Análisis de Datos y Boxplots Funcionales en la Clasificación de Perfiles Glucémicos

Ing. Nelson Ariza Morales
Centro de Investigación en Matemáticas. Unidad Monterrey
Email: nelson_ariza@cimat.mx

***Abstract*— Este estudio profundiza en el análisis de datos funcionales, con un enfoque en su generación y aplicación en la medicina, particularmente en la interpretación de pruebas de tolerancia a la glucosa. Se explica cómo se originan datos a partir de mediciones continuas, y se presenta la implementación de boxplots funcionales para visualizar las variaciones en estos complejos conjuntos de datos. Además, se explora el uso del clustering funcional, aplicando el método K-means adaptado para identificar patrones en las respuestas glucémicas. Los hallazgos subrayan la eficacia de técnicas estadísticas avanzadas en el análisis de datos funcionales, sentando una base firme para investigaciones futuras y su aplicación en el ámbito medico.**

Palabras clave: Datos Funcionales, Profundidad de curva, Boxplot, Clustering, Medicina.

I. Introducción

El análisis de datos funcionales, que trata con estructuras de datos donde cada observación es una función, curva o trayectoria, es esencial en disciplinas que abordan fenómenos dinámicos. Este tipo de datos es prevalente en áreas tan diversas como la biomedicina, donde se monitorean respuestas fisiológicas, la meteorología y la economía.

Para analizar datos funcionales, es fundamental emplear modelos matemáticos adecuados. Ramsay y Silverman (2005) han proporcionado métodos paramétricos, mientras que Ferraty y Vieu (2006) han desarrollado técnicas no paramétricas detalladas. Adicionalmente, la regresión cuantílica ha encontrado aplicaciones extensas en economía, como discuten Fitzenberger, Koenker y Machado (2002). Paralelamente, se han desarrollado métodos visuales que facilitan la representación y comprensión de los datos, resaltando sus características y revelando patrones interesantes (Hyndman y Shang, 2010).

En este estudio, nos centramos en la profundidad de banda para datos funcionales, como fue explorada recientemente por López-Pintado y Romo (2009), que permite ordenar una muestra de curvas desde el centro hacia afuera, introduciendo así una medida para definir cuantiles funcionales y evaluar la centralidad o atipicidad de una observación. Los boxplots funcionales, que emergen de este enfoque, ofrecen una herramienta de visualización atractiva y efectiva para los datos funcionales, extendiendo el boxplot clásico al dominio funcional (Sun y Genton, 2011).

Este artículo aborda la construcción de boxplots funcionales y boxplots funcionales mejorados, así como la regla de detección de valores atípicos asociada, demostrando la capacidad de visualización de los boxplots funcionales cuando se aplican a datos de pruebas de tolerancia a la glucosa. Estas pruebas, que registran la respuesta del cuerpo a la glucosa, producen datos ideales para el análisis mediante técnicas de datos funcionales, permitiendo una mejor comprensión y manejo de condiciones relacionadas con la glucosa, como la diabetes y la prediabetes. Dentro de la organización se incluye la explicación de la definición de profundidad de banda para datos funcionales y su versión modificada, la construcción de boxplots funcionales, y la aplicación de técnicas de clustering, todo dentro del contexto de un estudio sobre tolerancia a la glucosa.

II. Marco Teórico

La riqueza de los datos funcionales reside en su capacidad para capturar y modelar procesos continuos. Autores como Jacques y Preda (2014) han articulado meticulosamente la base metodológica que sustenta el análisis de datos funcionales, describiendo cómo una variable funcional aleatoria X toma valores en un espacio infinito-dimensional. Este enfoque transforma un conjunto de observaciones $\{X_1, ..., X_n\}$ en instancias de un proceso estocástico $\{X_t\}_{t\in T}$ en un espacio de Hilbert L^2, usualmente definido sobre un intervalo temporal T.

La base del análisis de datos funcionales se centra en el modelo fundamental:

$$x_i = f(t_i) + \epsilon_i$$

donde y_i son las observaciones, $f(t_i)$ representa la función que deseamos estimar, y ϵ_i denota los errores de medición, asumiendo que $\epsilon_i \sim N(0, \sigma^2)$ y son independientes. Este modelo nos plantea el desafío de estimar $f(t)$ a partir de datos ruidosos y discretos. La estimación que se menciona de $f(t)$ se plantea como sigue:

$$f(t) = \sum_{j=1}^{K} c_j \phi_j(t)$$

Donde las ϕ_k son las funciones bases y c_j son sus coeficientes.

II-A. Representación en Bases de Funciones

El primer paso en la FDA es la reconstrucción de la forma funcional a partir de observaciones discretas, lo cual se logra mediante la expansión en bases. Una base de funciones $\{\phi_j\}$ permite representar cualquier función como una combinación lineal de estas bases.

Bases Monomiales

Las bases monomiales utilizan polinomios de la forma:

$$\Phi(t) = (1, t, t^2, \ldots, t^k)$$

Aunque simples, estas bases pueden ser numéricamente inestables, especialmente con observaciones no uniformemente espaciadas.

Bases de Fourier

Ideal para datos periódicos, las bases de Fourier se componen de senos y cosenos de frecuencia creciente:

$$\Phi(t) = (1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \ldots)$$

donde ω define el período de oscilación. Esta base es natural para describir datos periódicos como ciclos anuales de temperatura.

B-Splines

Las B-splines son especialmente útiles para datos no periódicos, ofreciendo una representación suave y flexible:

$$f(t) = \sum_{j=1}^{K} c_j \phi_j(t)$$

donde $\phi_j(t)$ son las funciones base B-spline y c_j son los coeficientes ajustables. Estas bases son capaces de adaptarse a la localidad y suavidad de los datos funcionales.

