



Clasificación de series de tiempo con datos funcionales y técnicas de Machine Learning: una aproximación para el Índice de Desarrollo Humano

Fredy Alexander Rivera Gómez

Universidad Santo Tomás
Facultad de Estadística
División de Ciencias Económicas y Administrativas
Bogotá, D.C., Colombia
2023

Clasificación de series de tiempo con datos funcionales y técnicas de Machine Learning: una aproximación para el Índice de Desarrollo Humano

Fredy Alexander Rivera Gómez

Tesis o trabajo de grado presentada(o) como requisito parcial para optar al título de:
Magister en Estadística Aplicada

Director(a):
Wilmer Pineda Rios Ph.D.

Línea de Investigación:
Nombrar la línea de investigación en la que enmarca la tesis o trabajo de investigación
Grupo de Investigación:
USTAdística

Universidad Santo Tomás
Facultad de Estadística
División de Ciencias Económicas y Administrativas
Bogotá, D.C., Colombia
2023

Dedicatoria

Principalmente doy gracias a Dios, por ser el inspirador y darme la fortaleza necesaria para continuar con este proceso y culminar con éxito un peldaño más de mi formación académica.

A mi familia, por su amor, paciencia, colaboración y sacrificio durante todo este tiempo que ha llevado mi formación intelectual. Gracias a ustedes he logrado llegar hasta aquí y convertirme en el profesional que hoy soy.

Agradecimientos

Agradezco a Dios por permitirme llevar a feliz término todos los estudios de mi carrera profesional.

Agradezco a la Universidad Santo Tomás por abrirme las puertas en su seno científico, y de esta manera llevar a cabo este proceso de maestría, brindándome la oportunidad de avanzar en mis proyectos personales.

A los docentes que me brindaron sus conocimientos, tiempo y su apoyo para seguir adelante día a día.

A mi asesor de tesis Wilmer Pineda por haberme permitido contar con su capacidad intelectual y conocimiento académico. Asimismo, agradezco por haber tenido toda la disposición y paciencia para guiarme durante el desarrollo de este proyecto de grado.

También agradezco a todos los que fueron mis compañeros de clase durante la maestría, ya que su compañerismo y colaboración, aportaron en gran medida al desarrollo de muchas actividades, que con entusiasmo se llevaron adelante en este proceso de formación.

Ante todo, deseo expresar agradecimiento a mi familia, por apoyarme cuando más lo necesité, y por extender su mano en momentos difíciles y brindarme su amor cada día.

Finalmente agradezco a quien lee este apartado y más de este proyecto, del cual espero sirva como fuente de conocimiento y que pueda contribuir al desarrollo de otros trabajos de grado o estudios relacionados; incluyendo este folio como parte importante de su repertorio de información.

Resumen

La desigualdad y el desarrollo humano son dos aspectos sociales que han perdurado a lo largo del tiempo, captando el interés de naciones y diversos organismos internacionales en la medición del progreso y desarrollo de la humanidad. La Organización de las Naciones Unidas, como entidad de cooperación internacional dedicada al diseño soluciones integrales para asegurar el desarrollo responsable de las naciones, estableció en 1990 el Índice de Desarrollo Humano (IDH), a través del Programa de las Naciones Unidas para el Desarrollo (PNUD). Este índice se erige como un instrumento para evaluar el avance de la humanidad. En este trabajo se desarrolla lo necesario para llevar a cabo la aplicación de la técnica de clasificación estadística con datos funcionales orientada a caracterizar funciones de datos del índice de desarrollo humano y evaluar la eficiencia de esta técnica aplicando diversos modelos de clasificación con este enfoque, empleando el entorno R y algunos de los paquetes disponibles en la actualidad **fda.usc** (Febrero-Bande and Oviedo de la Fuente, 2012), **classiFunc** (Maierhofer and Pfisterer, 2017), entre otros.

Palabras clave: Clasificación, series de tiempo, datos funcionales, machine learning funcional, desarrollo humano, FGLM, FGSAM, FGKAM, análisis de componentes principales funcionales (ACPF).

Abstract

Inequality and human development are two social aspects that have endured over time, capturing the interest of nations and various international organizations in measuring the progress and development of humanity. The United Nations Organization, as an international cooperation entity dedicated to designing comprehensive solutions to ensure the responsible development of nations, established the Human Development Index (HDI) in 1990, through the United Nations Development Program (UNDP). This index stands as an instrument to evaluate the progress of humanity. In this work, what is necessary to carry out the application of statistical classification techniques with functional data is developed, in order to characterize data functions of the human development index and evaluate the efficiency of this technique under different classification models with this approach. For this, the R environment is used and the different packages currently available ‘fda.usc (Febrero-Bande and Oviedo de la Fuente, 2012), classiFunc ‘(Maierhofer and Pfisterer, 2017), among others.

Keywords: Classification, time series, functional data, functional machine learning, human development, FGLM, FGSAM, FGKAM, functional principal component analysis (FPCA)

Lista de Figuras

2-1	<i>Tasa de mortalidad en un año (Por cada 1000 personas)</i>	8
2-2	<i>Esperanza de vida al nacer (Años)</i>	9
2-3	<i>Relación INB per cápita, años promedio de escolaridad y esperanza de vida.</i>	11
3-1	<i>Ejemplo de temperaturas en Canadá</i>	16
3-2	<i>Expectativa de los años de escolaridad a nivel mundial</i>	16
3-3	<i>KNN - Dinamic time warping</i>	23
3-4	<i>Dinamic time warping</i>	23
3-5	<i>Máquina de soporte de vectores.</i>	28
4-1	<i>Tasas de depuración de registros por conjunto de datos</i>	34
4-2	<i>Datos funcionales - Índice de desarrollo humano (IDH)</i>	40
4-3	<i>Índice de educación</i>	41
4-4	<i>Expectativa de años de escolaridad</i>	41
4-5	<i>Esperanza de vida al nacer</i>	42
4-6	<i>Ingreso nacional bruto per cápita</i>	42
4-7	<i>Bagplot funcional</i>	44
4-8	<i>Análisis funcional de Componentes Principales</i>	45
4-9	<i>Método de Silueta</i>	46
4-10	<i>Método Elbow</i>	46
4-11	<i>Método de Gap Statistic</i>	47
4-12	<i>No. Óptimo de clústeres</i>	47
4-13	<i>Clústers Índice de desarrollo humano</i>	48
4-14	<i>Distribución geográfica de los clústers</i>	49
4-15	<i>Clasificación Índice de desarrollo humano (IDH)</i>	50
4-16	<i>Boxplot Índice de desarrollo humano (IDH)</i>	51
4-17	<i>Boxplot Índice de educación</i>	53
4-18	<i>Boxplot Expectativa de años de escolaridad</i>	54
4-19	<i>Boxplot usuarios de internet</i>	55
4-20	<i>Boxplot Esperanza de vida al nacer</i>	56
4-22	<i>Geografía de clústeres.</i>	56
4-21	<i>Boxplot Desigualdad en la expectativa de vida</i>	57
4-23	<i>Geografía mortalidad infantil ONU.</i>	57
4-24	<i>Boxplot tasa de mortalidad infantil</i>	59
4-25	<i>Boxplot tasa de mortalidad hombres adultos</i>	60
4-26	<i>Boxplot tasa de mortalidad mujeres adultas</i>	60

4-27	<i>Incidencia de tuberculosis</i>	61
4-28	<i>Boxplot Acceso a tecnologías y combustibles limpios para cocinar</i>	62
4-29	<i>Boxplot Producto Interno Bruto</i>	63
4-30	<i>Boxplot Ingreso Nacional Bruto</i>	64
4-31	<i>Boxplot Ratio de dependencia de jóvenes</i>	64
4-32	<i>Boxplot Empleo Vulnerable</i>	65
5-1	<i>Comparativo de eficiencia en técnicas clasificación empleadas</i>	81

Lista de Tablas

4-1	<i>Conjuntos de datos por categoría</i>	32
4-2	<i>Descripción de las covariables seleccionadas asociadas al IDH</i>	35
4-2	<i>Descripción de las covariables seleccionadas asociadas al IDH</i>	36
4-2	<i>Descripción de las covariables seleccionadas asociadas al IDH</i>	37
4-2	<i>Descripción de las covariables seleccionadas asociadas al IDH</i>	38
4-3	<i>Países con curvas atípicas</i>	43
4-4	<i>Tabla comparativa de correlaciones</i>	66
5-1	Algoritmos de clasificación empleados	69
5-2	Márgenes de precisión. Modelos funcionales KNN - NP Kernel por covariable funcional	71
5-3	Parámetros modelo FGLM	72
5-4	Parámetros modelo FGSAM	73
5-5	Parámetros modelo FGKAM	74

1 Introducción

El Índice de Desarrollo Humano (IDH), concebido por el Programa de las Naciones Unidas para el Desarrollo (PNUD), ha sido un instrumento fundamental para evaluar y comparar el bienestar humano en todo el mundo. Este indicador compuesto, que considera aspectos de salud, educación e ingresos económicos, ha impulsado investigaciones de gran relevancia en el ámbito socioeconómico en las últimas décadas. Sin embargo, el análisis tradicional del IDH se ha basado en enfoques estáticos y tabulares, lo que limita su capacidad para capturar las tendencias temporales y la variabilidad funcional inherente que tiene el desarrollo humano.

Este estudio involucra técnicas de clasificación con análisis de datos funcionales y la historia de los indicadores del índice de desarrollo humano. A lo largo de seis capítulos, se presenta las mejores prácticas y técnicas estadísticas que permiten abordar este desafío. En el Capítulo 2, se contextualiza sobre el IDH como un indicador crucial en la medición del desarrollo humano, destacando su evolución y relevancia global.

En el Capítulo 3, se presentan conceptos estadísticos clave a través de una revisión bibliográfica de las técnicas de análisis de datos más empleadas en el contexto de la clasificación de curvas funcionales. Se examinó diferentes enfoques y metodologías en el campo de la estadística y el análisis de datos funcionales. La revisión abarcó desde técnicas descriptivas que permiten una comprensión detallada de la variabilidad funcional hasta métodos avanzados diseñados para abordar la clasificación de funciones en un contexto multiclase.

Dentro de esta revisión, se exploró en detalle las técnicas descriptivas aplicadas a datos funcionales destacando su utilidad en la caracterización de curvas y funciones. Asimismo, se analizó los enfoques de modelado para series temporales funcionales, enfocándose en su capacidad para capturar patrones de evolución temporal en datos funcionales. Además, se revisó los métodos de análisis de datos funcionales que han demostrado ser efectivos para la clasificación, incluyendo la adaptación de algoritmos de clasificación tradicionales al contexto funcional, así como técnicas específicas para datos funcionales, como los clasificadores basados en distancias funcionales y los métodos de Machine Learning adaptados a datos funcionales. Esta revisión proporciona marco teórico sólido para la aplicación posterior de técnicas de clasificación en el análisis de datos funcionales a las covariables relacionadas con el cálculo del Índice de Desarrollo Humano.

El Capítulo 4 se centra en el entendimiento de los datos, En el proceso de este estudio, se llevó a cabo una etapa fundamental de entendimiento de los datos funcionales asociados al Índice de Desarrollo Humano (IDH). Inicialmente, se trabajó con un total de 52 conjuntos de

datos que contenían información funcional sobre múltiples dimensiones del IDH en distintos países. Para garantizar la calidad y relevancia de los datos, se llevó a cabo análisis exploratorios que permitieron identificar valores faltantes y posibles valores atípicos que permitieron evaluar la calidad de fuentes de datos a trabajar.

De igual forma, se desarrolló algunos análisis de correlación con el fin de encontrar las relaciones entre las diversas covariables funcionales y las funciones del IDH. Este proceso de depuración y selección de características culminó en la consolidación de 18 datasets con información de covariables funcionales más importantes asociadas al cálculo del IDH para llevar a cabo los análisis posteriores.

Además de la selección de variables, se llevó a cabo un análisis funcional de componentes principales (FPCA) y un análisis de clúster funcional. El FPCA permitió reducir la dimensionalidad de los datos funcionales e identificar patrones y estructuras subyacentes en las curvas del IDH. Posteriormente, con la información obtenida con el FPCA se llevó a cabo un análisis clúster que agrupó las curvas del IDH en tres clústeres distintos, proporcionando una clasificación de países en función de su desarrollo humano. Estos clústeres resultaron ser de gran utilidad para la aplicación de métodos de clasificación que se explican en detalle en el capítulo 5.

El Capítulo 5, titulado "Modelado con Datos Funcionales", se entregan los resultados obtenidos al aplicar técnicas de clasificación de datos funcionales a las curvas funcionales del IDH. En esta fase, se exploraron diversas metodologías de clasificación de datos funcionales mediante el uso de herramientas y paquetes disponibles en el entorno de programación R, para esto se empleó los paquetes identificados para datos funcionales disponibles como son: `classiFunc`, `fda.usc`, y `FDboost`; que ofrecen una variedad de métodos de clasificación para abordar datos funcionales.

Se implementa tanto métodos de clasificación univariados como multivariados, permitiendo una evaluación del rendimiento de los modelos. Entre los métodos univariados se incluyeron el KNN funcional y el KNN kernel, que se basan en la proximidad funcional entre observaciones, aplicando para esto la deformación dinámica temporal (DTW).

Por otro lado, los métodos multivariados funcionales, como el Modelo funcional lineal generalizado (FGLM), el modelo funcional aditivo generalizado (FGSAM) y modelo funcional kernel aditivo generalizado (FGKAM); fueron utilizados para abordar la clasificación en contextos multidimensionales y evaluar la relación conjunta entre múltiples covariables funcionales. Además, se exploraron métodos de Machine learning adaptados a datos funcionales, como son: Partición recursiva, Máquinas de vectores de soporte, Random forest, red neuronal, GLM con regularización Lasso o Elasticnet y Gradient Boosting Machine, lo que permitió la comparación y selección de los enfoques más adecuados para clasificar las curvas del IDH.

Finalmente, Capítulo 6, presentamos nuestras conclusiones, resumiendo las contribuciones y los hallazgos clave de este estudio, y discutimos las implicaciones para futuras investigaciones en el campo del análisis de datos funcionales aplicado al IDH. Este trabajo

aborda un enfoque innovador para evaluar y comprender el desarrollo humano a través de un análisis detallado de datos funcionales y métodos de clasificación avanzados, proporcionando una base sólida para futuros avances en esta rama.

2 El desarrollo humano

La desigualdad, un fenómeno social arraigado a lo largo de la historia, se manifiesta de manera cada vez más evidente en muchos países alrededor del mundo. En estas naciones, es frecuente escuchar quejas de personas pertenecientes a diversas comunidades, ya que la desigualdad supera sus expectativas de vida. Esto se refleja en las limitadas oportunidades para la supervivencia, el acceso a servicios básicos, la posibilidad de tener vivienda y alimentación adecuadas, así como la capacidad de acceder a una educación superior. En la mayoría de los casos, estas limitaciones resultan en la incapacidad de obtener un ingreso digno, según lo señalado por el (PNUD, 2019).

Los hechos anteriores y sus consecuencias han representado un gran desafío para los entes internacionales como la organización de las naciones unidas (ONU). Con el fin de supervisar y mitigar sus consecuencias, la ONU ha dedicado esfuerzos sustanciales a desarrollar mecanismos que permitan cuantificar este fenómeno. Este compromiso se traduce en la creación del Índice de Desarrollo Humano (IDH) por parte del Programa de las Naciones Unidas para el Desarrollo (PNUD) en 1990, una iniciativa diseñada para orientar el establecimiento de políticas a medio y largo plazo. El propósito subyacente es mantener el progreso de las naciones más vulnerables, como lo subraya el informe del (PNUD, 2019).

El Índice de Desarrollo Humano (IDH) emerge como una herramienta fundamental para evaluar y comparar el progreso socioeconómico de las naciones en todo el mundo, a cargo del Programa de las Naciones Unidas para el Desarrollo (PNUD). El IDH va más allá de las métricas económicas tradicionales al considerar aspectos claves del bienestar humano, como la esperanza de vida, la educación y el nivel de vida. Este índice proporciona una visión integral de la calidad de vida en diferentes países, al centrarse en aspectos fundamentales que afectan directamente a las personas. A través del cálculo y la comparación de los valores del IDH, se puede determinar cómo las naciones están abordando cuestiones esenciales de salud, educación e ingresos, y los logros, así como los desafíos para avanzar hacia un desarrollo sostenible. El IDH se ha convertido en un referente global para medir y monitorear el progreso humano, proporcionando datos valiosos para la formulación de políticas, la toma de decisiones y la promoción de un desarrollo más equitativo y centrado en las personas en todo el planeta.

En el contexto de Colombia, el Índice de Desarrollo Humano (IDH) actúa como una herramienta clave de las dinámicas socioeconómicas y en las mejoras en la calidad de vida de su población diversa. El IDH colombiano también incorpora medidas de esperanza de vida al nacer, años promedio de escolaridad, años esperados de escolaridad y el ingreso nacional bruto per cápita.

La situación del Índice de Desarrollo Humano (IDH) en Colombia, evidencia una realidad compleja y en evolución constante. A lo largo de las últimas décadas, el país ha experimentado mejoras notables en varios indicadores clave del IDH, como la esperanza de vida y la educación. Sin embargo, persisten desafíos especialmente en términos de reducción de la desigualdad y diferencias regionales. Mientras las zonas urbanas han experimentado avances significativos en términos de desarrollo humano, las áreas rurales y apartadas enfrentan mayores obstáculos para acceder a servicios esenciales como la educación y la atención médica.

2.1. Índice de Desarrollo Humano (IDH)

El IDH fue diseñado para medir el progreso de cada país, teniendo en cuenta tres dimensiones básicas para el desarrollo del ser humano: salud, educación e ingresos; y cada dimensión es representada por mediciones que se realizan sobre aspectos esenciales de la vida del ser humano, como son: la esperanza de vida al nacer, la esperanza de escolarización y el ingreso per cápita, respectivamente. El IDH se expresa en una escala entre cero y uno, siendo 0 el menor desarrollo posible alcanzado por un país y 1 el mayor desarrollo que un país puede alcanzar. (López, 2019).

Para entender el cálculo del IDH es necesario comprender como se obtienen los resultados de los indicadores subyacentes que lo componen. En particular, el cálculo del IDH de un país se obtiene de la media geométrica de los resultados de los índices normalizados en cada una de las dimensiones (salud, educación e ingresos)(PNUD, 2020), y su expresión matemática es:

$$IDH = (I_{Salud} * I_{Educación} * I_{Ingresos})^{(1/3)}$$

donde,

I_{Salud} : Es el índice de la dimensión de salud

$I_{Educación}$: Es el índice de la dimensión de educación.

$I_{Ingresos}$: Es el índice de la dimensión de Ingresos.

2.1.1. Índice de esperanza de vida: índice de salud

Para el **DANE** la esperanza de vida corresponde al promedio de años que una persona puede vivir, siempre y cuando se mantengan las tendencias de mortalidad existentes en un determinado período. En este sentido, la esperanza de vida de un país se calcula teniendo en cuenta el máximo de años que una persona puede vivir y el mínimo en años que se considera

una persona puede vivir. De la misma manera que en el IDH, la escala para este indicador es entre 0 y 1, siendo 1 el máximo valor y 0 el valor considerado como mínimo.

