Sevick MA., Fink D., Dankner R., Chetrit A., Roth J., Buysschaert M., Bergman M.; The 1-hour post-load glucose level is more effective than HbA1c for screening dysglycemia; Acta Diabetol 53: 543. 2016. https://doi.org/10.1007/s00592-015-0829-6

[19] James G., Witten D., Hastie T., Tibshirani R., An introduction to statistical learning (Vol. 112). New York: Springer. 2013

[20] Kaipio J., Somersalo E. Statistical and Computational Inverse Problems, Springe. 2004.

[21] Kanauchi, M., Kimura, K., Kanauchi, K., Saito, Y. Beta?cell function and insulin sensitivity contribute to the shape of plasma glucose curve during an oral glucose tolerance test in non-diabetic individuals. International journal of clinical practice. 2005. 59(4), 427-432.

[22] Khan ZAW., Vidyasagar S., Varma DM., Nandakrishna B., Holla A., Binu VS. The clinical and biochemical profiles of patients with IFG. International Journal of Diabetes in Developing Countries. 2019. 39(1), 94-99.

[23] Kuschinski, N. Statistical analysis of OGTT results. Doctoral dissertation, Ph. D thesis, Centro de Investigación en Matemáticas AC, Guanajuato, México. 2019.

[24] Levaillant, M., Lièvre, G., & Baert, G. Ending diabetes in Mexico. The Lancet. 2019. 394(10197), 467-468.

[25] Meza, R., Barrientos-Gutierrez, T., Rojas-Martinez, R., Reynoso-Noverón, N., Palacio-Mejia, L. S., Lazcano-Ponce, E., & Hernandez-Avila, M. Burden of type 2 diabetes in Mexico: past, current and future prevalence and incidence rates. Preventive medicine. 2015. 81, 445-450.

[26] Miao, H., Dykes, C., Demeter, L. M., & Wu, H. Differential equation modeling of HIV viral fitness experiments: model identification, model selection, and multimodel inference. Biometrics. 2009. 65(1), 292-300.

[27] Miao, H., Xia, X., Perelson, A. S., & Wu, H. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM review. 2011. 53(1), 3-39.

[28] Morbiducci, U., et al. "Predicting the metabolic condition after gestational diabetes mellitus from oral glucose tolerance test curves shape..^nnals of biomedical engineering. 2014. 42.5: 1112-1120.

[29] Nocedal, J., & Wright, S. J. (Eds.). Numerical optimization. New York, NY: Springer New York. 1999.

[30] Organización mundial de la salud. Diabetes. Recuperado en: https://www.who.int/es/news-room/fact-sheets/detail/diabetes

[31] Organización mundial de la salud. Global report on diabetes. Recuperado en: https://www.who.int/publications/i/item/9789241565257. 2016

[32] Palumbo P., Ditlevsen S., Bertuzzi A., De Gaetano A. Mathematical modeling of the glucose-insulin system: A review; Mathematical biosciences. 2013. 244(2), 69-81.

[33] Pillonetto G, Sparacino,G., Cobelli C. Numerical non-identifiability regions of the minimal model of glucose kinetics: superiority of Bayesian estimation; Math. Biosci. 2003. 184 , pp. 53-67

[34] PyMC3. Recuperado en: https://docs.pymc.io/en/v3/

[35] Rathee S. ODE models for the management of diabetes: A review; Int J Diabetes Dev Ctries. 2017. 37: 4. https://doi.org/10.1007/s13410-016-0475-8

[36] Rauf M., Sevil E., Ebru C., Cemil C. Early diagnosis of gestational diabetes mellitus during the first trimester of pregnancy based on the one-step approach of the International Association of Diabetes and Pregnancy Study Groups. International Journal of Diabetes in Developing Countries. 2018. 38(1), 20-25.

[37] Soto-Estrada, G., Moreno Altamirano, L., García-García, J. J., Ochoa Moreno, I., & Silberman, M. Trends in frequency of type 2 diabetes in Mexico and its relationship to dietary patterns and contextual factors. Gaceta sanitaria. 2018. 32, 283-290.

[38] Stuart A. M. Inverse problems: a Bayesian perspective. Acta Numerica. 2010. 19 : 451-559.

[39] Tschritter, O., Fritsche, A., Shirkavand, F., Machicao, F., Häring, H., Stumvoll, M. Assessing the shape of the glucose curve during an oral glucose tolerance test. Diabetes care. 2003. 26(4), 1026-1033.

[40] Vargas, P., Moreles, M. A., Pena, J., Monroy, A., & Alavez, S. Estimation and SVM classification of glucose-insulin model parameters from OGTT data: A comparison with the ADA criteria. International Journal of Diabetes in Developing Countries. 2021. 41(1), 54-62.

[41] Wieland, F. G., Hauber, A. L., Rosenblatt, M., Tönsing, C., & Timmer, J. On structural and practical identifiability. Current Opinion in Systems Biology. 2021.

[42] World Health Organization. Definition, diagnosis and classification of diabetes mellitus and its complications: report of a WHO consultation. Part 1, Diagnosis and classification of diabetes mellitus. 1999.

[43] World Health Organization (WHO). Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus. 2011. Recuperado en: `https://www.who.int/diabetes/publications/report-hba1c_2011.pdf`