Estas bases transforman los datos discretos en funciones continuas que facilitan el análisis posterior, utilizando técnicas estadísticas avanzadas para explorar y entender los fenómenos representados por los datos.

Esta metodología fue propuesta por De Boor (1977) donde tiene en cuenta una serie de bases tipo $\phi_j(t)$ para cumplir con los siguientes puntos:

- Cada una de las $\phi_j(t)$ es una función spline definida por un orden m y una sucesión de nodos t.
- Cualquiera de las diferentes combinaciones lineales que hagamos con esta serie de funciones base $\phi_j(t)$ será un spline.
- Dentro de cualquiera función spline que haya sido definida por su orden m con nodos t, puede expresarse como una combinación lineal de este tipo de funciones.

Otras bases

Las wavelets también ofrecen alternativas valiosas, especialmente cuando los datos no asumen periodicidad o cuando se requieren propiedades como la compacidad local

en la representación.

$$\phi_{kj}(t) = 2^{k/2}\phi(2^k t - k)$$

Siendo j y k números enteros. El uso estratégico de estas bases nos permite abordar la complejidad de los datos funcionales, transformando efectivamente el espacio infinito-dimensional a una representación finita que facilita el análisis posterior mediante técnicas estadísticas avanzadas.

II-B. Estadísticas Descriptivas

Las técnicas estadísticas clásicas, como la media y la varianza, son fundamentales en el análisis de datos multivariantes y pueden adaptarse eficazmente al análisis de datos funcionales. En este contexto, estas medidas descriptivas adquieren una nueva dimensión al ser aplicadas a funciones completas en lugar de a conjuntos de datos escalares.

Media Funcional

La media funcional se define como el promedio de un conjunto de funciones y se calcula de la siguiente manera:

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$$

Esta expresión puede ser expandida en términos de una expansión en bases de funciones, lo que permite una representación más detallada:

$$\bar{x}(t) = \sum_{j=1}^{p} \bar{a}_j \phi_j(t)$$

donde $\bar{a}_j$ es el promedio de los coeficientes de la base $\phi_j(t)$ correspondientes a la j-ésima función base, calculado como:

$$\bar{a}_j = \frac{1}{n}\sum_{i=1}^{n} a_{ij}$$

Varianza Funcional

La varianza funcional mide la dispersión de las funciones alrededor de la media funcional y se define como:

$$Var_x(t) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i(t) - \bar{x}(t))^2$$

La desviación estándar funcional, entonces, es simplemente la raíz cuadrada de la varianza funcional.

Covarianza Funcional

La covarianza funcional evalúa cómo dos puntos en diferentes instancias temporales s y t covarían a lo largo de las funciones observadas. Se define la superficie de covarianza como:

$$C(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i(s) - \bar{x}(s))(x_i(t) - \bar{x}(t))$$

A partir de la cual se puede derivar la función de correlación:

$$r(s,t) = \frac{C(s,t)}{\sqrt{C(s,s)C(t,t)}}$$

Estas definiciones proporcionan un marco para describir la variabilidad y las interdependencias dentro de un conjunto de datos funcionales, extendiendo los conceptos de medidas de tendencia central y variación del análisis multivariante al análisis de datos funcionales.

Profundidad de banda

La profundidad de banda para datos funcionales es un concepto clave que permite ordenar curvas de muestra en un espacio funcional según su centralidad o atipicidad. Este enfoque fue significativamente avanzado por López-Pintado y Romo (2009) y Sun y Genton (2011), quienes proporcionaron métodos robustos para evaluar la centralidad de curvas en el análisis de datos funcionales.

Sun y Genton

Sun y Genton introducen una metodología para ordenar curvas funcionales basada en la profundidad de banda. Definen la profundidad de una curva $x_i(t)$ como la frecuencia con la que $x_i(t)$ queda dentro de las bandas formadas por otras curvas en la muestra:

$$BD_J(x_i, P) = \sum_{j=2}^{J} P\{G(x_i) \subset B(X_1, \ldots, X_j)\}$$

donde $B(X_1, \ldots, X_j)$ es una banda delimitada por j curvas aleatorias y $G(x_i)$ es el gráfico de la función $x_i(t)$. La

profundidad de banda muestra cuán representativa o central es una curva respecto al resto de la muestra.

La profundidad de banda $BD_J(x, P)$ para una curva específica $x(t)$ con respecto a una medida de probabilidad P es calculada considerando todas las posibles bandas formadas por j curvas aleatorias, donde J es un número fijo tal que $2 \leq J \leq n$ (el número total de curvas).

Esta metodología resulta ser muy robusta debido a que define la profundidad de banda para una curva en contar cuántas veces el gráfico de la curva objetivo $y(t)$ está completamente contenido dentro de las bandas formadas por combinaciones de otras curvas en el conjunto, lo que podría darnos en algún punto poca información ya que imaginemos por ejemplo que una curva esta dentro de todo el intervalo por mas del 90 % del tiempo y solo en un lapso sale, estaríamos descartando inmediatamente considerarla.

Lopez y Pintado

López-Pintado y Romo proponen una variante más flexible, la profundidad de banda modificada (MBD), que considera la proporción del dominio en el que una curva está dentro de la banda formada por otras curvas:

$$MBD_n^{(j)} = \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} \lambda_r\{A(x; x_{i_1}, \ldots, x_{i_j})\}$$

donde λ representa la medida de Lebesgue sobre el intervalo I, y la suma se extiende sobre todos los subconjuntos de j curvas.