La dimensión de salud I_{Salud} está definida por el cálculo del Índice de esperanza de vida, que se calcula utilizando la siguiente expresión:

$$IEV = \frac{IV - \min(EV)}{\max(EV) - \min(EV)}$$

siendo,

IEV: El índice de esperanza de vida

IV: El índice de vida real

$\max(EV)$: La edad máxima de años de esperanza de vida

$\min(EV)$: La edad mínima de años de esperanza de vida

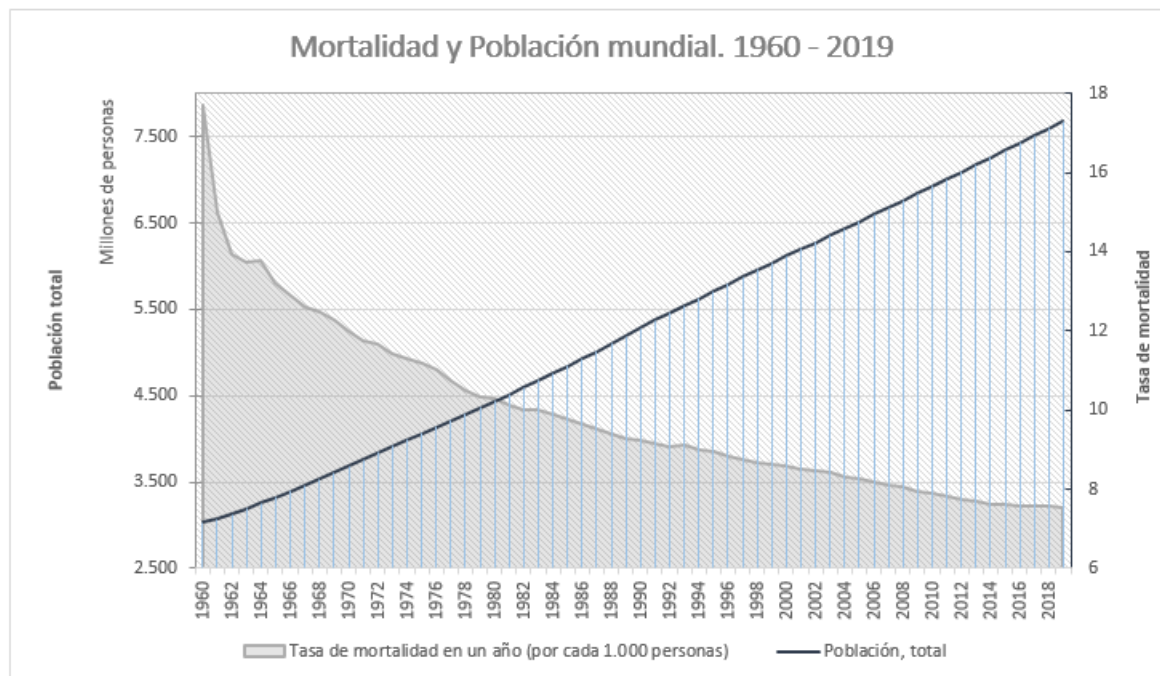
2.1.2. Esperanza de vida, tasas de mortalidad y crecimiento de la población.

La segunda guerra mundial fue un hecho sin precedentes que marcó el desarrollo del mundo; con la culminación de este importante hito en la historia de la humanidad, las tasas de mortalidad descendieron de manera vertiginosa y en consecuencia se dio un aumento acelerado de la población como se puede observar en la figura **2-1**.

De igual manera, los progresos en el ámbito médico han alcanzado el notable logro de incrementar la longevidad humana y reducir las tasas de mortalidad infantil. Como era de anticipar, las condiciones de vida para la especie humana han experimentado mejoras significativas. En consecuencia, el índice de esperanza de vida ha experimentado un aumento vertiginoso, generando como resultado un desafío en términos de densidad demográfica excesiva y, por ende, dando lugar al fenómeno de la sobrepoblación humana en ciertos países del mundo.

En el último siglo, la esperanza de vida se alargó en más de cinco años; según cifras del banco mundial la esperanza de vida para el año 2019 superó los 72 años, siendo el país de nacimiento una variable influyente en la esperanza de vida al nacer, que es mucho menor en algunos países del continente africano con condiciones sociales precarias (Mundial, 2021). Sin duda, la esperanza de vida está determinada por los niveles de mortalidad y estos a su vez se ven afectados por variables como el nivel económico, el acceso a la educación, las condiciones de salubridad, el acceso a los sistemas de seguridad social, el género y calentamiento global, entre otros.

Como se observa en la Figura **2-2**, el crecimiento de la población mundial es continuo

Figura 2-1: *Tasa de mortalidad en un año (Por cada 1000 personas)*

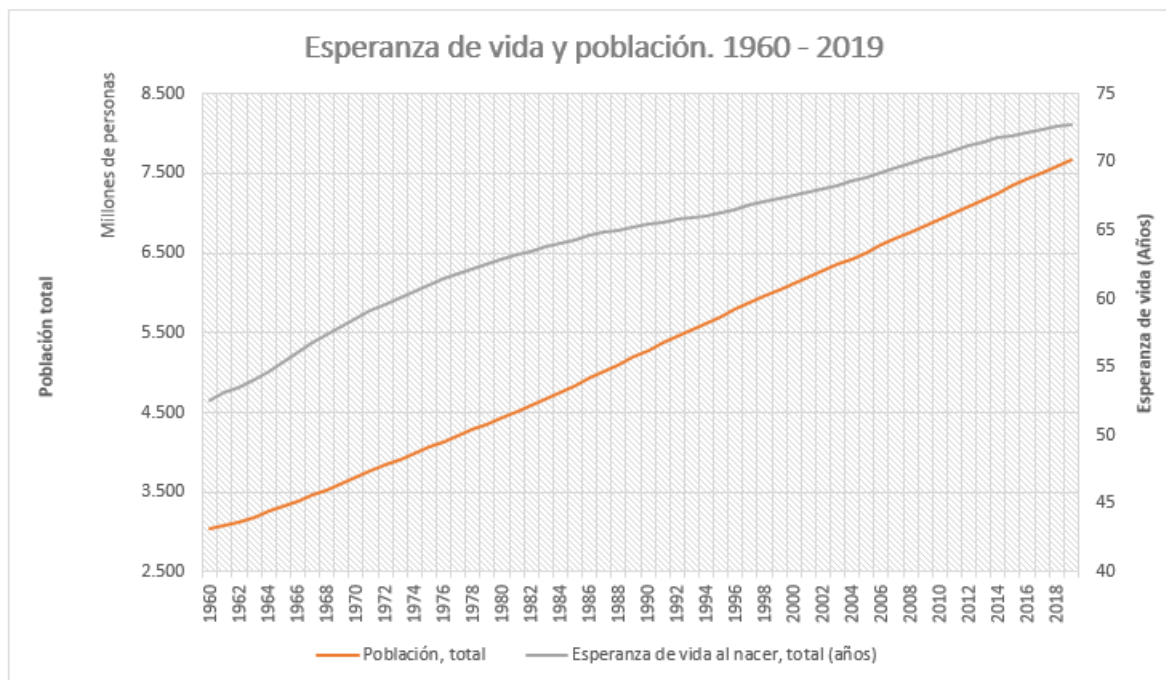
Nota: *Elaboración propia con datos tomados de Mundial (2023)*
<https://datos.bancomundial.org/>

y la esperanza de vida cada vez es mayor. En las últimas décadas, los gobiernos del mundo de una manera ortodoxa han enfocado esfuerzos en el desarrollo de políticas públicas dirigidas a controlar la fertilidad, obteniendo buenos resultados, pero los efectos de estas estrategias no han tenido los mejores frutos, dado que la población, aunque crece de una manera controlada se está envejeciendo. Para de las naciones unidas (2021), está previsto que la población mundial alcance los 8.500 millones de personas en 2030, 9.700 millones en 2050 y 11.200 millones en 2100. Por otra parte, se prevé que para 2050, una de cada seis personas en el mundo, tendrá más de 65 años.

2.1.3. Índice de esperanza de escolarización

Este índice busca medir el desarrollo educativo de un país y para ello compara esperanza en años de escolarización, frente a la escolaridad real alcanzada por un país. La dimensión de educación se mide a través de la escolaridad, cuya expresión se presenta a continuación:

$$IE = \frac{EE - \min(EE)}{\max(EE) - \min(EE)} + \frac{\bar{E} - \min(\bar{E})}{\max(\bar{E}) - \min(\bar{E})}$$

Figura 2-2: Esperanza de vida al nacer (Años)

Nota: *Elaboración propia con datos tomados de Mundial (2023)*
<https://datos.bancomundial.org/>

donde:

EE : Expectativa de años de escolaridad

$\bar{E}$: Años promedio de escolaridad

$\max(EE)$: Máximo de años de escolaridad

$\min(EE)$: Mínimo de años de escolaridad

La educación es un factor influyente en el desarrollo de un país. (PNUD, 2019), evidencia que las brechas en educación tienden a transferirse entre una generación y otra; además, quienes gozan de un determinado nivel de ingreso y educación, suelen casarse (o convivir) con parejas de una posición socioeconómica similar, y posterior a esto tienen hijos quienes son afectados por la situación socioeconómica de sus padres. De este modo en los países con mayor desigualdad, las brechas de educación tienden a reproducirse de manera intergeneracional limitando el desarrollo humano por décadas y hasta siglos.

2.1.4. Ingreso per cápita

Para la ONU, el nivel de vida se mide a través del Ingreso nacional bruto per cápita, en adelante INB per cápita, cuyos niveles son expresados en dólares con límites mínimos y máximos tratados en valores corrientes para este caso.

Entonces, la dimensión de ingresos se mide con el índice de ingresos (II) como sigue:

$$II = \frac{\log(\text{INB}_{\text{per cápita}}) - \log(\min(\text{INB}_{\text{per cápita}}))}{\log(\max(\text{INB}_{\text{per cápita}})) - \log(\min(\text{INB}_{\text{per cápita}}))}$$

donde:

II = Índice de ingresos

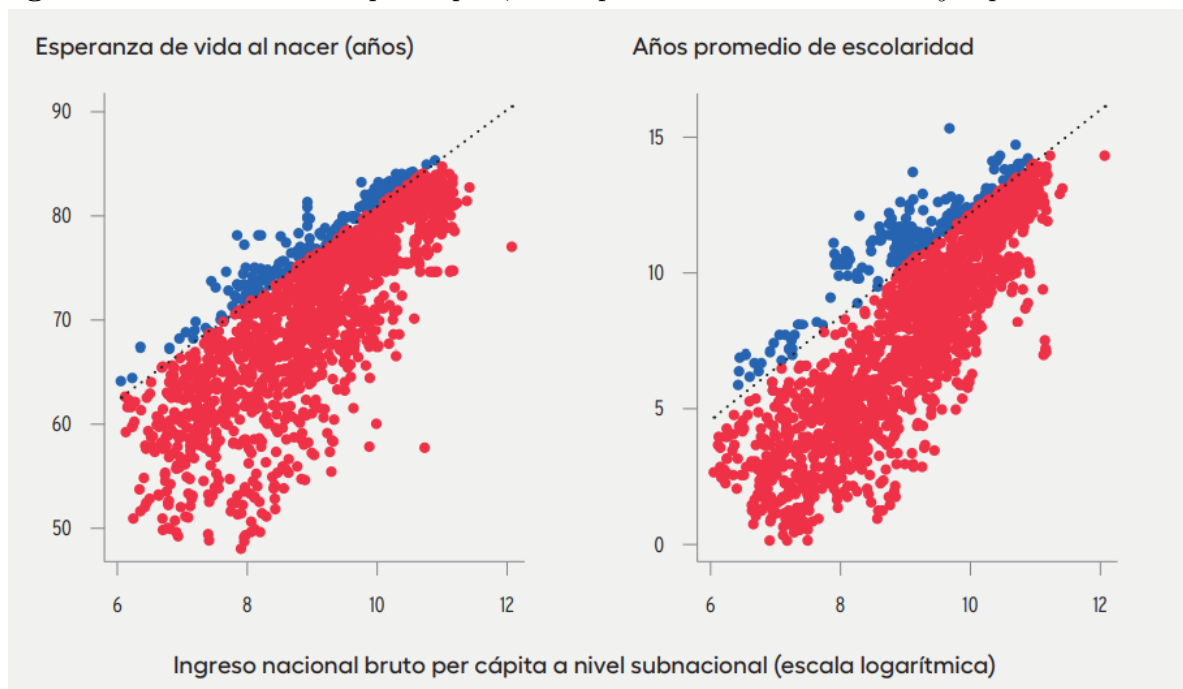
$\text{INB}_{\text{per cápita}}$ = Ingreso nacional bruto per cápita

Para muchos investigadores dedicados al estudio del bienestar del ser humano asociado con la distribución del ingreso en una sociedad, la medición de la desigualdad es de gran interés y como resultado se han planteado diferentes estrategias encaminadas a simplificar esta variable con el ánimo de tener resultados que permitan comparar la situación de los países en diversos puntos del tiempo. El INB per cápita es una variable diseñada en economía para medir el nivel de bienestar individual de una sociedad (Medina, 2001). En particular, el INB per cápita fue considerado por el PNUD como factor clave para análisis del desarrollo humano de un país, dada su estrecha relación con la escolaridad, el acceso a la salud, las tasas de mortalidad y las presiones que se ejercen sobre la situación ambiental del planeta.

Para (PNUD, 2020) cualquier nivel de ingreso muestra una amplia variación de los resultados en materia de salud y educación, haciendo hincapié, en que se podrían mejorar los dos tipos de resultados sin aumentar el ingreso como se evidencia en la figura 2-3). Existe un importante potencial en cada nivel de ingreso para aumentar el desarrollo humano mediante el cierre de las brechas en los logros relativos a la salud y la educación, lo que mejoraría la equidad en todas las dimensiones.

En todo caso, se observa como el ingreso guarda una relación directa con la esperanza de vida al nacer, así como con los años promedio de escolaridad, siendo como ya se mencionó, un determinante en el desarrollo humano.

Figura 2-3: *Relación INB per cápita, años promedio de escolaridad y esperanza de vida.*



Nota: Tomado de Programa de las Naciones Unidas para el Desarrollo PNUD (2020)

2.2. Métodos estadísticos aplicados a IDH y sus componentes

Como se ha visto, la necesidad de monitorear desarrollo humano mediante el cálculo correcto de sus indicadores es relevante. Los gobiernos y los organismos internacionales deben incorporar técnicas estadísticas y modelos avanzados, que permitan modelar el comportamiento de dichos indicadores con una baja incertidumbre y así dar inicio a la formulación temprana de estrategias que permitan mitigar impactos sociales, económicos y ambientales negativos, mejorando los niveles de vida para los países más vulnerables.

Aunque el interés de muchos investigadores en esta área siempre ha estado vigente, la atención frente al IDH es cada vez mayor si se tiene en cuenta que esta problemática es de connotación mundial. La literatura al respecto muestra una variedad de estudios que van desde la aplicación de técnicas estadísticas clásicas de estimación, hasta la elaboración de modelos robustos para predecir y/o modelar el comportamiento de la población, las tasas de mortalidad, la esperanza de vida, las tasas de fecundidad, entre otras variables importantes; sin embargo, obtener resultados precisos sigue siendo todo un desafío.

En el año 1992 Lee y Carter presentaron su modelo para hacer predicciones de las tasas de mortalidad de los Estados Unidos. En el modelo se combinó el uso de métodos estadísticos de series de tiempo y la distribución de tasas de mortalidad en función de la

edad. De acuerdo con (Lee, 2000) el método describe el logaritmo de una serie temporal de tasas de mortalidad específicas por edad como la suma de un componente específico por edad que es independiente del tiempo y otro componente que es el producto de un parámetro variable en el tiempo que refleja el nivel general de mortalidad, posteriormente, la estimación del parámetro variable en el tiempo se modeló y pronosticó como una serie de tiempo estocástica utilizando métodos estándar.

El método de (Lee and Carter, 1992) puede representarse como:

$$\ln(m_{x,t}) = a_x + b_x k_t + \epsilon_{x,t}$$

donde:

$m_{x,t}$: Tasa de muertes promedio de la edad x en el periodo t .

a_x : Parámetro de forma del perfil de edad x (intercepto).

b_x : Parámetro de cambio del perfil de edad x (pendiente).

k_t : Índice de mortalidad en el periodo t .

$\epsilon_{x,t}$: Error residual de la edad x y el periodo t

En 2001 Ronald Lee y Timothy Miller, efectuaron una variación al método Lee-Carter, haciendo una revisión y un diagnóstico sobre el método original de Lee-Carter diseñado en 1992. En la evaluación se demostró que el método se comporta mejor cuando se ajusta k_t al índice de esperanza de vida al nacer.

En (Akinbobola and Saibu, 2004) mediante un modelo vectorial autorregresivo que utilizó datos de Nigeria, se demostró que una tasa de desempleo en descenso mejora el desarrollo humano y en consecuencia se reduce la pobreza, así mismo, se evidenció una relación inversa que existe entre el incremento en el gasto público y la tasa de desempleo.

Más adelante (Hyndman and Booth, 2008) a través de análisis de datos funcionales con coeficientes de series de tiempo, modelaron el logaritmo de mortalidad específica por edad y logaritmo para la tasa de fertilidad, que posteriormente, fue ampliado por (Hyndman and Shang, 2009), en su investigación para modelar el logaritmo de la expectativa de vida específica por edad, la migración neta y el tamaño de la población de Australia.

En (Shang et al., 2013), se propone un enfoque de análisis de datos funcionales para llevar a cabo predicciones de mortalidad, fecundidad, emigración e inmigración, con el propósito de obtener previsiones de población, a partir de datos históricos de la población de reino unido. El método lleva a cabo la aproximación de las funciones a través de componentes principales funcionales y aplica transformaciones de Box-Cox a cada componente para eliminar la heterocedasticidad. Para los autores, el método de análisis de datos funcionales

multinivel es versátil en cuanto a que permite modelar múltiples conjuntos de funciones que pueden correlacionarse entre grupos, adaptando esta técnica para el análisis de una muestra de funciones de dos subpoblaciones que están correlacionadas.

Además, como complemento al modelo proponen la optimización los parámetros utilizando el algoritmo Bayesiano de muestreo de la cadena de Markov Monte Carlo aprovechando la conexión que para Staicu (2010) existe entre el modelo funcional multinivel y los métodos jerárquicos.

Por otra parte, en (Shang, 2015) se presenta la aplicación de una variedad de métodos para predecir la esperanza de vida al nacer y las tasas de mortalidad en 20 países desarrollados. En el estudio se hace uso del método Lee-Carter (LC), el método de Lee-Miller (LM), el método caminata aleatoria con derivada (RWD), el método Hyndman-Ullah (HU) y el método Hyndman-Ullah ponderado (HUw), la información contenía datos específicos de sexo y edad. Finalmente (Shang, 2015) muestra los resultados arrojados por cada uno de los métodos y las evidencias de que el método Lee-Miller y el método de caminata aleatoria arrojaron los pronósticos puntuales y de intervalo más precisos sobre las expectativas de vida específicas por edad en mujeres y hombres, respectivamente para cada modelo.

Luego en (Shang, 2016) se emplea un método de datos funcionales multinivel con el objetivo de hacer una previsión de mortalidad y esperanza de vida sobre una muestra de países desarrollados, utilizando datos específicos por edad y sexo. En el estudio se sugiere suavizar las tasas de mortalidad a través de la aplicación de splines de regresión penalizados; en general los métodos de datos funcionales multinivel demostraron producir mejores resultados que los métodos clásicos de Lee Carter y Li-lee, y de igual forma los resultados sobrepasaron a los métodos funcionales para análisis independientes. Adicionalmente, el autor aplica un modelo Bayesiano jerárquico para estimar los resultados de esperanza de vida por sexo.

(Deprez et al., 2017) en su trabajo aplica técnicas de Machine Learning y modelos clásicos para predecir las tasas de mortalidad en Suiza. Dentro de las técnicas clásicas empleadas se encuentran los modelos de Lee-Carter y el modelo Renshaw – Haberman, además se efectúan análisis con árboles de regresión empleando Gradient Boosting que es un método de ensamble de Machine Learning para árboles de decisión. Finalmente, en el estudio se buscó detectar las debilidades de los métodos clásicos, así como estimar las causas de mortalidad y sus probabilidades de ocurrencia haciendo uso de un modelo de probabilidades de Poisson.