La medida de Lebesgue, en el contexto de la MBD, se utiliza para medir çuánto"de la función $y(t)$ cae dentro de las bandas formadas por otras funciones en el conjunto de datos. No se trata solo de contar cuántos puntos caen dentro de las bandas (como se haría con una medida más simple, como la medida de conteo), sino de evaluar la proporción del intervalo total I durante el cual $y(t)$ permanece dentro de estas bandas.

La proporción del tiempo que $y(t)$ pasa en la banda para un conjunto específico de j curvas es $\lambda_r(y)$, que se calcula como la razón entre la medida de Lebesgue de los puntos donde $y(t)$ está dentro de la banda y la medida total de I:

$$\lambda_r(y) = \frac{\lambda(A_j(y))}{\lambda(I)}$$

Al dividir la medida de Lebesgue de la región donde $y(t)$ está dentro de la banda por la medida de Lebesgue del intervalo total I, se obtiene una proporción o porcentaje. Esta medida es particularmente útil para evitar empates excesivos y para destacar curvas que son consistentemente centrales a lo largo de su dominio.

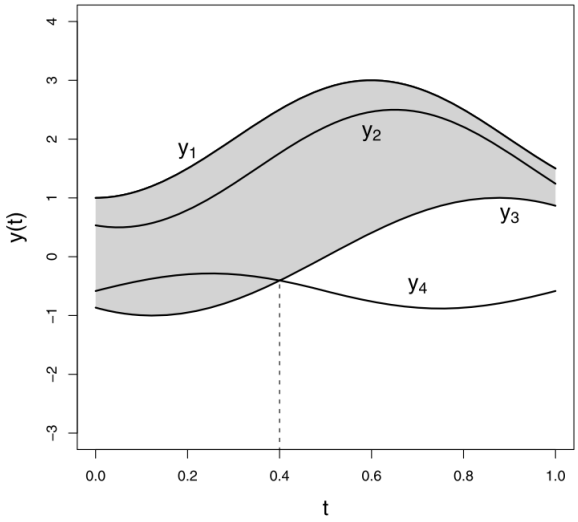

Fig. 1. Ejemplo de BD y BDM

Estos métodos proporcionan herramientas robustas para el análisis exploratorio de datos funcionales, permitiendo identificar curvas típicas y atípicas dentro de un conjunto de observaciones. La profundidad de banda ayuda a entender la estructura subyacente de los datos en un espacio funcional, facilitando la interpretación de fenómenos complejos representados por curvas. Esta metodología a diferencia de la de Sun&Genton, vemos que ahora si toma en consideración el caso previsto que se comentaba de la salida parcial de una curva en el caso de estudio que se este haciendo, ya que mide la proporción del tiempo que se mantiene dentro del intervalo.

Para dejar más claro esta diferencia, notemos en la figura 1 el mismo ejemplo, nos esta refieriendo que si estudiaramos la curva y_2 en el intervalo de las curvas y_1 y y_3 vemos que

si cumple la propuesta de BD, ya que esta completamente contenida, y sumaria a la profundidad de la curva, pero si trataramos la curva y_4 en el mismo intervalo, en las condiciones de BD, no se estaria considerando aporte a la profundidad de la curva, pero en la metodología BDM, si estariamos sumando porcentaje ya que parte de la curva si esta contenida en el intervalo.

II-C. Boxplot Funcional

La metodología de los boxplots funcionales extiende el concepto tradicional de boxplots a datos funcionales. Inspirados en la idea del rango intercuartílico (IQR) utilizado en estadísticas descriptivas clásicas, los boxplots funcionales adaptan estas nociones para manejar funciones completas en lugar de puntos discretos.

En el contexto de los datos funcionales, la "región central del 50 %" está formada por las curvas que caen dentro de los límites establecidos por la proporción α de las curvas más profundas del conjunto de datos. Para el caso del 50 %, se seleccionan las curvas que están entre el mínimo y el máximo de las $\lceil n/2 \rceil$ curvas más centrales, donde $\lceil n/2 \rceil$ es el menor entero no menor que la mitad del número total de curvas. La definición matemática de esta región central, basada en la profundidad de banda, es:

$$C_{0,5} = \{(t, y(t)) :$$

$$\min_{r=1,\ldots,\lfloor n/2 \rfloor} y[r](t) \leq y(t) \leq \max_{r=1,\ldots,\lfloor n/2 \rfloor} y[r](t)\}$$

Esta envolvente del 50 % actúa como el IQR en los boxplots tradicionales, proporcionando una indicación útil sobre la dispersión de la mitad central de las curvas, siendo robusta frente a valores atípicos y extremos.

Los 'bigotes' del boxplot funcional se extienden desde esta región central hasta el máximo alcance de los datos, excluyendo valores atípicos. Los valores atípicos se identifican utilizando una extensión del criterio IQR, similar a los boxplots clásicos:

$$Bigotes = C_{0,5} \pm 1{,}5 \times \text{IQR}_{\text{funcional}}$$

Curvas que se extienden más allá de estos 'Bigotes' son marcadas como potenciales valores atípicos.

En Python, la construcción de boxplots funcionales puede llevarse a cabo utilizando la biblioteca `scikit-fda`.

La adaptación de los boxplots a datos funcionales es crucial debido a la naturaleza inherente de los datos, que se componen de funciones completas en lugar de puntos discretos. Mientras que los boxplots convencionales son eficaces para resumir conjuntos de datos unidimensionales mediante estadísticas resumen como mediana, cuartiles y detección de valores atípicos, los boxplots funcionales manejan la complejidad añadida de los datos que varían continuamente.

En el análisis de datos funcionales, aplicar directamente boxplots convencionales a puntos discretos seleccionados de las funciones (como evaluar las funciones en tiempos específicos) puede resultar en una pérdida significativa de información. Esta aproximación ignora la correlación intrínseca y la estructura de dependencia entre los puntos evaluados y no captura adecuadamente la variabilidad global de las funciones.

II-D. Clustering Funcional

El análisis de clustering o agrupación es una colección de técnicas de clasificación no supervisada destinadas a agrupar objetos o segmentar conjuntos de datos en subconjuntos denominados clusters. Estos métodos buscan asignar objetos similares que comparten características comunes al mismo cluster. El clustering es esencial en muchos campos del análisis de datos funcionales, no solo para la exploración de datos sino también para clasificación e inferencia, como demuestran Flores, Lillo y Romo (2014) al utilizar medidas de profundidad para realizar pruebas de homogeneidad.

En datos funcionales, las métricas de distancia deben capturar diferencias en el comportamiento global de funciones completas en lugar de puntos discretos. La distancia L^p es comúnmente utilizada para este fin:

$$d(f,g) = \left(\int (f(t) - g(t))^p \, dt \right)^{1/p}$$

Para $p = 2$, esta distancia es una generalización de la distancia Euclídea al espacio de funciones, considerando la integral de las diferencias cuadradas entre funciones. Esta métrica es especialmente útil para datos funcionales porque considera la diferencia entre funciones en todo su dominio, proporcionando una medida integral de la disimilitud.

III. Implementación y resultados

Este estudio utiliza una base de datos que recoge resultados de la prueba de tolerancia oral a la glucosa (OGTT) realizada a pacientes durante 120 minutos. La OGTT evalúa cómo el cuerpo procesa la glucosa tras un ayuno nocturno, midiendo los niveles de glucosa en sangre en los tiempos 0, 30, 60, 90 y 120 minutos después de consumir una solución glucosada. Estas mediciones trazan curvas de respuesta glucémica que son cruciales para diagnosticar y entender las alteraciones en el metabolismo de la glucosa.

Además de las curvas de glucosa, la base de datos incluye variables que caracterizan a los pacientes, como:

- Índice de Masa Corporal (IMC): Indica la relación entre el peso y la altura del paciente.
- Edad: Esencial para analizar cómo la tolerancia a la glucosa varía con la edad.
- Condición Médica: Estado glucémico del paciente (normal, prediabetes, diabetes tipo 2).
- Variables Clínicas Adicionales: Presión arterial y niveles de colesterol, entre otros, que pueden influir en el metabolismo de la glucosa.

El análisis se centrará en las mediciones de glucosa, transformándolas en curvas funcionales mediante técnicas como splines. Este enfoque permite la aplicación de métodos estadísticos funcionales, incluidos boxplots funcionales y clustering funcional, para identificar patrones comunes y diferencias entre grupos de pacientes.

El primer paso como lo hemos comentado en el marco teórico es la construcción del dato funcional, analizamos los datos de 1496 pacientes, de los cuales mostramos el comportamiento de la glucosa de los 10 primeros en la figura 2.

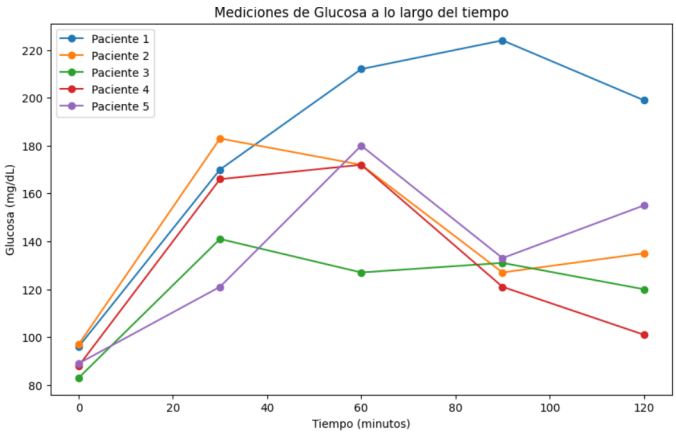

Fig. 2. Medición de tolerancia de glucosa a los primeros 5 pacientes

Como observamos, estos datos pueden ser codificados como datos funcionales ya que estan en función del tiempo todos los pacientes, para ello usamos B-Splines por su capacidad de soportar estos cambios de subida y bajadas. Es decir, aquellos datos que demuestran no ser periódicos.

Por lo que nuestros datos funcionales para nuestros mismos primeros 5 pacientes son los mostrados en la figura 3.

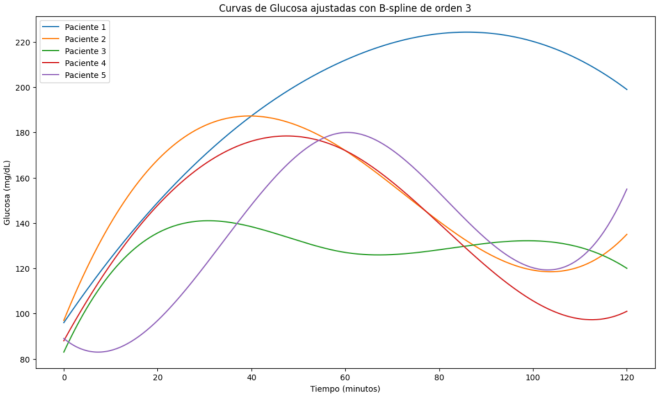

Fig. 3. Medición de tolerancia de glucosa en datos funcionales

Podemos notar que estos datos, siguen la misma tendencia a los datos mostrados en la figura 2. utilizando el grado 3 del polinomio utlizado. Dandonos como resultado un dato funcional el cual vamos a analizar con la metodología propuesta en

este artículo.