Sin duda, los algoritmos de Machine Learning han sido ampliamente aplicados en diferentes campos en los últimos años, debido a los excelentes resultados que han mostrado por su gran poder predictivo. Asimismo, los datos funcionales han ganado terreno en el campo del Big data por su aplicabilidad en conjuntos de datos con gran tamaño, además de ser una técnica bastante empleada para predecir en datos demográficos (Shang et al., 2013). Las técnicas de Machine Learning con entradas funcionales son un campo en exploración, pero a pesar de esto, los estudios existentes muestran buenos aportes como se evidencia en (Pfisterer et al., 2019) que hace una comparación entre diversos algoritmos de Machine

Learning con entradas de datos funcionales y optimización bayesiana de hiperparámetros para la clasificación de series de tiempo con datos funcionales. El estudio demostró que las predicciones alcanzaron un 92 % de precisión y que la optimización de hiperparámetros aportó a la precisión un 5 % adicional.

Estudios más recientes (Khan et al., 2020), muestran como técnicas de Machine Learning son útiles para estimar muertes por COVID en datos de una muestra de países desarrollados, tales como, Turquía, España, Suecia, Francia y Pakistán. Las predicciones de tasas de mortalidad se elaboran a partir de casos confirmados por COVID, implementando una extensión del modelo de Lee-Carter sobre redes neuronales para múltiples poblaciones; en este caso la optimización de hiperparámetros fue realizada aplicando un proceso gaussiano de regresión.

3 Conceptos estadísticos

3.1. Análisis de datos funcionales

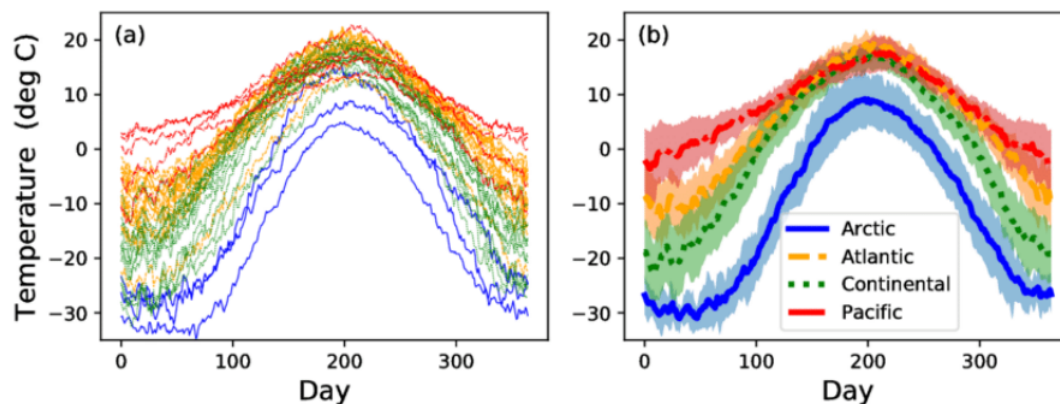
En esta sección se hace un recorrido, a través del marco teórico de los datos funcionales y los aspectos más importantes de esta técnica con el motivo de proporcionar las herramientas necesarias para el entendimiento de los análisis presentados posteriormente.

El análisis de datos funcionales es una técnica utilizada para estudiar la relación que existe entre una variable independiente y una o más variables dependientes, basándose en el análisis de funciones matemáticas para modelar y entender cómo cambian los datos a medida que varía la variable independiente. El término de datos funcionales se empezó a acuñar desde el año 1950 por Grenander en su publicación **Stochastic processes and statistical inference**, más tarde en 1958 Rao en **Some statistical methods for comparison of growth curves** retoma el termino, sin embargo quien empieza robustecer el método es Ramsay a partir de 1982.

Las aplicaciones de esta técnica son amplias y variadas, encontrándose desde la investigación científica hasta el análisis de datos en negocios para la toma de decisiones. En la investigación científica, el análisis de datos funcionales se utiliza para estudiar fenómenos complejos y entender cómo interactúan diferentes variables. El análisis estadístico de datos funcionales es un campo relativamente en desarrollo, siendo cada vez más utilizado en áreas en las que se necesita la construcción de modelos parsimoniosos, que faciliten el procesamiento y análisis de datos cuando se posee de grandes volúmenes de estos.

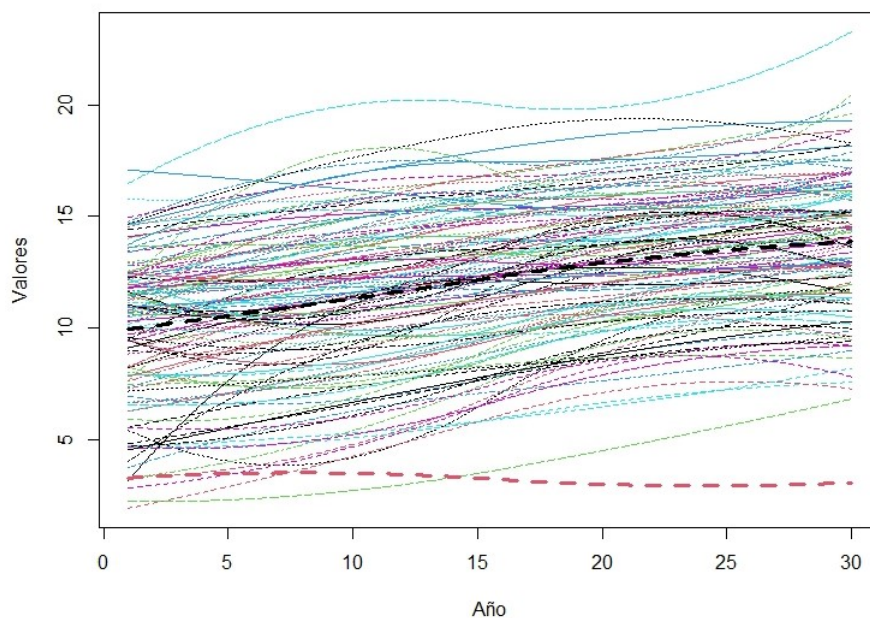
La potencia que tienen los datos funcionales se deriva en la búsqueda de poder estudiar fenómenos cuya estructura de datos es continua, como lo pueden ser los latidos del corazón, fenómenos como la precipitación, la temperatura y eventos relacionados con el tiempo y/o variables demográficas como el crecimiento de la población, los índices de vida y mortalidad en este caso. Algunos ejemplos se muestran en las Figuras **3-1** y **3-2**

Figura 3-1: *Ejemplo de temperaturas en Canadá*



Nota: Tomado de *Méthodes statistiques pour données fonctionnelles multifarious*.
(Golovkine, 2021)

Figura 3-2: *Expectativa de los años de escoraridad a nivel mundial*



Nota: Elaboración propia con datos tomados de
<https://www.datos.bancomundial.org>

Para (Ramsay and Silverman, 2005) los objetivos del análisis de datos funcionales son esencialmente los mismos que los de cualquier otra rama de la estadística, en ese sentido el método persigue hacer representación de los datos de manera que faciliten la realización de un

análisis más detallado, mostrar datos para resaltar sus características, identificar patrones y variaciones, y explicar resultados de una variable dependiente usando información de entrada disponible en las variables independientes.

El análisis de datos funcionales tiene su unidad mínima en la función, en la que cada individuo replica sus datos en una muestra aleatoria de funciones (Wang et al., 2015), en otras palabras, los datos funcionales surgen cuando los datos asociados a una variable pueden representarse como una curva suave o una función (Kokoszka and Reimherr, 2017).

Para (Kokoszka and Reimherr, 2017) la forma más simple de datos funcionales se refiere a la observación de N funciones en un intervalo continuo t , como el tiempo o el espacio. Donde el espacio de Hilbert $L^2(\tau)$ se define como el espacio de funciones cuadrado integrables en el que se dan las diferentes interacciones de un proceso estocástico, que se puede representar mediante la expresión:

$$\begin{aligned} X_n(t_{j,n}) &\in R, \quad t_{j,n} \in [T_1, T_2] \\ n &= 1, 2, \dots, N \\ j &= 1, 2, \dots, J_n \end{aligned} \tag{3-1}$$

Si X es un elemento aleatorio de $L^2(\tau)$, se dice que es cuadrado integrable si:

$$\|E\| = E\left[\int X^2(t)dt\right]^{1/2} < \infty \tag{3-2}$$

Al igual que en el enfoque clásico, el desarrollo del análisis de datos funcionales requiere de adoptar una estrategia que se podría caracterizar como exploratoria, confirmatoria o predictiva (Ramsay and Silverman, 2005).

”La diferencia entre análisis exploratorios y confirmatorios tiende a ser el grado en que se utiliza la teoría de la probabilidad, en el sentido de que la mayoría de los análisis confirmatorios se resumen en una o más declaraciones de probabilidad. Los estudios predictivos son algo menos comunes y se enfocan en usar los datos disponibles para hacer una declaración sobre estados no observados, como el futuro”.

Ramsay and Silverman (2005).

3.1.1. Métodos descriptivos para el análisis de datos funcionales

De la misma manera que en el contexto clásico, el enfoque funcional aplica estadísticas análogas de resumen para media y varianza.

Media funcional

En el enfoque funcional la **media puntual** funcional está dada por:

$$\bar{X}_N(t) = N^{-1} \sum_{i=1}^N X_i(t) \quad (3-3)$$

Varianza funcional

Por otra parte, la **varianza puntual** funcional, se define mediante la siguiente expresión,

$$var_X(t) = (N - 1)^{-1} \sum_{i=1}^N [X_i(t) - \bar{X}_N(t)]^2 \quad (3-4)$$

Covarianza y correlación

(Ramsay and Silverman, 2005), presentan en un grupo diferente a la covarianza y correlación, y las presentan como funciones que resumen la dependencia de los registros.

En el caso funcional, la **covarianza** se define como:

$$cov_x(t_1, t_2) = (N - 1)^{-1} \sum_{i=1}^N \{x_i(t_1) - \bar{x}(t_1)\} \{x_i(t_2) - \bar{x}(t_2)\} \quad (3-5)$$

Finalmente, la correlación funcional está asociada a la expresión análoga de la correlación de Pearson como sigue:

$$corr_x(t_1, t_2) = \frac{cov_x(t_1, t_2)}{\sqrt{var_x(t_1)var_y(t_2)}} \quad (3-6)$$

Sin embargo, identificar una relación lineal entre variables funcionales no siempre es posible, por tanto cobra relevancia medir la relación no lineal o correlación monótona existente entre funciones aplicando otro tipos de métodos, tales como, el coeficiente de correlación de Spearman definido por la siguiente ecuación:

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n} \quad (3-7)$$

3.1.2. Funciones básicas

Para (Ramsay and Silverman, 2005) el primer paso para trabajar con datos funcionales es poder expresarlos a través de una función básica de expansión $X(t) \in L^2$ como:

$$X(t) \approx \sum_{k=1}^K C_k \phi_k, \quad 1 \leq n \leq N \quad (3-8)$$

Donde, ϕ_k son funciones básicas del vector ϕ y los coeficientes C_k son elementos del vector C de largo K .

Para Ramsay and Silverman (2005) **”las funciones básicas son como ondas o splines, o funciones seno y coseno que comparten algunas propiedades de forma, por lo que pueden aproximarse como combinaciones lineales de algunas M formas básicas.”**

Las funciones B-spline son una técnica de suavizado de datos funcionales, que se utilizan para aproximar una serie de tiempo o una curva en un conjunto de puntos de control. Estas funciones se basan en la interpolación de polinomios de grado bajo entre los puntos de control, lo que permite suavizar los datos sin afectar significativamente la forma general de la curva. Es decir, que la suavización funcional es una técnica utilizada para eliminar el ruido de una serie de tiempo o una curva y su objetivo es el de obtener una representación más suave de los datos originales sin perder información importante.

Existen diferentes métodos de suavizado funcional tales como, el suavizado por spline, el suavizado por kernel y el suavizado mediante la transformada de Fourier. No obstante, en este estudio se empleará en concreto la base de B-splines propuesta por (de Boor, 1978) para la construcción de los datos funcionales.

3.1.3. Análisis de componentes principales para datos funcionales

La técnica de análisis de componentes principales (PCA) al igual que en la estadística clásica es un método que se emplea en datos funcionales con el fin de reducir la dimensionalidad de un conjunto de datos funcionales.

El objetivo principal de FPCA es encontrar el conjunto de componentes principales que explican la mayor variabilidad en los datos. Estos componentes son combinaciones lineales de las variables originales y son ordenados de manera tal que la primera componente principal explica la mayor cantidad de variabilidad, seguida por la segunda, y así sucesivamente.

El análisis de componentes principales de datos funcionales (FPCA), es una técnica considerada de las más importantes debido a la gran cantidad de bondades que presenta al

permitir reducir la dimensionalidad de los datos funcionales, detectar patrones que no son visibles fácilmente, además de permitir obtener una estructura de covarianza mucho más informativa que puede reemplazar inclusive por completo un resultado confuso obtenido a través de una análisis de la función de varianza y covarianza Ramsay and Silverman (2005). Es importante tener en cuenta que FPCA no preserva las relaciones temporales o espaciales en los datos funcionales, por lo que no es adecuado para todos los casos de uso.

La forma análoga del caso clásico de PCA para datos funcionales asume que para este caso los valores de las variables son comparables con los valores de las funciones.

En PCA multivariado el análisis produce un vector ponderación que es el que define la dirección de la variabilidad propia de cada componente, mientras que el enfoque funcional cada componente principal es proporcionada por una componente funcional de peso $\beta(s)$ definida en el mismo rango s del dato funcional Ramsay and Silverman (2005). Por lo tanto, los valores de los componentes principales están dados por f_i como sigue.

$$f_i = \int \beta x_i = \int \beta(s)x_i(s)ds \quad (3-9)$$

3.2. Series de tiempo y análisis de datos funcionales

Las series son un tipo común de datos funcionales y pueden ser consideradas como funciones discretas al largo del tiempo. En otras palabras, las series de tiempo existen cuando se observa el mismo objeto funcional durante un período de tiempo Pfisterer et al. (2019). La revisión de la literatura permite evidenciar que la teoría de series de tiempo funcionales es incipiente, ahora bien, se han estudiado los modelos autorregresivos análogos a los modelos ARMA(p, q) para datos escalares, tal como, el modelo ARCH(1):

$$Y_s(t) = \int \psi(s, r)Y_{t-1}(r)dr + \epsilon_t(S) \quad (3-10)$$

En 2009 (Hyndman and Shang, 2009), se propone un método para pronosticar series de tiempo; el método plantea el uso de análisis funcional de componentes principales; para descomponer la serie de tiempo funcional en un número de componentes principales y sus scores y regresión funcional ponderada de mínimos cuadrados parciales; para obtener los pronósticos de series funcionales, estos dos enfoques permiten obtener funciones fluidas que asignan pesos más altos a los datos más recientes. Por último, los intervalos de predicción se obtienen a través de un método de Bootstrap no paramétrico.

Por otro lado, en (Pfisterer et al., 2019) emplean algoritmos de aprendizaje automático que se pueden aplicar a datos funcionales y técnicas clásicas de Machine Learning para

datos escalares; sobre conjuntos de datos con series de tiempo. En este estudio se complementa el trabajo realizado por (Bagnall et al., 2017) y se incorporan técnicas de aprendizaje supervisado en análisis de datos funcionales para resolver problemas de clasificación de series de tiempo. El estudio arrojó resultados con precisiones óptimas en un 95 % de los casos y además mostró, como los métodos funcionales generan mejores resultados de clasificación que técnicas avanzadas de clasificación adaptadas para datos escalares, así mismo, se comprobó que los resultados mejoraron la precisión en un 5.4 % cuando la optimización de los hiperparámetros se realizó empleando modelos jerárquicos bayesianos (Pfisterer et al., 2019).

3.2.1. Clasificación funcional de series de tiempo

La clasificación de series temporales es una técnica en desarrollo que se diferencia del objetivo de hacer una previsión o un pronóstico, partiendo una la búsqueda de patrones y comportamientos que permitan identificar diferencias entre series y así poder determinar la clase de una serie temporal en cuestión. Basicamente lo que se busca con esta técnica es asignar una etiqueta o categoría a una serie temporal en función de sus características o comportamientos observados a lo largo del tiempo.

Un problema común de clasificación de series de tiempo, se trabaja con observaciones secuenciales de una variable a lo largo de períodos de tiempo buscando clasificar a qué categoría pertenece la serie temporal en función de ciertos patrones identificables en los datos.

La clasificación de series de tiempo se aplica en una variedad de contextos. Algunos ejemplos de aplicaciones desarrolladas son:

- Las transacciones financieras a lo largo del tiempo pueden clasificarse para identificar patrones de comportamiento normales y anormales asociados con posibles fraudes.
- Las series temporales de datos de señales cardíacas pueden clasificarse para predecir si un paciente tiene o desarrollará enfermedades cardíacas.
- Usando datos de frecuencia cardiaca sus series temporales pueden permitir clasificar cada una de las curvas para identificar diferentes actividades humanas como caminar, correr, estar sentado, entre otras.

Algoritmos de para clasificar series de tiempo

La clasificación de series de tiempo es una tarea importante en muchas áreas y se ha aplicado en diferentes áreas, como la bolsa de valores y la medicina, entre otras. Aunque existe evidencia que se han probado diferentes técnicas de clasificación análogas al método estadístico clásico, para este estudio se realizará un recorrido sobre las técnicas de interés empleados en este caso.

Algoritmo de clasificación KNN - DTW (Vecino más cercano con deformación dinámica del tiempo)

Un enfoque que se ha desarrollado bastante en clasificación de series temporales por ser uno de los más intuitivos y ofrecer buenos resultados es el algoritmo KNN (k-nearest neighbors), que se basa en la idea de que una serie de tiempo funcional es similar a sus vecinos más cercanos en el espacio de características. Para aplicar el algoritmo KNN a las series de tiempo funcionales, se utiliza una métrica que es la función matemática que define la disimilaridad en los diferentes grupos de muestras Faouzi (2022).

La simplicidad y efectividad del algoritmo KNN lo han hecho ampliamente utilizado en aplicaciones como la clasificación de patrones en señales de ECG y la detección de anomalías en la transmisión de datos.

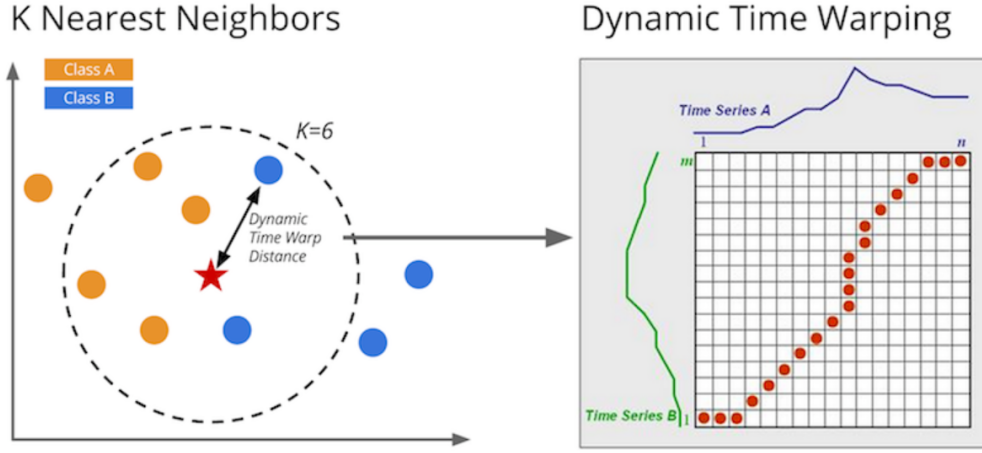
Asimismo, la literatura sugiere la aplicación del algoritmo KNN para clasificación de series de tiempo funcionales utilizando la métrica **KNN con DTW (Dynamic Time Warping)**, debido a que la misma permite corregir principalmente 2 aspectos relevantes en los que la distancia euclídea se queda corta a la hora de aplicarse a funciones.

- Para Faouzi (2022) la distancia euclídea está definida para dos vectores con la misma longitud, en tanto, que las series temporales de un conjunto de datos requieren de otro tipo de tratamiento, dado que suelen tener longitudes diferentes.
- De manera similar Faouzi (2022) menciona que la distancia euclídea compara los valores de ambas series temporales en cada punto temporal de forma independiente, mientras que los valores de las series temporales están correlacionados; considerando el mínimo de las distancias euclídeas entre la serie temporal más pequeña y los subsegmentos en proporción con la longitud de la serie temporal más grande, lo cual pueden no ser óptimo.