Al aplicar esta base al resto de nuestros datos funcionales, podemos llegar a la forma que se muestra en la figura 4.

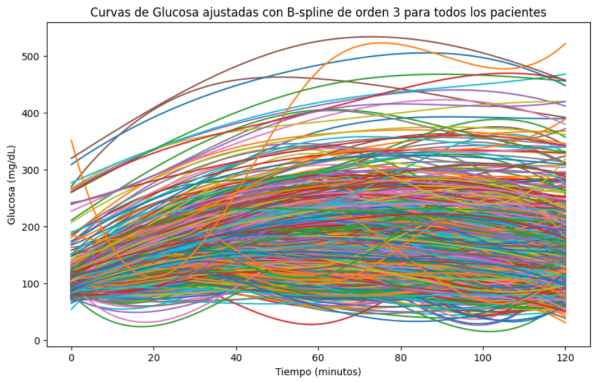

Fig. 4. Medición de tolerancia de glucosa en datos funcionales para todos los pacientes

A primera vista dentro de esta base de funciones, notamos ya unos pacientes que estan totalmente muy arriba de donde se concentran la mayor parte de los pacientes, pero hacer el análisis visual no es suficiente para afirmar esto, por lo que es necesario empezar a aplicar tenicas como el uso de Boxplot para identificar pacientes que pueden ser tomados como outliers, o que tienen condiciones totalmente diferentes a la mayoria de los pacientes. Por ello tomamos como primera opción hacer el un Boxplot a toda la base de datos, para analizar a primera vista, si existen estos posibles outliers.

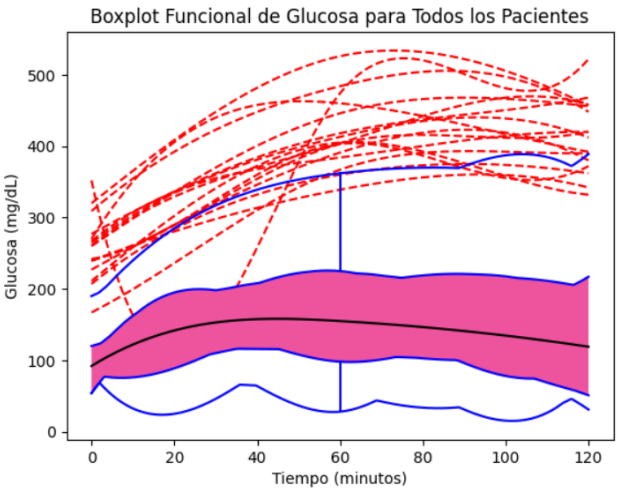

Fig. 5. Boxplot funcional para todos los pacientes

En la figura 5, en donde estamos analizando todos los datos sin hacer niguna de las particiones (Como se analizará mas adelante) notamos que existen bastantes valores atípicos, esto es aun impresionante sabiendo ya que se calcularon las profundidades de banda con respecto a todos los datos. Contamos que al respecto de los valoresa tipicos son 16 pacientes.

Resulta interesante, en los datos medicos que el comportamiento entre hombres y mujeres sea diferentes al momento de hacer un estudio, para este caso, se hace la hipotesis que es así y por eso se propone analizar los datos tanto en mujeres como en hombres, al filtrar la base de datos y analizarlos por separados, deberiamos esperar tener mejores predicciones al momento de querer encontrar valores atípicos.

Para esto primero analizamos el caso de los hombres, en la figura 6, para el caso de los hombres deberiamos esperar que tuviera menos ruido al momento de querer encontrar datos atipicos, debido a que estamos separando ya por generos, y vemos que para solo el caso de hombres encontramos 6 valores atipicos, que seria conveniente ahora realizar un estudio, para saber por que o si tienen condiciones especiales este tipo de pacientes.

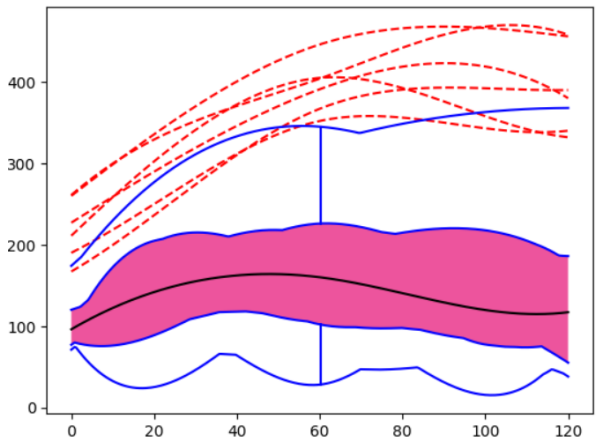

Fig. 6. Boxplot funcional para los pacientes hombres

Para intentar resolver la duda anterior, atendiendo a los valores atipicos podemos hacer una visualización a solo estos datos y alguna de sus condiciones que puedan ser causa de

que salieran irregulares estos pacientes.

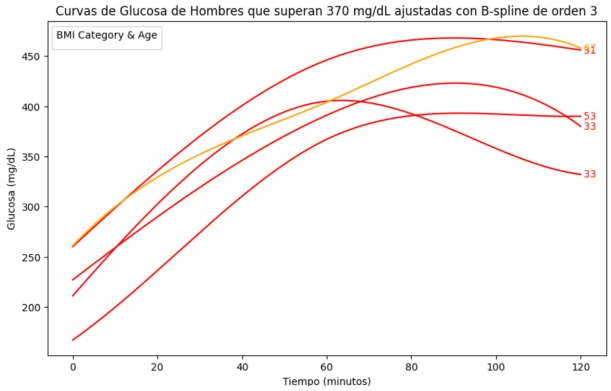

Fig. 7. Exploración de pacienes hombres que superan 370 mg/dL de glucosa, con variables claves

Para el supuesto anterior, hemos recurrido a solo haccer un analisis de dos variables que clinicamente pueden afectar al rendimiento de un paciente durante esta prueba como lo es la Edad y el BMI, que es la condicion fisica del paciente o su indice de masa corporal. Para inciar el analisis, se introdujeron colores segun la categoría que se encuentran segun el indice BMI, 'Menor que 18.5' como 'Peso bajo', 'Entre 18.5 y 24.9' es 'Normal', 'Entre 25 y 29.9' es ' Sobrepeso' y 'Mayor que 30' como 'Obesidad', entonces vemos en la figura 7, que en estos outliers que nosotros estamos viendo todos resultan tener una condicion fisica de obesidad, lo que pudiera ser causante de su comportamiento diferente a los demas pacientes.

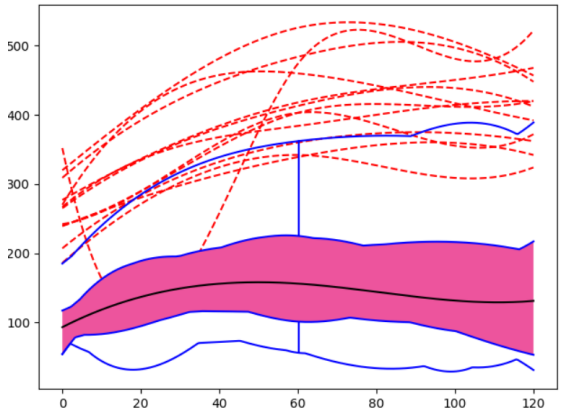

Fig. 8. Boxplot funcional para los pacientes mujeres

Para el caso de las mujeres, como podemos verlo en la figura 8, podemos observar de la misma forma que nos encontramos con valores atípicos, pero resultan ser mayores que los de hombres, por lo que seria importante hacer un analisis mas profundo como el que se hizo para hombre para dar una posible respuesta a su comportamiento de estos datos.

Siguiendo con el mismo análisis en mujeres podemos ver que en la figura 9, donde representamos los valores atípicos con algunas de sus variables medicamente representativas, la edad y su condición física. De primera instancia podemos encontrar que la mayor parte de estos pacientes con datos atípicos muestran un sobrepeso, además de que dejando de lado 2 de estos pacientes, la edad supera los 45 años, estas dos variables pueden ser factores importantes al momento de obtener los resultados clínicos para este estudio.

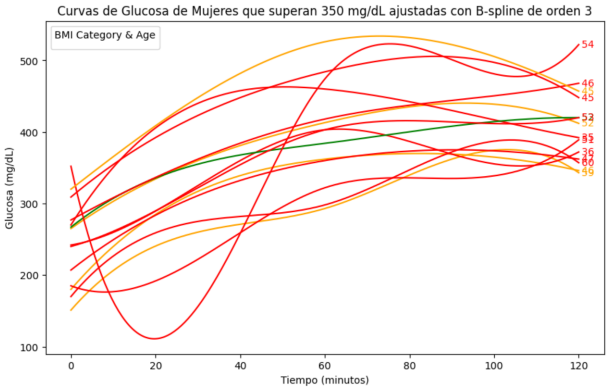

Fig. 9. Exploración de pacienes mujeres que superan 350 mg/dL de glucosa, con variables claves

Siguiendo el análisis basado en la construcción de los Boxplot, se propone identificar posibles irregularidades en pacientes que fueron catalogados con una de las 3 condiciones medicas que nos ofrece este estudio segun los resultados de la prueba:

- Normal
- Prediabetico
- Diabetes Tipo 2

Par este caso, se realiza ahora un filtro respetando este catalogo, y al momento de aplicar Boxplot funcional, obtenemos lo siguiente, donde los ejes representan la glucosa en el eje y,

y el tiempo de prueba transcurrido en el eje x:

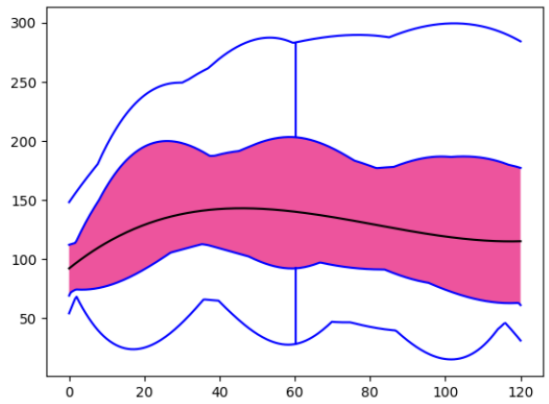

Fig. 10. Boxplot para pacientes con condición normal, en el eje y se muestra el puntaje de glucosa y en el eje x el tiempo transcurrido de la prueba

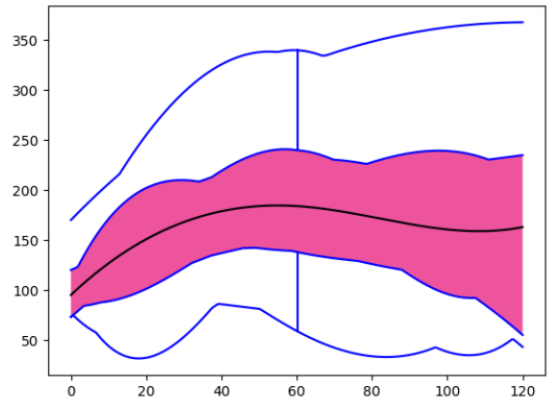

Fig. 11. Boxplot para pacientes con condición de Prediabeticos, en el eje y se muestra el puntaje de glucosa y en el eje x el tiempo transcurrido de la prueba

Como era de esperarse en este estudio por condiciones de los pacientes, los boxplot son mas regulares, es decir tienden a no tener datos atípicos y es claramente algo bueno, si no podria ser un error en tanto el trato o clasificación de los datos que se le esta dando al hacer la prueba.