Dinamic time warping (DTW)

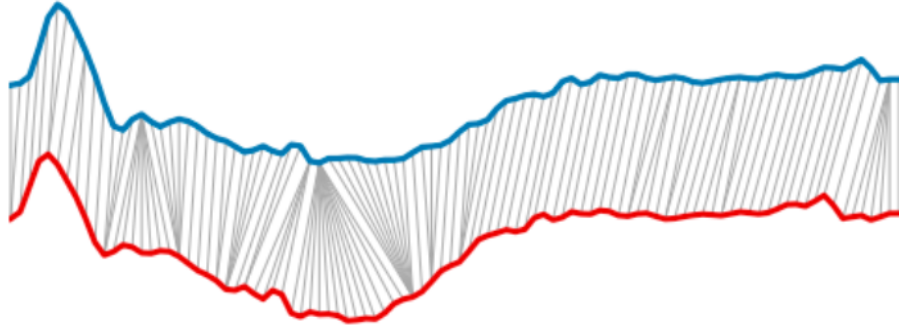
Como se ha mencionado la medida de distancia DTW es recomendada por su adaptabilidad en series de tiempo funcionales y por la mejora en los resultados con KNN. Según (Bagnall et al., 2015) en diferentes estudios se ha comprobado que 1-NN DTW a través de validación cruzada es bastante preciso y difícilmente de superar por otras técnicas de clasificación de series de tiempo figura **3-3**. Sin embargo, es importante mencionar que DTW puede conllevar a un costo computacional muy alto y la clasificación puede tardar mucho tiempo. De acuerdo con (DEVBAY., 2020) para cada nueva serie de tiempo se debe calcular la distancia a todas las series temporales en los datos de entrenamiento ver figura **3-4**.

DTW es en sí mismo un problema de optimización, lo que genera lentitud en el cálculo de la distancia y para un gran conjunto de datos que consta de millones de series temporales de datos, debe repetirse una y otra vez. Por lo tanto, no escala más allá de problemas bastante

Figura 3-3: *KNN - Dynamic time warping*

Nota: Tomado de <https://github.com/markdregan/K-Nearest-Neighbors-with-Dynamic-Time-Warping/> (Mark, 2018)

pequeños.

Figura 3-4: *Dynamic time warping*

Nota: Tomado de <https://developersbay.se/time-series-classification-an-overview/#ucr/DEVBAY.> (2020)

Sean $X = (x_1, \dots, x_n) \in \mathbb{R}^n$ and $Y = (y_1, \dots, y_m) \in \mathbb{R}^m$ dos series de tiempo con longitudes n y m , con un matriz de costo C de tamaño $n \times m$ que consiste en el par de valores de las dos series X y Y , con:

$$\forall i, j \in \{1, \dots, n\} \times \{1, \dots, m\}, C_{ij} = f(x_i, y_j)$$

donde f es una función que evalúa el costo entre un par de números reales, general-

mente esta función está dada por la diferencia entre las distancias euclídeas, así:

$$\forall x, y \in \mathbb{R}, f(x, y) = (x - y)^2$$

La deformación dinámica temporal (DTW) estaría dada por un score DTW y un costo asociado por trayectoria que se define con la expresión:

$$\text{DTW}(X_{:i}, Y_{:j}) = C_{i,j} + \min\{\text{DTW}(X_{:i1}, Y_{:j1}), \text{DTW}(X_{:i1}, Y_{:j}), \text{DTW}(X_{:i}, Y_{:j1})\}$$

Algoritmo KNN Kernel

La clasificación de series de tiempo funcionales también puede realizarse utilizando algoritmos de clasificación basados en Kernel. Este enfoque se basa en la transformación de las series de tiempo a un espacio de características de alta dimensión, donde se aplica una función Kernel para medir la similitud entre ellas. Este enfoque ha demostrado ser efectivo en tareas de clasificación complejas, especialmente cuando se tienen series de tiempo de formas variables o no estructuradas.

Los métodos Kernel se basan en funciones de núcleo que miden similaridad entre cualquier par de entradas (Faouzi, 2022)

Modelo de clasificación series temporales funcionales con regresión de gradiente potenciado

Otro enfoque para la clasificación de series de tiempo funcionales se construye mediante el uso de algoritmos de gradiente funcional. Este enfoque se basa en la optimización de una función objetivo que refleja la similitud entre las series de tiempo y sus respectivas categorías. El gradiente de esta función se utiliza para guiar la clasificación, permitiendo la asignación de cada serie de tiempo a la categoría o clase que maximiza la similitud.

El Modelo de clasificación de series temporales funcionales con regresión de gradiente potenciado es un algoritmo de aprendizaje automático que se utiliza para predecir una variable categórica a partir de una serie temporal funcional. La regresión de gradiente potenciado es una técnica de optimización que permite ajustar los parámetros del modelo de manera eficiente.

La fórmula básica del modelo puede ser representada como:

$$y = f(X, \beta) + \epsilon$$

donde:

y es la variable categórica por predecir.

X es la serie temporal funcional.

β es el vector de parámetros a ajustar.

f es la función de regresión.

ϵ es el error residual.

El proceso de entrenamiento consiste en encontrar los valores óptimos de los parámetros β que minimizan la función de pérdida. Una vez entrenado, el modelo se puede utilizar para hacer predicciones en nuevos datos.

Este enfoque es efectivo en tareas de clasificación de series de tiempo de alta dimensionalidad y formas complejas, especialmente cuando se tienen restricciones de memoria o tiempo de procesamiento limitados.

Para (Bühlmann and Hothorn, 2007) el modelo funcional de gradiente potenciado se estima mediante la combinación iterativa de modelos simples análogo al modelo clásico y puede verse como un método que se comporta a través del descenso de gradiente.

Modelo de clasificación funcional lineal generalizado

El modelo de clasificación funcional lineal generalizado (GLM, por sus siglas en inglés) es un modelo estadístico que se utiliza para modelar relaciones lineales entre variables predictoras y una variable objetivo categórica. La forma general de un GLM se puede representar matemáticamente como:

$$y = f(E(y)) = f(X\beta)$$

donde y es la variable objetivo, X es la matriz de variables predictoras, β es un vector de coeficientes, y f es una función de enlace que se utiliza para modelar la relación entre las variables.

El GLM permite describir relaciones lineales entre las variables y la variable objetivo, lo que lo hace adecuado para una amplia variedad de problemas de clasificación multiclase. Sin embargo, es importante tener en cuenta que el GLM asume una relación lineal entre las variables, por lo que puede ser insuficiente para modelar relaciones complejas.

Modelo de clasificación funcional aditivo generalizado

El modelo de clasificación funcional aditivo generalizado (FGAM, por sus siglas en inglés) es un modelo estadístico que permite realizar la clasificación de datos mediante una combinación no lineal de funciones básicas. Este modelo se utiliza para modelar relaciones complejas entre variables predictoras y una variable objetivo-categorica. Los componentes básicos de un GAM son las funciones aditivas, que permiten describir relaciones no lineales entre las variables y la variable objetivo. Además, el modelo es flexible y permite controlar la complejidad de la relación entre las variables, lo que lo hace adecuado para una amplia variedad de problemas de clasificación.

El Modelo de clasificación funcional multinomial aditivo generalizado (MGAM, por sus siglas en inglés) es una extensión del Modelo de clasificación funcional aditivo generalizado (FGAM) que se utiliza para modelar relaciones complejas entre variables predictoras y una variable objetivo-categorica con más de dos categorías. La forma general de un MGAM se puede representar matemáticamente como:

$$P(y = k|x) = f_k(x) = \exp(g_k(x)) / \sum_{j=1}^K \exp(g_j(x))$$

donde $P(y = k|x)$ es la probabilidad de que la variable objetivo y sea igual a la categoría k dado un vector x de variables predictoras, g_k es una función aditiva que permite describir relaciones complejas entre las variables y la variable objetivo, y K es el número total de categorías en la variable objetivo.

El MGAM permite describir relaciones complejas entre las variables y la variable objetivo, lo que lo hace adecuado para una amplia variedad de problemas de clasificación multiclase. Cada función aditiva en el modelo puede ser una combinación no lineal de las variables predictoras y permite controlar la complejidad de la relación entre las variables

Modelo de clasificación funcional con partición recursiva para regresión

El Modelo de clasificación funcional con partición recursiva para regresión (RPART, por sus siglas en inglés) es un modelo de árboles de decisión que se utiliza para modelar relaciones complejas entre variables predictoras y una variable objetivo-categorica. La forma general de un RPART se puede representar matemáticamente como:

$$y = f(x) = \sum_{m=1}^M w_m I(x \in R_m)$$

donde y es la variable objetivo, x es el vector de variables predictoras, I es una función indicadora que toma el valor 1 si x satisface la condición especificada, R_m es el conjunto de región m en el espacio de las variables predictoras, w_m es el valor predicho para la región m , y M es el número total de regiones en el modelo.

El RPART utiliza una técnica de partición recursiva para dividir el espacio de las variables predictoras en regiones homogéneas en términos de la variable objetivo. Cada región es representada por un valor de la variable objetivo y las regiones se dividen en subregiones hasta que se cumplen ciertas condiciones de parada. Esto permite describir relaciones complejas entre las variables y la variable objetivo, lo que lo hace adecuado para una amplia variedad de problemas de clasificación.

Modelo de clasificación funcional con máquina de soporte de vectores (SVM)

El modelo de clasificación funcional con máquina de soporte de vectores (SVM, por sus siglas en inglés) es un modelo de aprendizaje automático que se utiliza para modelar relaciones complejas entre variables predictoras y una variable objetivo-categorica. La forma general de un SVM se puede representar matemáticamente como:

$$f(x) = \text{sign}(w^T x + b)$$

$$c^*(x) = \sum_{i=1}^N \alpha_i K(x_i, x)$$

donde f es la función de decisión que determina la clase de la variable objetivo, x es el vector de variables predictoras, w es un vector de pesos que representa la dirección de la frontera de separación entre las clases, b es un término de sesgo, y sign es una función que toma el signo de su argumento.

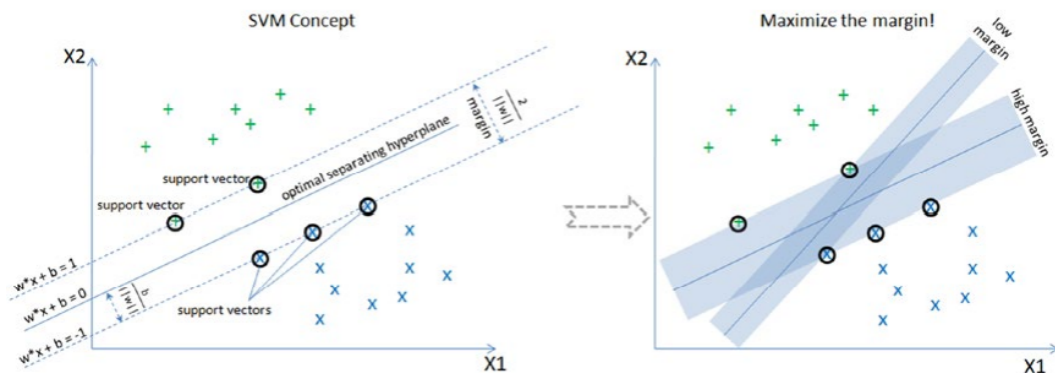
La figura **3-5** permite entender como SVM utiliza una técnica de optimización para encontrar una frontera de separación óptima que maximiza el margen entre las clases. Esto se puede visualizar como un hiperplano que separa las clases en el espacio de las variables predictoras. SVM es particularmente adecuado para problemas de clasificación en los que las variables predictoras tienen altas dimensiones o los datos están altamente correlacionados.

Las SVM fueron concebidos para para solucionar problemas de clasificación multivariante, sin embargo, también se han utilizado con buenos resultados en problemas que con enfoque funcional. En (Rossi and Villa, 2006) los autores explican cómo se utiliza SVM con este enfoque a través de una serie de datos con datos funcionales. Finalmente, concluyen que esta técnica es bastante precisa cuando se aplica a datos de naturaleza funcional.

El modelo de clasificación funcional con máquina de soporte de vectores multinomial es una extensión del Modelo de clasificación funcional con máquina de soporte de vectores (SVM) para manejar problemas de clasificación multiclase. En lugar de encontrar una sola frontera de separación, el SVM multinomial encuentra múltiples fronteras de separación para cada par de clases.

Una forma común de implementar el SVM multinomial es a través del uso de una técnica llamada 'one-vs-all' o 'one-vs-one'. En la técnica 'one-vs-all', se entrena un modelo

Figura 3-5: Máquina de soporte de vectores.



Nota:

Tomado de Mastering machine learning with python in six steps. (Swamynathan, 2019)

SVM para cada clase en el problema de clasificación y se elige la clase con la mayor probabilidad de ser correcta. En la técnica 'one-vs-one', se entrena un modelo SVM para cada par de clases y se elige la clase con la mayor cantidad de victorias en los modelos individuales.

El SVM multinomial se puede representar matemáticamente como:

$$f(x) = \operatorname{argmax}_{i=1}^K w_i^T x + b_i$$

donde:

f es la función de decisión que determina la clase de la variable objetivo.

x es el vector de variables predictoras.

w es un vector de pesos que representa la dirección de la frontera de separación para cada clase.

b es un término de sesgo.

K es el número de clases en el problema de clasificación.

Modelo de clasificación funcional de red neuronal (nnet)

El Modelo de clasificación funcional de red neuronal es un enfoque de aprendizaje automático para la clasificación de series temporales. Este modelo se basa en la teoría de las redes neuronales y consiste en una red de nodos interconectados que se ajustan a través de un proceso de entrenamiento supervisado para producir una predicción de la clase de una serie temporal dada.

La fórmula general para una red neuronal se puede expresar como:

$$y = f(W * X + b)$$

donde:

y es la salida de la red neuronal.

f es la función de activación.

W son los pesos de la red neuronal.

X son las características de la serie temporal.

b es el sesgo de la red neuronal.

Durante el entrenamiento, los pesos W y el sesgo b se ajustan para minimizar el error entre la salida prevista y la salida real de la red neuronal. La función de activación se utiliza para regular la salida de la red neuronal y puede ser una función sigmoideal, tangente hiperbólica o rectificadora.

Modelo de clasificación funcional empleando Naive Bayes

El Modelo de clasificación funcional Naive Bayes multinomial es un enfoque de aprendizaje automático para la clasificación de series temporales que utiliza el teorema de Bayes para calcular la probabilidad de cada clase dada una serie temporal. Este modelo asume que las características de la serie temporal son independientes entre sí, lo que se conoce como la suposición de naive Bayes.

La fórmula para el modelo de clasificación funcional Naive Bayes multinomial se puede expresar como:

$$P(C_i|X) = P(X|C_i) * P(C_i)/P(X)$$

donde:

$P(C_i|X)$ es la probabilidad de que la serie temporal X pertenezca a la clase i .

$P(X|C_i)$ es la probabilidad de que la serie temporal X tenga un conjunto de características dado que pertenece a la clase i .

$P(C_i)$ es la probabilidad a priori de la clase i .

$P(X)$ es la probabilidad a priori de la serie temporal X .

La probabilidad $P(X|C_i)$ se puede estimar utilizando una distribución multinomial, que se ajusta a través de un proceso de entrenamiento supervisado con los datos de entrenamiento. La probabilidad a priori $P(C_i)$ se puede estimar a partir de la frecuencia de cada clase en los datos de entrenamiento. Finalmente, la clase con la probabilidad más alta se selecciona como la clase prevista para la serie temporal X .

3.2.2. Precisión de un modelo de clasificación multiclase

3.2.3. Clasificación funcional de series de tiempo

Para un clasificador multiclase la precisión de una etiqueta k mide la relación que hay entre el número de elementos clasificados de manera correcta. Es una métrica que mide proporción de aciertos de las proyecciones hechas por el modelo en comparación con el total de predicciones realizadas. Su relevancia como métrica de un clasificador está determinada su resultado sea cercano a 1 en un modelo en donde los valores de las etiquetas estén balanceadas o equilibradas.

$$\text{Precisión} = \frac{\sum_{i=1}^k \text{Verdaderos positivos}_i}{\sum_{i=1}^k (\text{Verdaderos positivos}_i + \text{Falsos positivos}_i)}$$

Donde:

- k es cada una de la clases.
- Verdaderos positivos (TP_i) son las instancias correctamente clasificadas como clase i .
- Falsos positivos (FP_i) son las instancias incorrectamente clasificadas como la clase i .

4 Entendimiento de los datos

En esta sección se describe el paso a paso llevado a cabo para la obtención de los datos utilizados en este estudio; involucra la identificación de las fuentes de información, la exploración de los múltiples conjuntos de datos que contienen información asociada al cálculo del IDH, el estado de los datos, así como la calidad de estos mismos. Además del proceso llevado a cabo para la selección de variables que permiten la clasificación de series de tiempo funcionales.

Como se resaltó anteriormente, el IDH fue creado por la ONU con el propósito de mantener el desarrollo humano. Por lo tanto, la exploración bibliográfica inicial, la consulta de reportes, la descarga de conjuntos de datos y demás insumos empleados en este análisis fueron obtenidos en su mayoría, a través del data center que está disponible en la página web del Programa de las Naciones Unidas para el Desarrollo (PNUD) (Programme, 2023).

Las bases de datos contienen información histórica sobre los resultados del IDH, con niveles de información por país. Asimismo, a través de la exploración de dicha url se detectó la existencia de una amplia variedad de reportes, con datos de diferentes variables asociadas con desarrollo humano. No obstante, algunos conjuntos de datos de las variables analizadas dentro de este estudio fueron obtenidas por consulta del repositorio del Banco Mundial (Mundial, 2023), que fue la segunda gran fuente de obtención de información sobre la temática en desarrollo humano.

4.1. Descripción de los datos

Los datos crudos fueron obtenidos en formato Excel, en total se exploró 54 conjuntos de datos con información relevante sobre el IDH. El análisis exploratorio demandó evaluar por separado cada dataset en términos de su estructura de datos, la cantidad de información disponible, la calidad de los datos y la integridad de estos mismos.

En este punto se encontraron diferencias en cuanto a la temporalidad de la información reportada, es decir, mientras algunos datasets cuentan con datos reportados de manera anualizada, en otros casos los datos se encuentran disponibles de manera quinquenal o con disponibilidad en periodos de tiempo diferentes.

Por último, se desarrolla un análisis para datasets con disponibilidad de información anualizada por país desde el año 1990 hasta 2019. La mayor parte de los datasets seleccio-

nados cuentan con registros de información de 193 países, sin ser esta la generalidad; ya que en algunas de las bases de datos seleccionadas; aunque se registra esa cantidad de países al explorar su contenido se encuentran valores faltantes en gran parte de los registros. A continuación, en la tabla **4-1** se relacionan las categorías en las que fueron clasificados los datasets según la temática con la que tienen que ver:

Tabla 4-1: *Conjuntos de datos por categoría*

Categoría	No de datasets
Comercio y flujos financieros	2
Desigualdad	5
Educación	7
Emisiones	4
Género	2
Ingresos	6
Migración y comunicación	4
Población	4
Pobreza	4
Salud	6
Seguridad humana	1
Sostenibilidad	7
Trabajo, empleo y vulnerabilidad	2

Nota: *Elaboración propia*

4.2. Preparación de los datos

Inicia con la exploración de datos que es un proceso fundamental para entender y sacar el máximo provecho a los conjuntos de datos con los que se cuenta. Debido a la variedad de conjuntos de datos seleccionados para el análisis del IDH, fue necesario emplear el manejo de listas anidadas en R. Lo anterior, permitió obtener eficiencias en el momento de generar el código y ejecutarlo, las listas anidadas se presentan cuando dentro de un objeto lista en R, se almacena en otro objeto del mismo tipo (RCODER., 2023), facilitando su manipulación y el análisis cuando se cuenta con una alto volumen de conjuntos de datos.