Obtenemos algo interesante en el caso de pacientes clasificados como Diabetes tipo 2, se presentan 4 casos atípicos, ahora presentamos estos datos de diferente forma para ver que es lo que pasa con ellos, conociendo sus variables como edad y BMI. De primera vista vemos que tambien se presentan condiciones como para las mujeres, tanto fisicas como por la edad, los 4 pacientes presentan condiciones mínimas sobrepeso, ademas de que la edad promedio de estos pacientes es de 53 años. Los cuales podrian ser un indicador clave de su comportamiento sobre los demás pacientes con Diabetes Tipo 2.

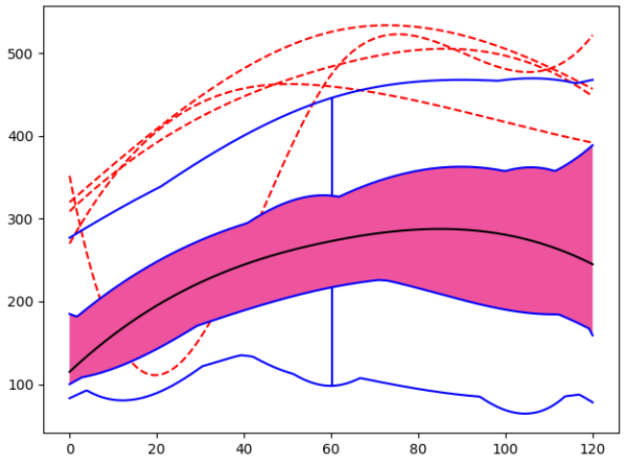

Fig. 12. Boxplot para pacientes con condición de Diabetes Tipo 2, en el eje y se muestra el puntaje de glucosa y en el eje x el tiempo transcurrido de la prueba

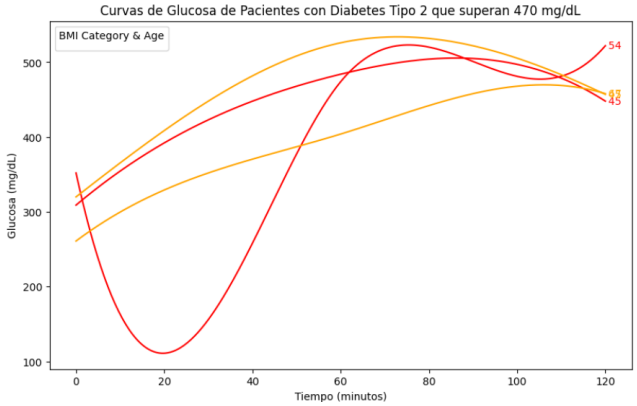

Fig. 13. Pacientes de la categoria Diabetes Tipo 2, con puntaje de glucosa mayor a 470 mg/dL, con variables como BMI y edad

Resulta importante, mencionar que estamos analizando datos clinicos y hay variables que no podemos controlar, tanto propias de cada persona como aquellas condiciones que se desarolla cada uno de los estudios, porque por más que hagamos el estudio siguiendo una metodología las condiciones propias del paciente y las del medico que aplica la muestra pueden ser variantes del resultado.

K means

Como se ha explicado, K means ayuda a agrupar de alguna forma objetos que tienen caracteristicas similares, creando un centro y midiendo distancias de los demas objetos para unirlos a un cluster, por ello, se ha propuesto esta metodologóia a fin de encontrar si existe este agrupamiento por condiciones físicas, ¿Por qué?, debido a que en el análisis de Boxplot logramos encontrar que por condiciones fisicas tenemos más regulado los pacientes atípicos. Por lo que resulta interesante, agruparlos por cluster para saber si existen caracteristicas propias que los definan como Normales, Prediabeticos y con Diabetes tipo 2.

Al aplicar a todos nuestras funciones de pacientes un clasificador Kmeans, se obtienen los siguientes centros.

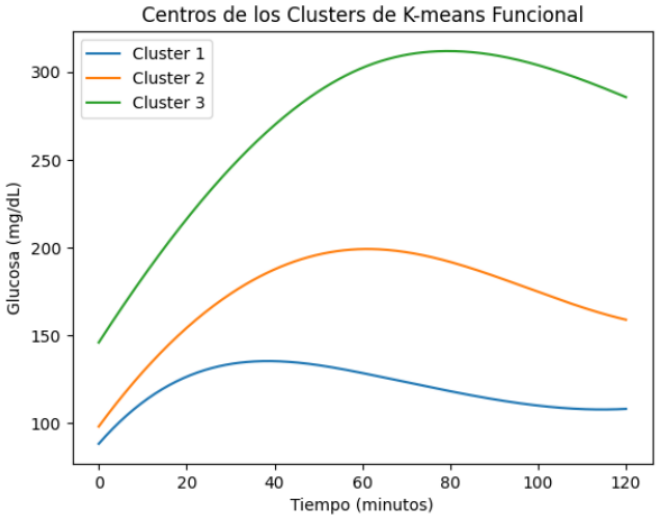

Fig. 14. Centros de los clusters con Kmeans a los datos funcionales.