4.2.1. Limpieza y selección variables a utilizar en el estudio

En este proceso, se aplicó con rigurosidad una serie de técnicas para evaluar los datos disponibles; identificando y tratando valores faltantes, atípicos o inconsistentes que puedan afectar la precisión de los resultados. En particular, se buscó preservar la integridad del conjunto de datos eliminando registros incompletos o duplicados, asegurando que cada país,

contenido en el dataset contara con información coherente y completa. La depuración de datos por integridad no solo mejoró la calidad del conjunto de datos, sino que también ayudó a establecer una base sólida para análisis más profundos y confiables, en relación con el desarrollo humano en diferentes regiones. Al eliminar las fuentes potenciales de error, como inconsistencias en los valores reportados, se crea una base más sólida para la toma de decisiones informadas y la comprensión de los patrones y tendencias que influyen en el índice de desarrollo humano en el tiempo y en la diversidad de geográfica.

Despues de cargar los datasets en el entorno de R, se llevó a cabo un proceso de limpieza que dió inicios con la identificación de campos vacíos para su correcta depuración, posteriormente se procedió con la inspección e identificación de campos con consonantes y/o signos y su respectiva depuración. También, se llevó a cabo procesos de depuración de registros con individuos (países) que no cumplieron con los mínimos requeridos para que ser considerados como aptos para la construcción de los datos funcionales.

En el proceso de selección variables se descartó múltiples conjuntos de datos, como consecuencia de llevar a cabo una depuración de registros de países con una proporción de valores faltantes que superaba el 15 % del total de los datos esperados según la temporalidad del dataset. Esta estrategia permitió clasificar los datasets desde el punto de vista integridad de la información contenida: en datasets para excluir del estudio y datasets útiles para el mismo. Se consideró como regla de exclusión aquellos datasets en los cuales se eliminó el 25 % de los registros contenidos con respecto al archivo original. El resultado de tasas de depuración y la decisión por dataset es como se muestra en la tabla **4-1**.

Figura 4-1: Tasas de depuración de registros por conjunto de datos

Nombre covariable funcional	# Registros base inicial	# Registros después de quitar registros con valores faltantes	Tasa de depuración de valores faltantes	Decisión
A1_Foreign direct investment, net inflows (% of GDP).xlsx	193	149	22,80%	Mantener
A2_Rural population with access to electricity (%) (1).xlsx	195	104	46,67%	Excluir
B1_Coefficient of human inequality.xlsx	189	125	33,86%	Excluir
B3_Inequality in life expectancy (%).xlsx	191	179	6,28%	Mantener
B4_Inequality in income (%).xlsx	190	127	33,16%	Excluir
B5_Inequality in education (%).xlsx	191	133	30,37%	Excluir
B7_Expectancy in life at birth.xlsx	194	194	0,00%	Mantener
C1_Education index.xlsx	189	144	23,81%	Mantener
C2_Expected years of schooling (years).xlsx	193	172	10,88%	Mantener
C3_Expected years of schooling, female (years).xlsx	192	142	26,04%	Excluir
C4_Expected years of schooling, male (years).xlsx	192	141	26,56%	Excluir
C5_Government expenditure on education (% of GDP).xlsx	192	10	94,79%	Excluir
C6_Percentage of primary schools with access to the internet.xlsx	191	4	97,91%	Excluir
C7_Percentage of secondary schools with access to the internet.xlsx	191	3	98,43%	Excluir
D10_Área selvática (kilómetros cuadrados).xlsx	192	188	2,08%	Mantener
D7_Natural resource depletion (% of GNI).xlsx	190	105	44,74%	Excluir
D8_Use of fertilizer nutrient nitrogen (N), per area of cropland (kg per hectare)	190	140	26,32%	Excluir
D9_CO2 emissions (kt).xlsx	192	188	2,08%	Mantener
E1_Total unemployment rate (female to male ratio).xlsx	191	179	6,28%	Mantener
E2_Youth unemployment rate (female to male ratio).xlsx	191	180	5,76%	Mantener
F1_Gross national income per capita, female (2017 PPP \$).xlsx	189	166	12,17%	Mantener
F2_Gross national income per capita, male (2017 PPP \$).xlsx	188	167	11,17%	Mantener
F3_GDP per capita, PPP (current international \$).xlsx	192	156	18,75%	Mantener
F4_Gross national income (GNI) per capita (constant 2017 PPP \$).xlsx	191	189	1,05%	Mantener
F5_Income index.xlsx	191	189	1,05%	Mantener
F7_Oil rents (% of GDP).xlsx	192	188	2,08%	Mantener
G1_International inbound tourists (thousands).xlsx	191	140	26,70%	Excluir
G2_Internet users, total (% of population).xlsx	195	153	21,54%	Mantener
G3_Mobile phone subscriptions (per 100 people).xlsx	195	177	9,23%	Mantener
G5_Mobile celular subscriptions (per 100 people).xlsx	195	179	8,21%	Mantener
I1_Human Development Index (HDI).xlsx	189	144	23,81%	Mantener
I10_Young age (0-14) dependency ratio (per 100 people ages 15-64).xlsx	190	183	3,68%	Mantener
I2_Population ages 65 and older (millions).xlsx	190	183	3,68%	Mantener
I3_Total population (millions).xlsx	194	193	0,52%	Mantener
I4_Urban population (%).xlsx	194	192	1,03%	Mantener
J1_Multidimensional poverty index (MPI).xlsx	189	0	100,00%	Excluir
J3_Population vulnerable to multidimensional poverty (%) (1).xlsx	189	0	100,00%	Excluir
J4_Working poor at PPP\$3.20 a day (% of total employment).xlsx	191	134	29,84%	Excluir
K3_Mortality rate, infant (per 1,000 live births).xlsx	195	191	2,05%	Mantener
K5_Mortality rate, female adult (per 1,000 people).xlsx	191	182	4,71%	Mantener
K6_Mortality rate, male adult (per 1,000 people).xlsx	191	182	4,71%	Mantener
K7_Tuberculosis incidence (per 100,000 people).xlsx	195	187	4,10%	Mantener
K8_HIV prevalence, adult (% ages 15-49).xlsx	190	139	26,84%	Excluir
K9_Inmunización contra sarampión.xlsx	192	188	2,08%	Mantener
L2_Homicide rate (per 100,000 people).xlsx	192	74	61,46%	Excluir
M1_Adjusted net savings (% of GNI).xlsx	189	114	39,68%	Excluir
M3_Research and development expenditure (% of GDP).xlsx	190	63	66,84%	Excluir
M4_Rural population with access to electricity (%).xlsx	195	104	46,67%	Excluir
M5_Skilled labour force (% of labour force).xlsx	192	42	78,13%	Excluir
M6_Total debt service (% of exports of goods, services and primary income)	189	101	46,56%	Excluir
M7_Access to electricity, urban (% of urban population).xlsx	189	181	4,23%	Mantener
M8_Access to clean fuels and technologies for cooking, urban (% of urban)	189	179	5,29%	Mantener
N3_Unemployment, total (% of labour force).xlsx	190	179	5,79%	Mantener
N4_Vulnerable employment (% of total employment).xlsx	191	180	5,76%	Mantener

Nota: Elaboración propia

Para concluir, este proceso permitió decantar 32 datasets como útiles que son los que contaron con una mayor completitud de información de individuos de dentro del periodo de tiempo establecido para el análisis (1990 - 2019).

4.3. Preparación de los datos y construcción de los datos funcionales

Posteriormente, se compiló los 32 datasets seleccionados hasta este punto, cuyas características eran suficientes para trabajar con datos funcionales. Como se ha mencionado los datos fueron tomados del repositorio de información del PNUD. En la tabla 4-2 se proporciona una breve descripción las variables cuya información está contenida en cada uno de los datasets seleccionados.

Tabla 4-2: Descripción de las covariables seleccionadas asociadas al IDH

Variable	Dataset	Descripción
Índice de desarrollo humano	I1 Human Development Index (HDI).xlsx	Medida resumida del logro promedio en dimensiones clave del desarrollo humano: una vida larga y saludable, estar bien informado y tener un nivel de vida decente.
Índice de educación	C1 Education index.xlsx	Medida de la tasa de alfabetización de adultos (con una ponderación de dos tercios) y la tasa bruta combinada de matriculación en primaria, secundaria y terciaria (con una ponderación de un tercio).
Expectativa de vida al nacer	B7 Expectancy in life at birth.xlsx	Número de años que viviría un recién nacido si los patrones de mortalidad prevalecientes en el momento de su nacimiento se mantuvieran iguales a lo largo de su vida.
Expectativa de años de escolaridad	C2 Expected years of schooling (years).xlsx	Número promedio de años completos de educación de la población de 25 años o más de un país de, excluyendo los años dedicados a repetir grados individuales.
Índice de ingreso	F5 Income index.xlsx	Medida de los ingresos totales recibidos
Producto interno bruto per cápita	F3 GDP per capita, PPP (current international\$).xlsx	Este indicador proporciona valores per cápita para el producto interno bruto (PIB) expresados en dólares internacionales corrientes convertidos por el factor de conversión de paridad del poder adquisitivo (PPA).
Ingreso nacional bruto per cápita	F4 Gross national income (GNI) per capita (constant 2017 PPP\$).xlsx	Es el ingreso nacional bruto (GNI) convertido a dólares internacionales utilizando tasas de paridad del poder adquisitivo.
Acceso a electricidad - zona urbana	M7 Access to electricity, urban (% of urban population).xlsx	Es el porcentaje de la población urbana con acceso a electricidad.

Tabla 4-2: *Descripción de las covariables seleccionadas asociadas al IDH*

Variable	Dataset	Descripción
Acceso a tecnologías y combustibles limpios para cocinar	M8 Access to clean fuels and technologies for cooking, urban (% of urban population).xlsx	Proporción de la población urbana que utiliza principalmente tecnologías y combustibles limpios para cocinar.
Área selvática	D10 Forest area (% of land area).xlsx	El área forestal bajo rodales naturales o árboles plantados de al menos 5 metros in situ, sean productivos o no, y excluye los rodales de árboles en sistemas de producción agrícola (por ejemplo, en plantaciones de frutas y sistemas agroforestales) y árboles en parques y jardines urbanos.
Desempleo total (% Fuerza de trabajo)	N3 Unemployment, total (% of labour force).xlsx	Proporción de la fuerza laboral que no tiene trabajo pero que está disponible y buscando empleo.
Desigualdad en esperanza de vida	B3 Inequality in life expectancy (%).xlsx	Es la desigualdad en la distribución de la esperanza de vida según los datos de las tablas de supervivencia estimadas utilizando el índice de desigualdad de Atkinson.
Emisiones de CO2	D9 CO2 emissions (kt).xlsx	Emisiones de dióxido de carbono son las derivadas de la quema de combustibles fósiles y la fabricación de cemento. Incluyen el dióxido de carbono producido durante el consumo de combustibles sólidos, líquidos y gaseosos y la quema de gas.
Empleo vulnerable	N4 Vulnerable employment (% of total employment).xlsx	Trabajadores familiares auxiliares y los trabajadores por cuenta propia como porcentaje del empleo total.
Incidencia de tuberculosis	K7 Tuberculosis incidence (per 100,000 people).xlsx	Número estimado de casos de tuberculosis nuevos y recidivantes que surgen en un año determinado, expresado como tasa por 100.000 habitantes. Se incluyen todas las formas de TB, incluidos los casos en personas que viven con el VIH.
Ingreso nacional bruto per cápita, hombres	F2 Gross national income per capita, male (2017 PPP \$).xlsx	Indicador que proporciona valores per cápita para el ingreso nacional bruto obtenido por hombres.
Ingreso nacional bruto per cápita, mujeres	F1 Gross national income per capita, female (2017 PPP \$).xlsx	Indicador que proporciona valores per cápita para el ingreso nacional bruto obtenido por mujeres.

Tabla 4-2: *Descripción de las covariables seleccionadas asociadas al IDH*

Variable	Dataset	Descripción
Inmunización contra sarampion	k9 Inmunización contra sarampion.xlsx	La inmunización de niños mide el porcentaje de niños de entre 12 y 23 meses de edad que recibieron vacunas antes de los 12 meses o en cualquier momento antes de la encuesta. Se considera que un niño está inmunizado de manera adecuada contra el sarampion tras recibir una dosis de la vacuna.
Inversión extranjera directa	A1 Foreign direct investment, net inflows (% of GDP).xlsx	La inversión extranjera directa son las entradas netas de inversión para adquirir una participación gerencial duradera (10 por ciento o más de las acciones con derecho a voto) en una empresa que opera en una economía distinta a la del inversionista.
Población mayor a 65 años	I2 Population ages 65 and older (millions).xlsx	Población de 65 años y más como porcentaje de la población total. La población se basa en la definición de población de facto, que cuenta a todos los residentes independientemente de su estatus legal o ciudadanía.
Población total	I3 Total population (millions).xlsx	La población total se basa en la definición de población de facto, que cuenta a todos los residentes independientemente de su estatus legal o ciudadanía.
Población urbana	I4 Urban population (%).xlsx	La población urbana se refiere a las personas que viven en áreas urbanas según lo definido por las oficinas nacionales de estadística. Se calcula utilizando las estimaciones de población del Banco Mundial y las proporciones urbanas de las Perspectivas de urbanización mundial de las Naciones Unidas.
Ratio de dependencia de jóvenes	I10 Young age (0-14) dependency ratio (per 100 people ages 15-64).xlsx	La tasa de dependencia por edad es la relación entre los dependientes (personas menores de 15 años o mayores de 64) y la población en edad de trabajar (entre 15 y 64 años).
Rentas petroleras	F7 Oil rents (% of GDP).xlsx	Diferencia entre el valor de la producción de crudo a precios regionales y los costos totales de producción.
Suscripción a telefonía móvil	G3 Mobile phone subscriptions (per 100 people).xlsx	La suscripciones a servicios de telefonía en un país por cada 100 personas.

Tabla 4-2: Descripción de las covariables seleccionadas asociadas al IDH

Variable	Dataset	Descripción
Suscripciones a celular móvil	G5 Mobile cellular subscriptions (per 100 people).xlsx	Los abonos de telefonía móvil celular son abonos a un servicio público de telefonía móvil que proporciona acceso a la RTPC mediante tecnología celular. El indicador incluye (y se divide en) el número de suscripciones de pospago y el número de cuentas de prepago activas (es decir, que se han utilizado durante los últimos tres meses). El indicador se aplica a todas las suscripciones celulares móviles que ofrecen comunicaciones de voz. Se excluyen las suscripciones a través de tarjetas de datos o módems USB, las suscripciones a servicios públicos de datos móviles, los servicios privados de radio móvil troncalizado, telepunto, radiobúsqueda y telemetría.
Tasa de desempleo jóvenes	E2 Youth unemployment rate (female to male ratio).xlsx	La relación entre el desempleo juvenil femenino y el masculino es el porcentaje de las tasas de desempleo juvenil femenino y masculino.
Tasa de desempleo total	E1 Total unemployment rate (female to male ratio).xlsx	Porcentaje de las tasas de desempleo total femenino y masculino.
Tasa de mortalidad infantil	K3 Mortality rate, infant (per 1,000 live births).xlsx	Número de niños que mueren antes de cumplir un año de edad, por cada 1.000 nacidos vivos en un año determinado.
Tasa de mortalidad, hombres adultos	K6 Mortality rate, male adult (per 1,000 people).xlsx	La tasa de mortalidad de adultos, hombres, es la probabilidad de morir entre los 15 y los 60 años, es decir, la probabilidad de que un hombre de 15 años muera antes de cumplir los 60 años, si está sujeto a las tasas de mortalidad específicas por edad de los valores especificados. año entre esas edades.
Tasa de mortalidad, mujeres adultas	K5 Mortality rate, female adult (per 1,000 people).xlsx	La probabilidad de que una mujer de 15 años muera antes de cumplir los 60 años, si está sujeto a las tasas de mortalidad específicas por edad de los valores especificados. año entre esas edades.
Total de usuarios internet	G2 Internet users, total (% of population).xlsx	personas que han utilizado Internet (desde cualquier lugar) en los últimos 3 meses. Internet se puede utilizar a través de una computadora, teléfono móvil, asistente digital personal, máquina de juegos, TV digital, etc.

Nota: *Elaboración propia*

La preparación de datos también implica la estandarización de las siglas de los países, asegurando que las mismas referencias geográficas se utilicen en todos los conjuntos de datos y el estudio. En esta etapa se realiza un esfuerzo por unificar el nombre de cada país bajo el código ISO 3166-1 alpha3 establecido a nivel internacional. Al llevar a cabo estas tareas de integración y estandarización, se crea un conjunto de datos unificado y armonizado, listo para el análisis. Este proceso no solo facilita la comparación y el contraste de variables, sino que también sienta las bases para un análisis robusto y preciso de las diferencias y similitudes en el desarrollo humano entre países.

Una vez superadas las etapas de recolección de información, selección de variables y unificación de los datos; lo importante ahora es definir una muestra de los países (individuos) para los cuales se cuenta con información en cada uno de los datasets seleccionados. En este caso la selección de la muestra se realiza mediante un proceso de selección de muestra no probabilístico según (Hernández-Sampieri and Torres, 2018). Este proceso de cruce de variables tiene como objetivo construir un grupo de datasets cohesivo que permita explorar y comprender las relaciones que explican el IDH. Como resultado de este proceso se identificó la existencia de información completa de 119 países en los 32 datasets compilados.

En el marco del análisis descriptivo de datos funcionales, comunmente se adaptan métodos estadísticos tradicionales para obtener información valiosa sobre las relaciones entre variables funcionales. Un enfoque común es calcular las correlaciones entre pares de variables funcionales utilizando el coeficiente de correlación de Pearson adaptado para datos funcionales. Este coeficiente permite medir el grado de relación lineal entre dos variables a lo largo de sus dominios. Después de calcular las correlaciones, es una práctica común eliminar aquellas variables que presentan una correlación muy baja con respecto a la variable de interés. Esta estrategia de selección se basa en la idea que las variables altamente correlacionadas pueden proporcionar una comprensión más profunda y significativa de la relación entre el comportamiento de las funciones.

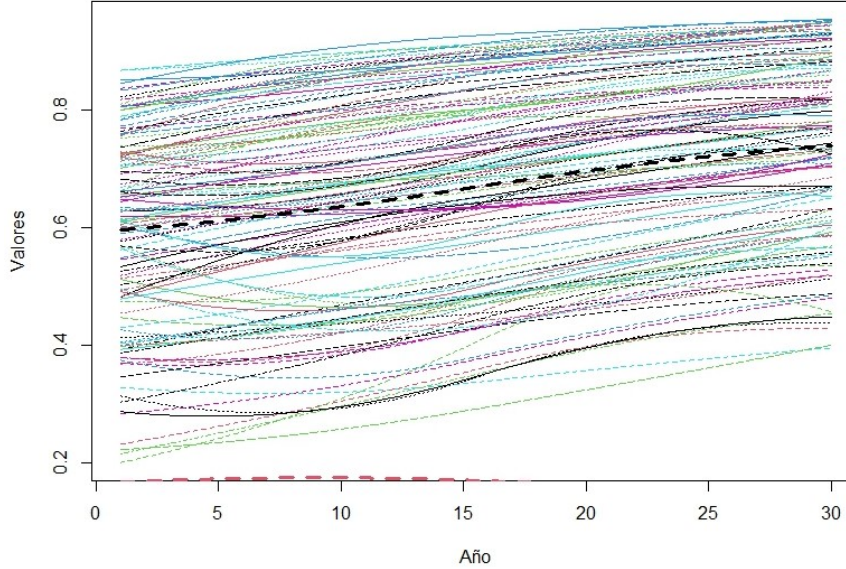
El análisis de datos funcionales aplicado a una variable en particular permite obtener una visión profunda de cómo esa variable evoluciona a lo largo del tiempo. Este enfoque es particularmente útil cuando tratamos con datos que representan fenómenos cambiantes y continuos. Para ilustrar esto ahora, se considera las variables principales del IDH y su comportamiento en 30 años.