Notemos una tendencia clara, los centros se dividen según la cantidad de glucosa que obtienen en la muestra, por lo que hemos forzado a ser 3 clusters, tratando de representar asi las condiciones médicas que pudieran clasificar a cada paciente. Para ello, entenderemos el cluster 1 como aquellos pacientes con condiciones Normales, el clusters 2 como la condición Prediabetes, y el clusters 3 como Diabetes tipo 2. Lamentablemente técnicas como Kmeans tienden a equilibrar la cantidad de objetos asignados a cada cluster, lo que es un problema para esta base de datos ya que tenemos un desbalance de 911 pacientes Normales, 481 Prediabeticos, y 104 con Diabetes tipo 2, lo que no ayuda al logaritmo clasificador. Aún asi presentamos la matriz de confusión la cúal nos dara mejor entendimiento de que es lo que pasa con cada clusters.

Observamos una tendencia clara a confundir entre los pacientes Prediabeticos con los pacientes normales, esto debido a su cercania de pacientes con una condición de la otra, lo cual resulta algo coherente hablando en terminos de salud, siguiendo la misma tendencia dentro de las condiciones de Prediabetes y Diabetes Tipo 2, también el algoritmo los suele confundir algunos de los casos, esto haciendo la propuesta que algunos pacientes estan ya mas cerca del los limites que los convierten a Diabeticos.

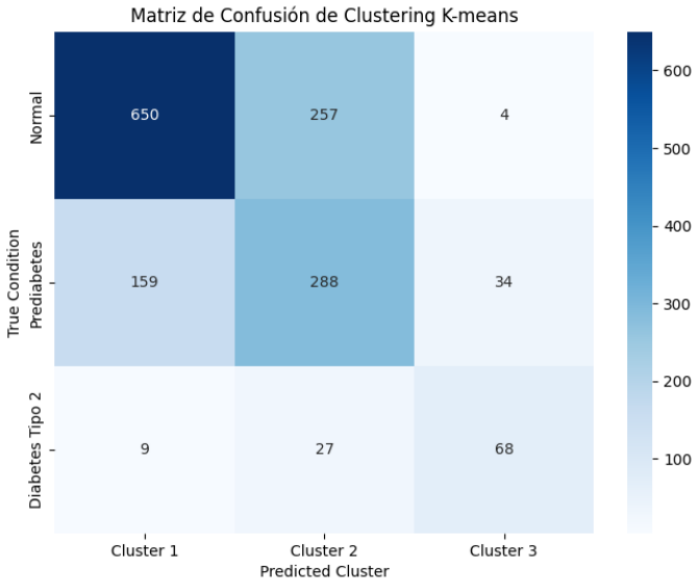

Fig. 15. Matriz de confusión comparando los pacientes asignados a cada clusters por kmeans con las etiquetas verdaderas de cada paciente

Estos resultados han demostrado tener un buen accuracy, conociendo las circunstancias llegando hasta un 68 % de precisión. Lo que nos da muestra que hay areas de oportunidad, como primero nivelar la base de datos, buscando realizar másа pruebas o tener acceso a una base de datos más completa, porque recordemos que estamos trabajando solo con una

sección de una base de datos extensa.

Segundo, el algoritmo Kmeans, no es muy bueno al trabajar con clusters desnivelados, lo que puede ser un desafio para el mismo algoritmo poder clasificarlos correctamente, lo ideal seria adaptar y probar otros algoritmos que sean capaces de abordar correctamente este desbalance entre clases de los datos. Lo cual seria beneficioso si quisieramos seguir explorando mayores posibilidades con esta misma base de datos.

IV. Conclusiones

Se ha realizado un fda, y se ha encontrado que las tecnicas aplicadas para los datos puntuales, estan también aplicadas para los datos funcionales, y esto nos ayuda a poder hacer un gran análisis, visto desde esta metodología para aquellos procesos dinamicos que dependan del tiempo o un espacio.

La construccion de los datos funcionales se ve afectada por el número de puntos en el espacio tiempo que estemos analizando, esto debido a que puede ser un problema el uso computacional, para el caso de estudios que se tomo, se consideraron 5 puntos de muestra para lo que la metodología demostro tener un buen rendimiento no teniendo tiempos de computo significativos, es importante hacer el análisis a una base de datos más grande para que podamos tener punto de comparación de este tipo de metodología.

También se ha demostrado como el Boxplot funcionales, han logrado darnos una variante para el analisis de datos atipicos, proporcionandonos mediante curvas un producto similar a los Boxplot tradicionales. Seguido de este análisis se planteo la implementación de clustering, donde se puede desmostrar que hay una adaptación que funciona bien de kmeans para datos funcionales, lamentablemente por problemas de balanceo de la base además de la propia naturaleza del problema por tener distancias cierto tipo de pacientes con otros.

V. Citas

- López-Pintado, S. y Romo, J. (2009). Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 53(4), 1562-1576.
- Sun, Y. y Genton, M. G. (2011). Functional Boxplots. *Journal of Computational and Graphical Statistics*, 20(2), 316-334.
- Ramsay, J.O., & Silverman, B.W. (2005). Functional Data Analysis (2nd ed.). Springer Series in Statistics. Springer.
- Hernando Bernabé, A. (2017). Development of a Python package for Functional Data Analysis. Universidad Autónoma de Madrid.
- Murillo González, L. (2021). Estudio de la evolución mundial del Covid-19 mediante análisis de datos funcionales. Trabajo de Fin de Máster, Universidad de Granada, Máster en Estadística Aplicada.
- Ferraty, F., & Vieu, P. (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer.
- Fitzenberger, B., Koenker, R., & Machado, J. A. F. (2002). Economic Applications of Quantile Regression. Journal of Economic Literature, 40(4), 1001-1046.
- Hyndman, R. J., & Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. Journal of Computational and Graphical Statistics, 19(1), 29-45.