Al representar estas mediciones como curvas funcionales, donde el eje x representa el periodo de tiempo, siendo los puntos de discretización los 30 años (1990 - 2019), y el eje y los valores específicos de cada una de las variables. De esta manera se puede visualizar la curva funcional que nos muestra como ha sido el comportamiento variable por país.

En La figura 4-2 se realiza la representación funcional del índice de desarrollo humano, y permite apreciar el comportamiento de este índice según el país. De la misma manera, las figuras 4-3 a 4-6, permiten visualizar el comportamiento funcional de una muestra compuesta

por datos de 119 países para cada uno de los índices principales asociados al IDH.

Figura 4-2: *Datos funcionales - Índice de desarrollo humano (IDH)*



Nota: Tomado de Programa de las Naciones Unidas para el desarrollo. (Programme, 2023)

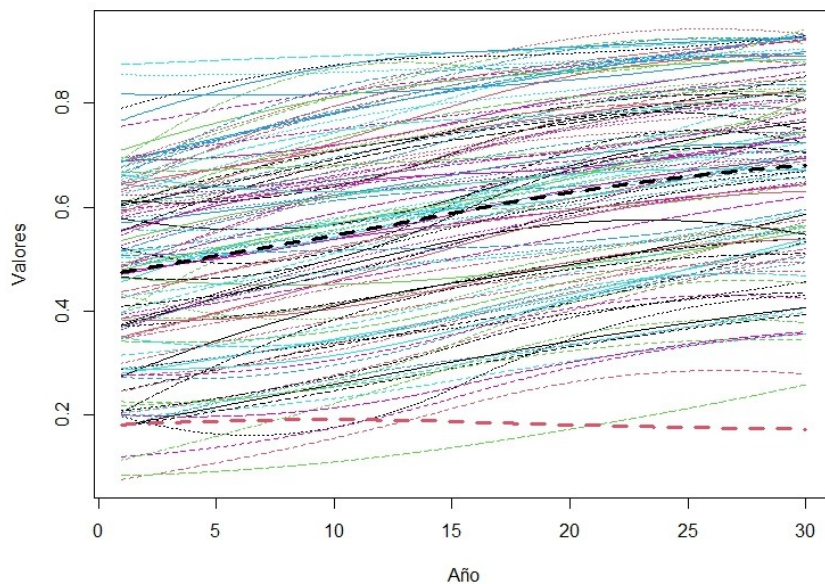
La representación funcional de los datos requiere de la realización previa de un suavizado, que permita detectar los patrones en los datos, para esto existe diferentes métodos que han sido desarrollados por los autores de del FDA. Los diferentes métodos pueden hallarse en (Ramsay and Silverman, 2005).

En la teoría de datos funcionales, las técnicas más empleadas para hacer suavizamiento son las bases de Fourier y B-splines. La primera es bastante conocida y se emplea en casos donde las funciones son periódicas. Sin embargo, en este caso se utilizó la técnica de B-splines cuya base aplica para funciones no periódicas, además que ha sido bastante acogida dadas sus ventajas computacionales al tomar una base de polinomios para el suavizamiento, para este ejercicio fue necesario fijar el parámetro d en 5 básicos con el cual se logró un buen suavizamiento de las curvas en todos los casos.

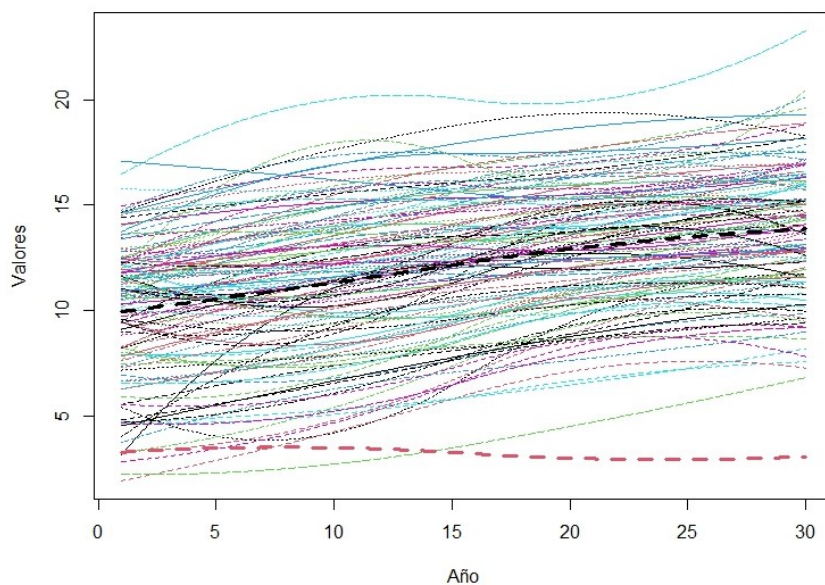
4.4. Análisis exploratorio de datos funcionales

En esta sección se realiza un análisis exploratorio de los datos aplicando cada una de las medidas de posición más utilizadas en datos funcionales. Entonces, se tiene que:

$$S_n = \{X_{ij}\}_{i=1}^n \quad (4-1)$$

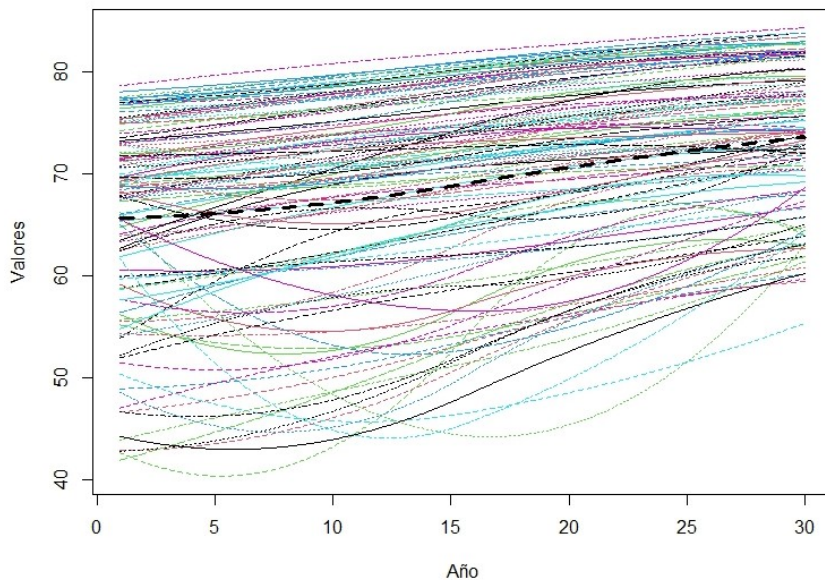
Figura 4-3: *Índice de educación*

Nota: Tomado de Programa de las Naciones Unidas para el desarrollo. (Programme, 2023)

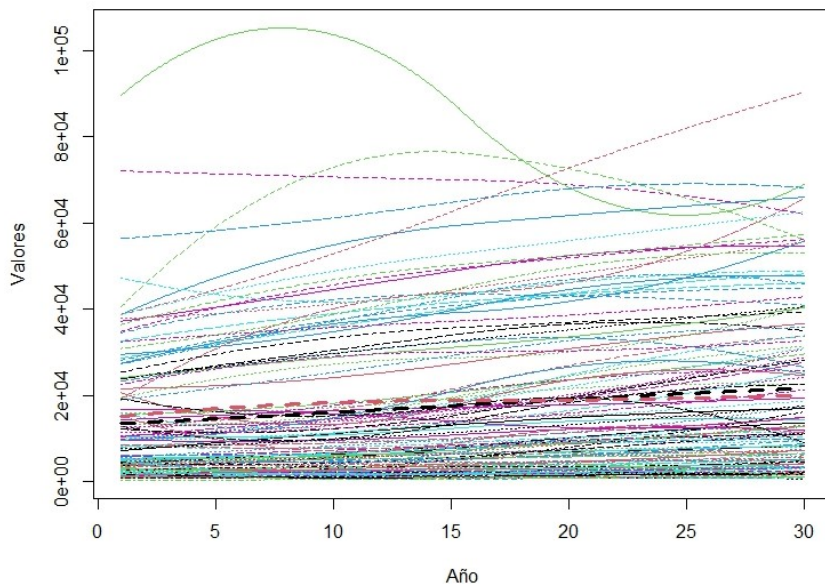
Figura 4-4: *Expectativa de años de escolaridad*

Nota: Tomado de Programa de las Naciones Unidas para el desarrollo. (Programme, 2023)

donde S es la muestra con un tamaño $n = 119$ países y $X_{i,j}$ es la variable funcional que corresponde al país i y el índice j a la variable del IDH que se quiere analizar.

Figura 4-5: *Esperanza de vida al nacer*

Nota: Tomado de Programa de las Naciones Unidas para el desarrollo. (Programme, 2023)

Figura 4-6: *Ingreso nacional bruto per cápita*

Nota: Tomado de Programa de las Naciones Unidas para el desarrollo. (Programme, 2023)

Por otra parte, la media funcional se define como el centro de gravedad de los datos. Tomando la fórmula para el cálculo de la media y la varianza funcionales; mencionadas en

el capítulo anterior. Se tiene que:

$$\bar{x}(t) = N^{-1} \sum_{i=1}^N x_i(t) \quad (4-2)$$

$$Var_x(t) = (N - 1)^{-1} \sum_{i=1}^N [x_i(t) - \bar{x}(t)]^2 \quad (4-3)$$

En las Figuras 4-2 y 4-3 hasta 4-6 muestran la media y la varianza muestral. Se observa que estamos ante un conjunto de curvas que tienen un comportamiento que es estable dado que la curva de la media mantiene su trayectoria en el tiempo y varianza es casi 0.

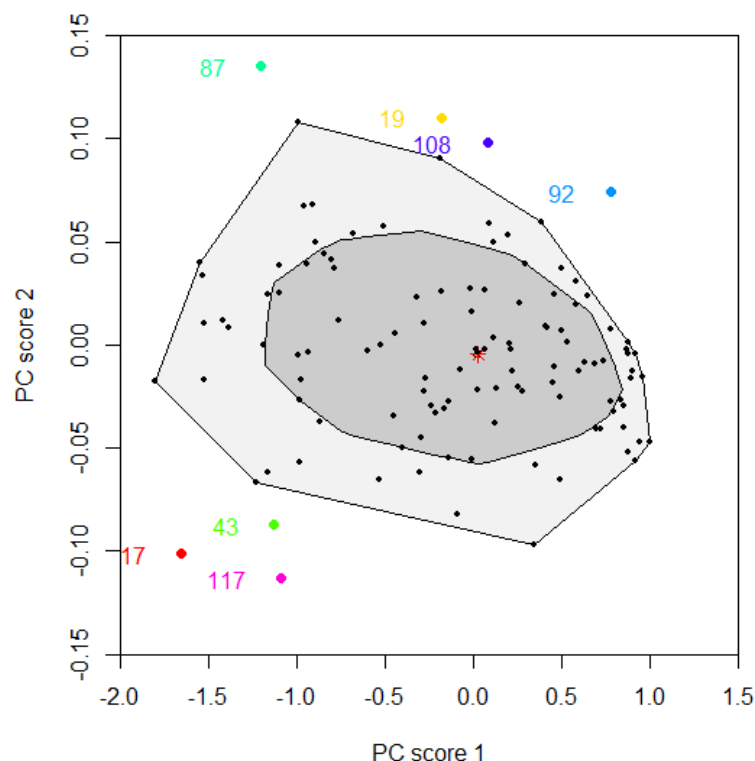
Por otra parte, la figura 4-7, muestra el bagplot funcional es una herramienta gráfica utilizada para visualizar y analizar la distribución de datos funcionales. Proporciona información sobre la dispersión, la forma y la presencia de posibles valores atípicos en un conjunto de datos bidimensional. El bagplot consta de una elipse central que indica la ubicación central de las curvas funcionales (asterisco rojo en la figura), que es similar al concepto de la mediana en un boxplot. Además, el bagplot posee un bolsa exterior (Outer Bag) y una bolsa interior (Inner bag), la bolsa exterior representa una región que puede representar una mayor dispersión de los datos funcionales cuando estos se encuentran en esta zona, por otra parte la bolsa interior representa una zona más compacta indicando menor dispersión cuando la mayoría de los datos funcionales se encuentran allí.

Los puntos atípicos (outliers), son todos los puntos fuera de la bolsa exterior. Estos puntos pueden ser valores extremos o datos que difieren significativamente de la tendencia general. En línea con lo anterior, se aprecia un total de 7 curvas funcionales atípicas pertenecientes a los países relacionados en la tabla 4-3

Tabla 4-3: *Países con curvas atípicas*

Consecutivo	Abreviatura	Nombre País
17	CAF	África Central
19	CHA	China
43	HTI	Haití
87	RWA	Ruanda
92	SGP	Singapur
108	TUR	Turquía
117	YEM	Yemen

Nota: *Elaboración propia*

Figura 4-7: *Bagplot funcional*

Nota: Elaboración propia.

4.5. Análisis funcional de componentes principales

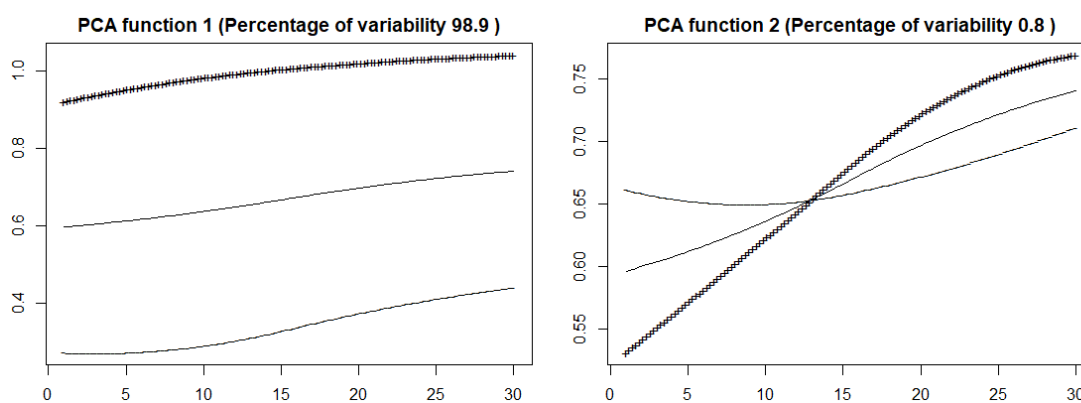
En esta sección se aplica el análisis funcional de componentes principales (FPCA) por sus siglas en inglés, a la muestra de países relacionados con el Índice de Desarrollo Humano (IDH) y curvas con los resultados del IDH en 30 años. En este escenario, el FPCA se convierte en una herramienta esencial para examinar la complejidad de los datos funcionales que representan las evoluciones cambiantes de los indicadores del IDH en cada nación, a lo largo del tiempo. Al aplicar el FPCA, es posible extraer los componentes principales que capturan las variabilidades más significativas entre los países en términos de su desarrollo. Estos componentes no solo permiten identificar los patrones de cambio en los indicadores clave, como la esperanza de vida, la educación y el ingreso per cápita, sino que también proporcionan una visión más clara de las relaciones y diferencias entre los países en función de su desarrollo.

Al reducir la dimensión de los datos funcionales a través de los componentes principales, se logra una representación simplificada que resalta las tendencias más relevantes, facilitando así la interpretación y el análisis comparativo entre naciones.

En medida que los modelos estadísticos se adaptan a la naturaleza funcional de las observaciones, se logra una comprensión de la variabilidad presente en el grupo de datasets seleccionado. Este enfoque no solo amplía nuestro conocimiento sobre las relaciones entre funciones, sino que también muestra el camino para una representación más clara y una toma de decisiones informadas basadas en los resultados obtenidos.

Mediante el uso de la función **pca.fd** de la librería **fda** en R, se identifican los tipos de variación de la muestra de funciones para el IDH. La función se configura con 2 armónicos teniendo en cuenta que el armónico 1 (componente principal) recoge el 98 % de la variación. Los valores propios que indican la importancia de cada modo de variación y las puntuaciones armónicas para funciones individuales figura 4-8.

Figura 4-8: *Análisis funcional de Componentes Principales*



Nota: *Elaboración propia.*

4.6. Análisis clúster

Posteriormente al FPCA, se llevó a cabo la aplicación de técnicas de agrupamiento con el fin de determinar los grupos de curvas funcionales óptimo. Esto permitió definir la muestra de países en tres grupos a saber.

1. Clúster 1: Países desarrollados.
2. Clúster 2: Países desarrollo intermedio.
3. Clúster 3: Países con bajo desarrollo.

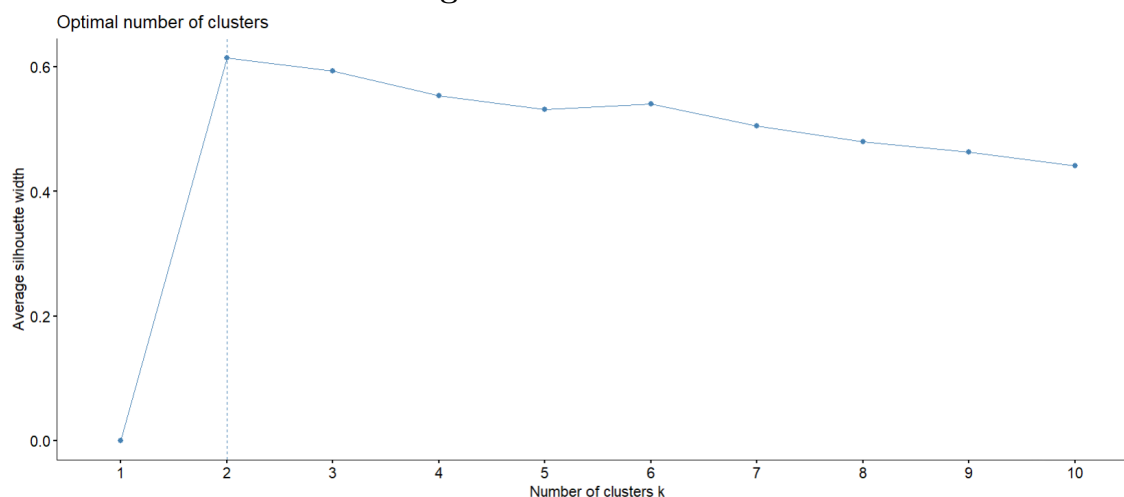
Estos grupos reflejan diferentes niveles de desarrollo humano y permiten una categorización significativa basada en la esperanza de vida, la educación y el ingreso per cápita, además de las variables que afectan cada una de estas variables macro.

4.6.1. Técnicas empleadas para la obtención del número de clústeres

Para la ejecución del análisis de clúster, fue determinante la función **fviz_nbcluster** del paquete **factoextra** en R, para llevar a cabo la aplicación de las técnicas como la silueta, el método del codo y gap stat en la definición de los grupos.

Mediante el uso de la técnica de la silueta, se evaluó la calidad de la agrupación en función de la distancia entre los puntos dentro de un mismo clúster y la distancia promedio a los clústeres vecinos arrojando un resultado de 2 grupos, lo que se observa en la figura 4-9

Figura 4-9: Método de Silueta

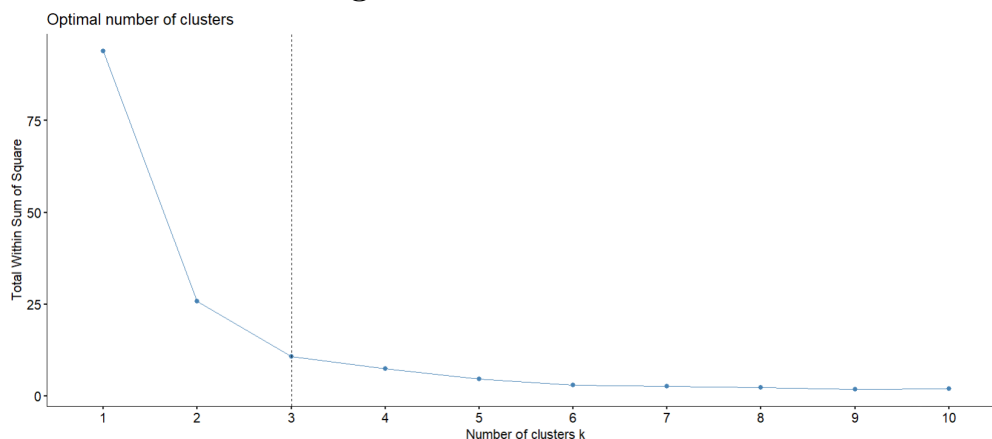


Nota:

Fuente: Elaboración propia.

En paralelo, se ejecutó el método del codo para determinar el número óptimo de clústeres basado en la variabilidad de los datos, obteniendo un resultado de 3 grupos 4-10.

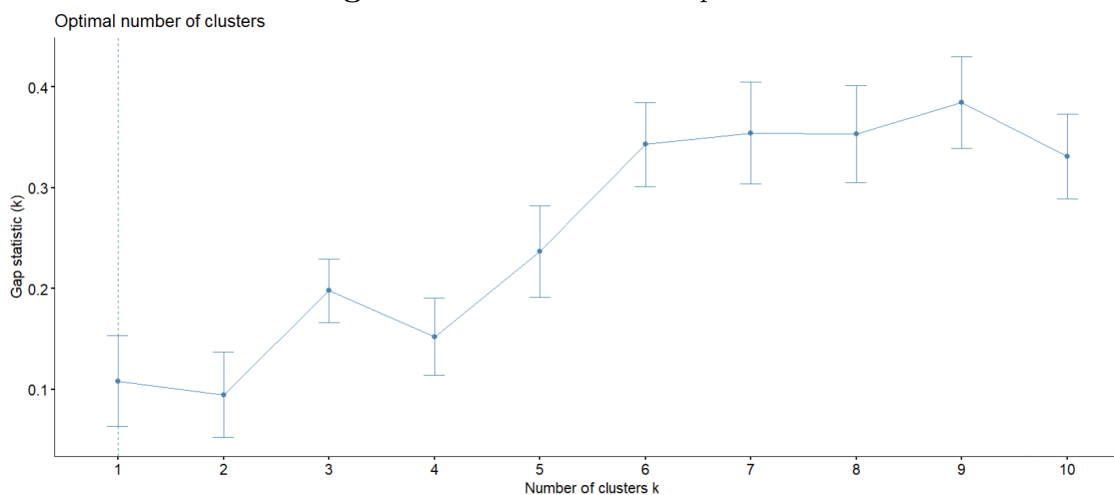
Figura 4-10: Método Elbow



Nota: *Fuente: Elaboración propia.*

Como las técnicas descritas anteriormente, arrojaron resultados diferentes, para obtener un criterio de decisión objetivo sobre el número óptimo de grupos a seleccionar, se llevó a cabo la aplicación de un tercer método de clusterización denominado **gap_stat**; que se basa en la búsqueda de la variación intraclúster para diferentes valores en el número de grupos. El resultado de esta técnica ahora fué 1 grupo **4-11**.

Figura 4-11: *Método de Gap Statistic*



Nota: *Fuente: Elaboración propia.*

Por consiguiente, se empleó el paquete NbClust de R que provee 30 índices diferentes, y proporciona el mejor resultado de agrupación clúster al evaluar diversos tipos de técnicas y medidas de distancia. La figura **4-12** permite identificar los resultados obtenidos con ayuda de este paquete.

Figura 4-12: *No. Óptimo de clústeres*

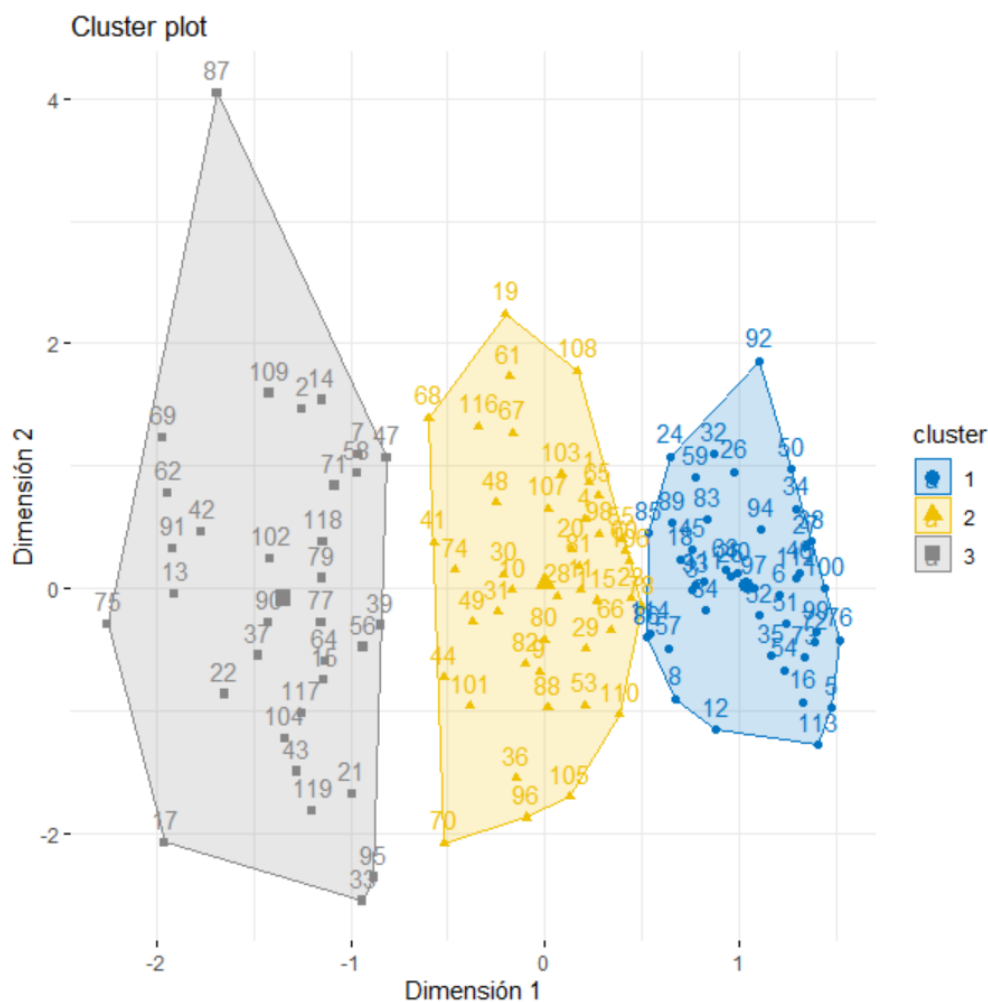
```
*****
* Among all indices:
* 7 proposed 2 as the best number of clusters
* 10 proposed 3 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 3 proposed 8 as the best number of clusters
* 1 proposed 10 as the best number of clusters
* 3 proposed 11 as the best number of clusters
* 2 proposed 15 as the best number of clusters
***** Conclusion *****
* According to the majority rule, the best number of clusters is 3
*****
```

Nota: *Elaboración propia con paquete Nbclust.*

Obtención de los clústeres

Con el método descrito en la sección anterior se llega a la conclusión de que la división en 3 grupos es la que permite obtener el mejor resultado en términos de captura de variabilidad; los resultados se soportan con el método Elbow y los resultados obtenidos con el paquete **NbClust**. Figura 4-13

Figura 4-13: *Clústers Índice de desarrollo humano*



Nota:

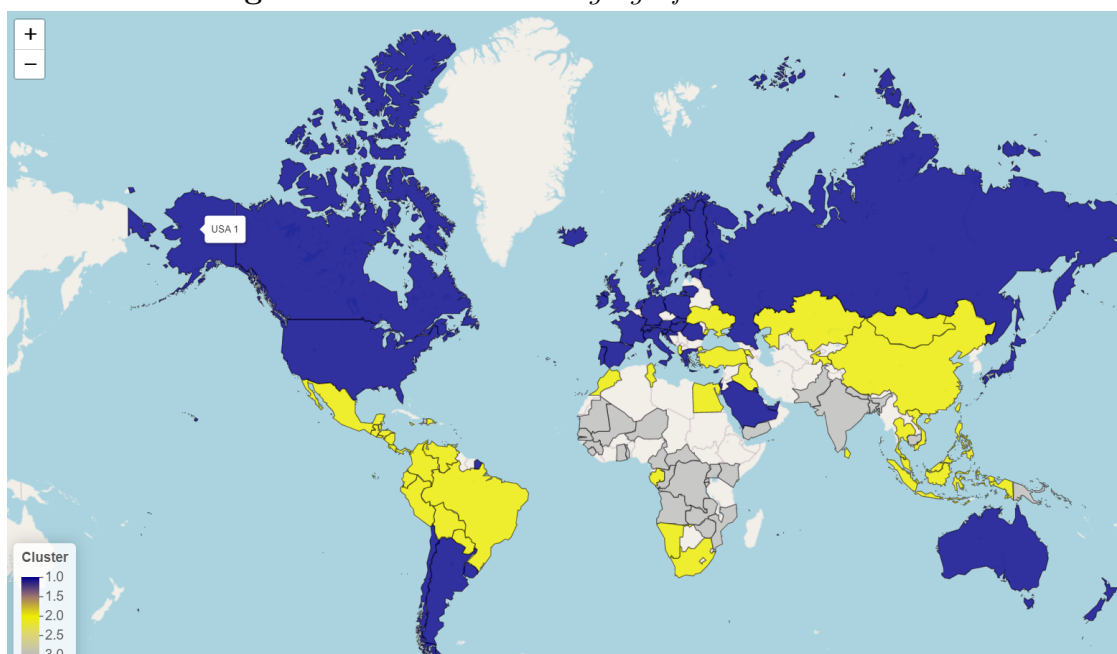
Elaboración propia

Análisis de los clústers

En esta parte se llevó a cabo la representación geográfica de los países, según el grupo obtenido a través del análisis clúster; además del análisis de las variables seleccionadas anteriormente. Lo anterior, con el objeto de identificar las variables más influyentes dentro cada agrupación.

Los países fueron distribuidos en 3 grupos clúster, con patrones significativos en la distribución global. En el clúster 1 se agrupan países de América del Norte, Europa, Australia y Rusia, llamando la atención que en este selecto grupo se encuentra a Chile y Argentina. En el clúster 2 incluyen países de América Latina, Asia y la parte sur de África, entre otros; y por último, en el clúster 3 se tiene las naciones del continente africano, con India y Pakistán en su mayoría. Ver figura 4-14.

Figura 4-14: *Distribución geográfica de los clústers*

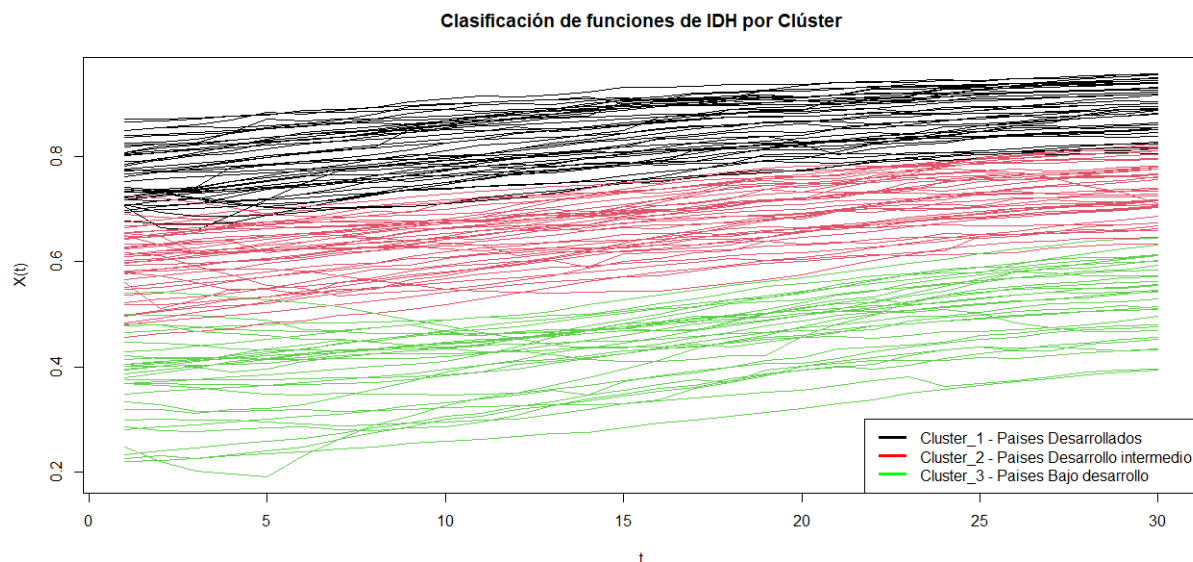


Nota: *Elaboración propia*

En este punto la perspectiva visual de los grupos mencionados, empiezan a mostrar similitudes y diferencias entre las regiones; resaltando principalmente la ubicación geográfica de países por grupo, y por consiguiente tener una idea general de la posible situación económica subyacente de cada país.

La figura 4-15 muestra como para el cluster 1 de países desarrollados las curvas funcionales tienen posiciones muy altas y que a medida que la función se encuentra más baja la probabilidad de pertenecer al grupo 2 aumenta; ocurriendo lo mismo con el grupo 2 y el grupo 3.

Adicionalmente, se profundizó en el análisis de las diferencias entre las 3 agrupaciones clúster de países obtenidas, llevando a cabo una descriptiva de los grupos a través de diagramas de caja. Estos proporcionan una visualización efectiva de las distribuciones y dispersiones de las variables clave relacionadas con el Índice de Desarrollo Humano (IDH); permitiendo tener una comprensión más precisa de cómo los países en cada clúster se desempeñan en términos de desarrollo y cómo algunas variables están relacionadas con factores

Figura 4-15: *Clasificación Índice de desarrollo humano (IDH)*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

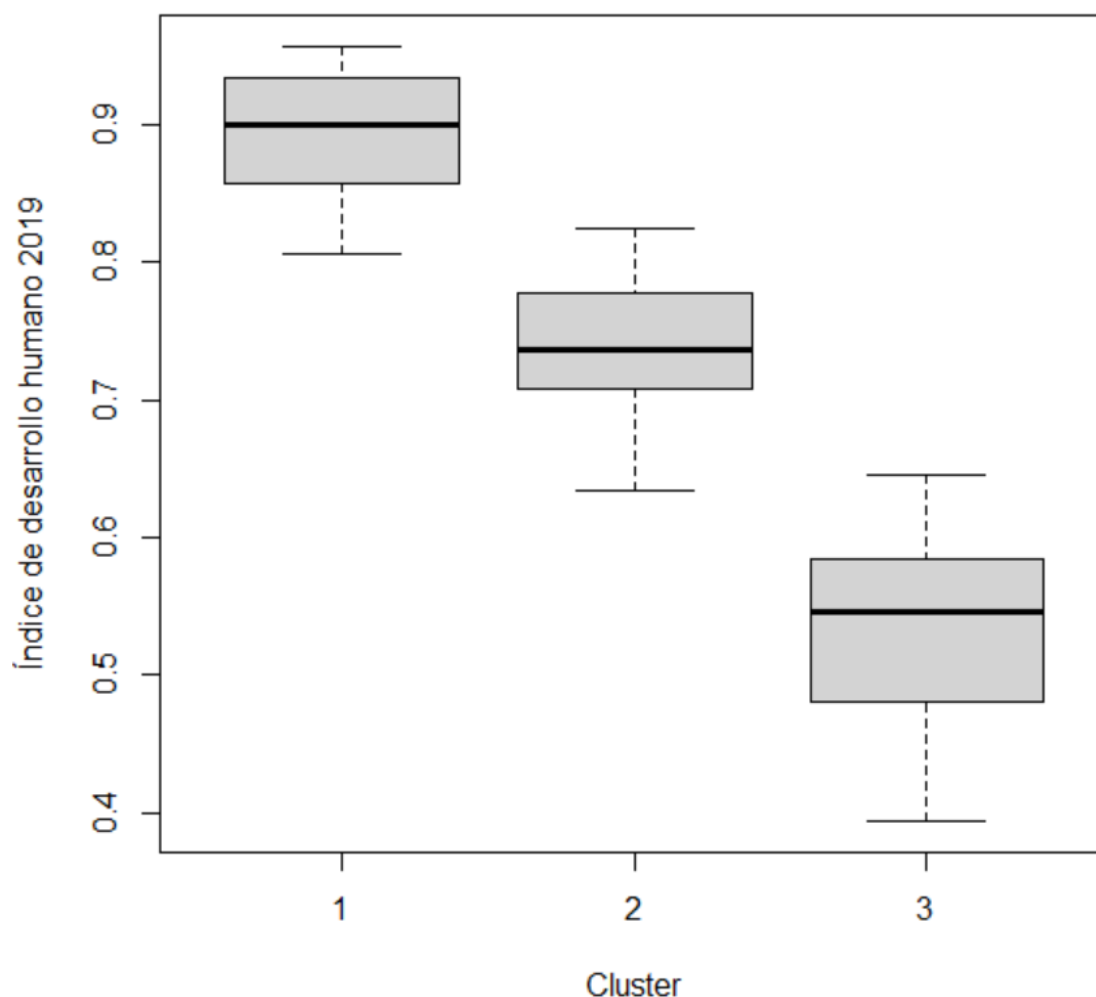
geográficos, económicos y culturales.

En algunas variables como la esperanza de vida al nacer, la educación y los ingresos; los diagramas de caja ponen al descubierto diferencias socioeconómicas entre las regiones. Los diagramas de caja permiten una comparación fácil de las medianas, rangos intercuartílicos y valores atípicos, entre los grupos. dando la posibilidad de identificar brechas o similitudes que explican la heterogeneidad de los grupos en el desarrollo humano global.

Al analizar en detalle los resultados de los diagramas de caja de cada uno los grupos clúster, se observan patrones distintivos en las distribuciones de los Índices de Desarrollo Humano (IDH). Ver figura 4-16.

En el Clúster 1, se evidencia una tendencia donde los valores del IDH son consistentemente más altos en comparación con los demás grupos (valores de IDH superiores a 0.8). Lo que sugiere que los países en el clúster 1 son países con un nivel de desarrollo humano superior (países desarrollados); como ya se dijo los países en este grupo pertenecen en su mayoría a Norteamérica y Europa, y otros tantos como Rusia, Australia, Japón, Nueva Zelanda y Arabia Saudita. Este grupo tiene una distribución asimétrica positiva con una mediana que se encuentra en 0.9.

Ahora bien, el clúster 2 muestra una distribución un poco más simétrica de los valores del IDH, que sugiere una variabilidad equilibrada entre los países en cuanto a su nivel de desarrollo humano. Este grupo concentra los países que denominaremos países de desarrollo intermedio; con una mediana de IDH ubicada en 0,75. Estos países se ubican en

Figura 4-16: *Boxplot Índice de desarrollo humano (IDH)*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

Latinoamérica y Asia, principalmente.

Finalmente, el clúster 3 se caracteriza por tener los valores de IDH más bajos y una dispersión alta en desarrollo humano, encontrando valores de desarrollo humano desde 0,4 hasta 0,65, que indica un desempeño bajo de los países en esta categoría, en términos de desarrollo humano (países con bajo desarrollo). Este grupo en su mayoría está compuesto por países del continente africano y la zona de India y Pakistán.

Un análisis específico del Índice de Educación (IE) permitirá identificar las brechas en los niveles educativos entre las distintas regiones. Los boxplot de cada uno de los clústeres muestran diferencias marcadas en los resultados del Índice de educación por grupos. Ver figura 4-17

En el clúster 1 - Países desarrollados, se evidencian valores altos y notoriamente superiores a los valores en los demás grupos. En este grupo la mediana alcanza un valor de 0.88 y el IE muestra valores superiores a 0,7; los países en esta agrupación poseen sistemas educativos sólidos y un acceso extendido a la educación.

El clúster 2 - Países con desarrollo intermedio, muestra una distribución un poco más normalizada en los valores del IE con una mediana en el IE igual 0,7, que sugiere que los países dentro de esta categoría tienen un rango diverso de niveles educativos y posibilidad de acceso a la educación. Este análisis del IE brinda una visión profunda de las diferencias educativas que subyacen en los grupos clúster identificados, revelando la importancia de la educación para el desarrollo humano.

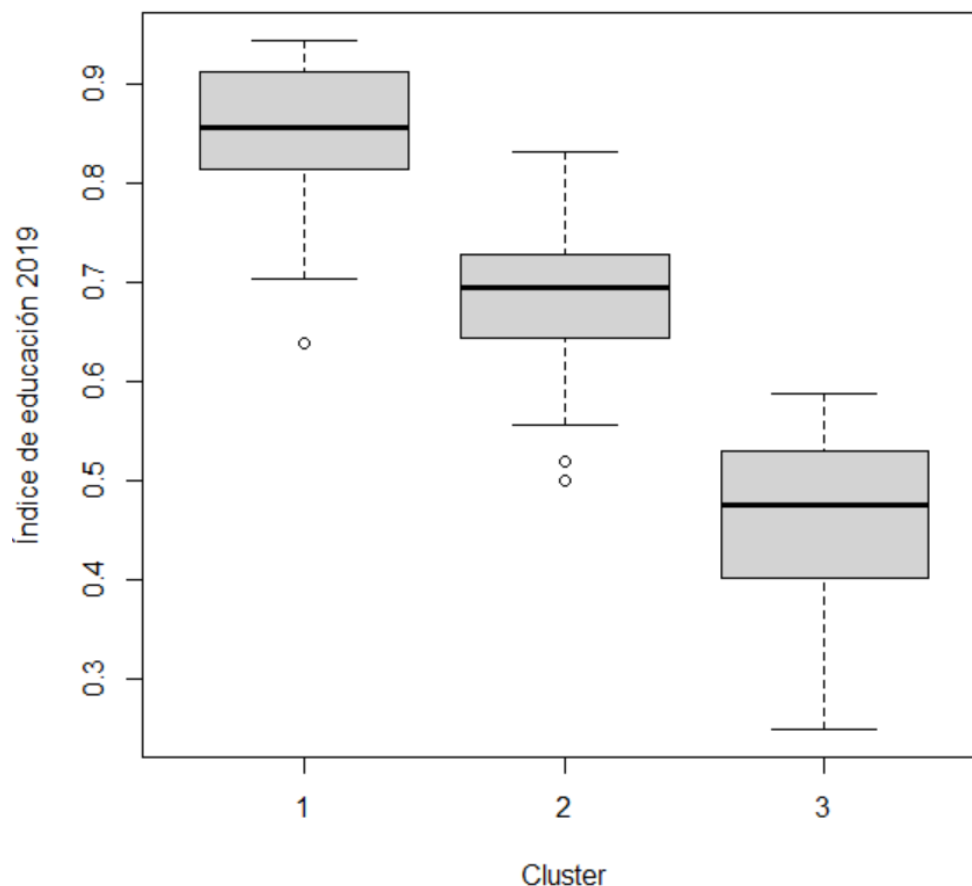
En contraste con el clúster 1; el clúster 3 - Países con bajo desarrollo, muestra la mediana más baja en el IE con una distribución asimétrica negativa y resultados dispersos, que en algunos casos son menores al 0,3 y van hasta 0,6 como máximo; la situación para estas naciones es difícil en cuanto a educación y presentan desafíos en términos de acceso y calidad educativa.

De la misma manera, al considerar dentro de este análisis la revisión del comportamiento de indicadores relacionados con el IE, tales como: la expectativa de años de escolaridad y el acceso de usuarios a internet, la tesis mencionada frente de las diferencias entre sistemas educativos se robustece, principalmente por los países dentro del grupo 3.

Dando un vistazo a los boxplot de expectativa de años de escolaridad **4-18**, para el clúster 1 - Países en desarrollo, se puede decir que su media supera los 16 años, es decir, que en el 50 % de los países de este grupo las personas alcanza estudios profesionales y de postgrado. En contraste, en el clúster 3 - Países con bajo desarrollo, la mediana esta alrededor de 11 años, con una distribución asimétrica negativa implica que el 50 % de los países en este grupo no alcanzan si quiera a completar sus estudios de secundaria. En este grupo se puede encontrar países con expectativa de años de escolaridad que apenas superan los 6 años. Por último, en el clúster 2 - Países con desarrollo intermedio, la expectativa de años de escolaridad presenta una mediana alrededor de 14 años y un mínimo de 10 años y una distribución un poco más simétrica.

También, se analizó los diagramas de caja de usuarios de internet **4-19**, se evidencia una disparidad en la accesibilidad digital. El primer clúster presenta una mediana que supera el 80 % de las personas, que indica una penetración alta de Internet. El clúster 2 muestra una mediana que supera 60 % de penetración con una amplia dispersión, y el clúster 3, cuya mediana apenas supera el 20 % de la población, denota una penetración muy baja de acceso a esta tecnología. Este análisis, en ese caso proporciona una visión holística de las diferencias en educación y conectividad digital entre los grupos, destacando una diversidad tecnológica a nivel global y amplias barreras en conectividad en países con bajo desarrollo.

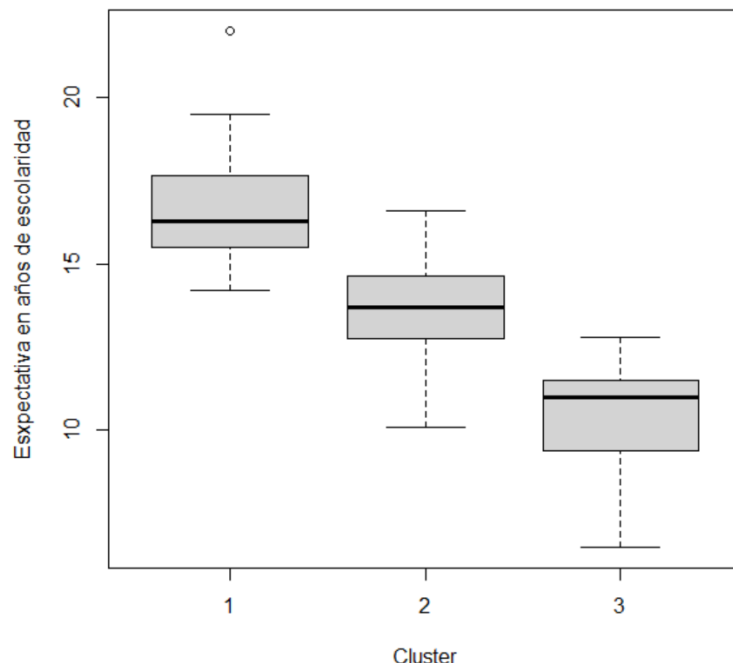
Ahora, se revisan los tres grupos clúster para el índice que mide la esperanza de vida

Figura 4-17: *Boxplot Índice de educación*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

al nacer (EVN), que refleja la calidad de vida y el bienestar de las poblaciones. Al examinar los diagramas de caja correspondientes a la esperanza de vida de cada clúster se identificar fácilmente el contraste de este indicador dejando al descubierto diferencias significativas en salud y longevidad. Ver figura 4-21

En el primer clúster, los valores de la esperanza de vida al nacer (EVN) son altos, ubicándose la mediana de la expectativa de vida en 82 años, que permite determinar de los países en esta categoría tienen poblaciones con una esperanza de vida más prolongada. En contraste, el tercer clúster muestra la mediana por debajo de 65 años de EVN, indicando una situación en la que los países tienen un nivel bajo de longevidad. En este grupo de países con bajo desarrollo, la expectativa de vida se encuentra entre los 55 años y los 74 años y la mediana de EVN es 18 años menos que la de los países desarrollados. Esto refleja problemas sociales como la dificultad de tener una buena alimentación, situaciones de desnutrición y maltrato infantil, el difícil acceso a un sistema de salud, la desigualdad social y la precariedad o carencia de vivienda, violencia, entre otros problemas sociales. El segundo clúster, por su

Figura 4-18: *Boxplot Expectativa de años de escolaridad*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

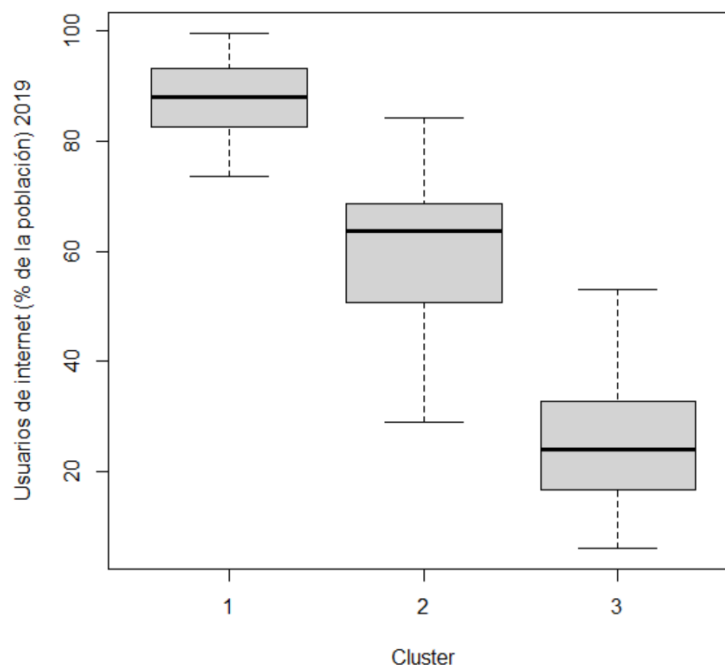
parte, muestra que el mínimo de años en la esperanza de vida al nacer es 11 años mayor que el mínimo de los países con bajo desarrollo, la mediana de este indicador se sitúa sobre 74 años. Figura 4-20.

Estos contrastes en la esperanza de vida al nacer en los tres grupos clúster, vislumbran una desigualdad en los niveles de salud y calidad de vida a nivel global, lo que es congruente con los resultados obtenidos en los diagramas de caja del indicador de desigualdad en la expectativa de vida.

En conjunto, este análisis destaca la diferencia entre la longevidad de las poblaciones y la urgente necesidad implementar medidas para eliminar las desigualdades en la salud a nivel global, así como la importancia de establecer políticas y mecanismos que promuevan un acceso equitativo a la atención médica y un aumento en la calidad de vida en todas las poblaciones.

Adicionalmente a los índices de esperanza de vida analizados anteriormente, se tienen representaciones graficas de los resultados de las variables asociadas a tasas de mortalidad e incidencia de tuberculosis.

Como definió anteriormente, la tasa de mortalidad infantil se refiere al número de muertes de niños menores de un año por cada 1,000 nacidos vivos en un año; este es un indicador importante de la salud y el bienestar de una población. Una tasa de mortalidad

Figura 4-19: *Boxplot usuarios de internet*

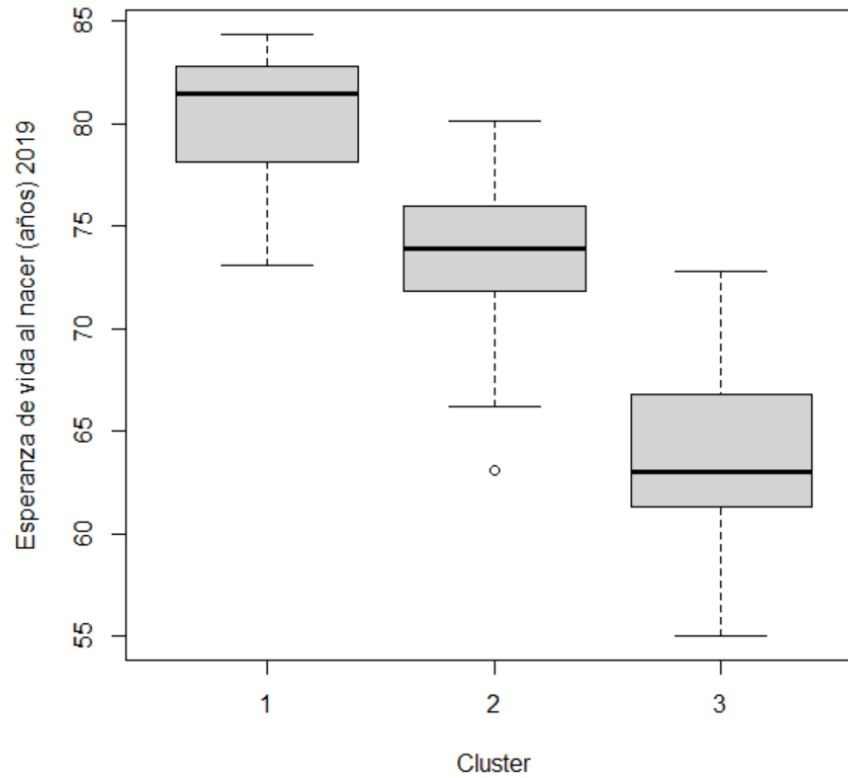
Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

infantil alta generalmente señala deficiencias en la atención médica materna, el acceso a servicios de salud y las condiciones socioeconómicas de un país. Este indicador posee ventajas comparativas relevantes que, a la hora de evaluar el impacto de las acciones en el bienestar y la supervivencia de los niños al desarrollar diversas actividades enfocadas en inmunización, agua potable, salud materna, nutrición, entre otros.

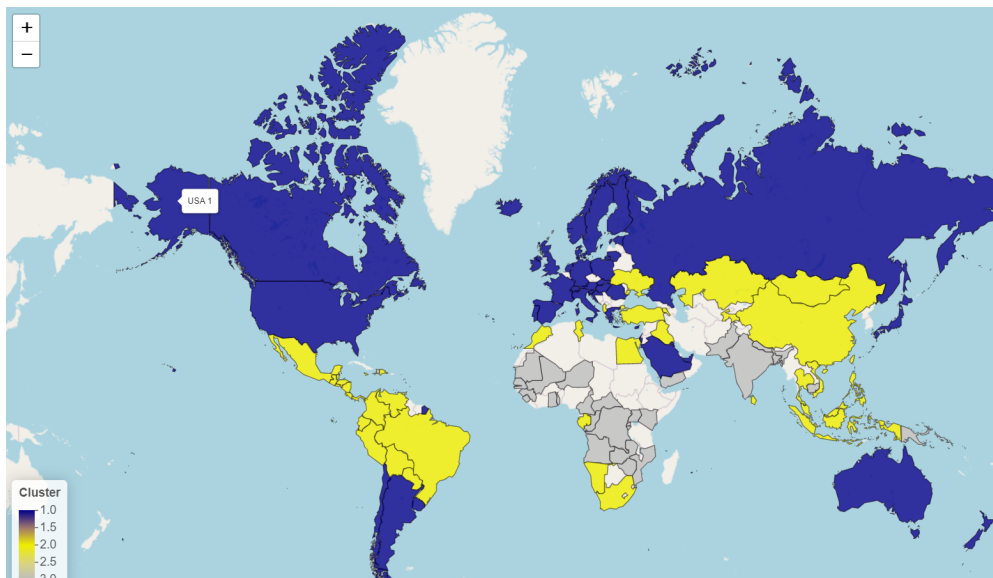
En este aspecto los grupos muestran diferencias marcadas. El clúster 1 - Países desarrollados muestra un diagrama de caja con una dispersión baja y una mediana alrededor de 5 muertes. El clúster 2 - países con desarrollo intermedio las muertes de infantes son considerablemente menores al grupo 2, sin embargo, su mediana se sitúa en 10 muertes doblando el valor del clúster 1.

Para finalizar este análisis, en el clúster 3 - países con bajo desarrollo, la distribución muestra una gran dispersión, con una mediana que se sitúa cercana a las 40 muertes de infantes y algunos países que superan las 70 muertes de niños. Figura 4-24

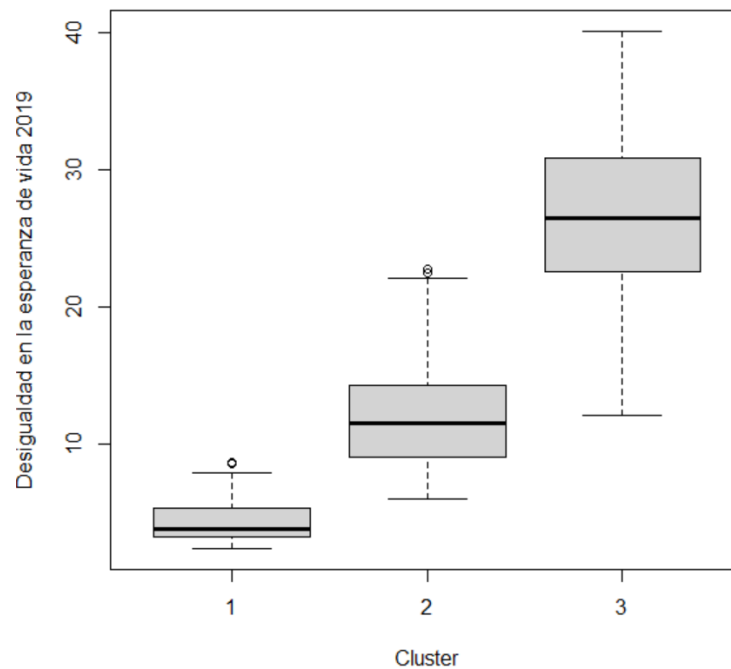
La figuras 4-23 y 4-22 permiten observar la existencia de una estrecha relación entre las tasas de mortalidad infantil y el comportamiento del IDH. La clasificación geográfica generada por la ONU con resultados de mortalidad infantil y los resultados obtenidos mediante la técnica de agrupación clúster empleada en este trabajo son congruentes.

Figura 4-20: *Boxplot Esperanza de vida al nacer*

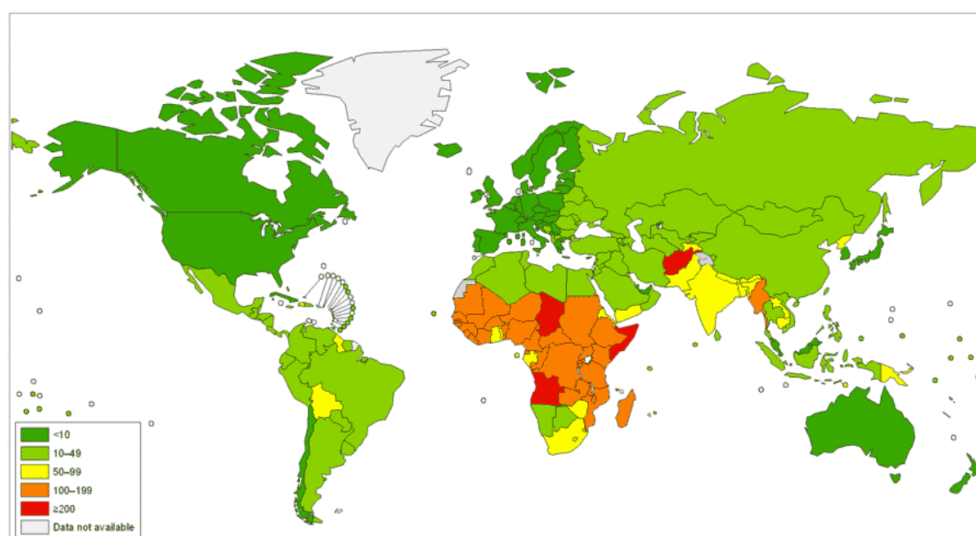
Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

Figura 4-22: *Geografía de clústeres.*

Nota: *Elaboración propia.*

Figura 4-21: *Boxplot Desigualdad en la expectativa de vida*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

Figura 4-23: *Geografía mortalidad infantil ONU.*

Nota: *Tomado de*
<https://www.humanium.org/es/enfoques-tematicos/salud/mortalidad-infantil/>

Además, las tasas de mortalidad en hombres y mujeres adultas también pueden tener un impacto en la esperanza de vida al nacer. En muchos países, las estadísticas han demostrado que las mujeres suelen tener una esperanza de vida mayor que los hombres. Esto se debe a varias razones, como diferencias biológicas, comportamientos de riesgo (como fumar y consumir alcohol), ocupaciones peligrosas y acceso diferencial a la atención médica. Cuando la tasa de mortalidad es más alta en un género específico, puede influir en la esperanza de vida general de la población.

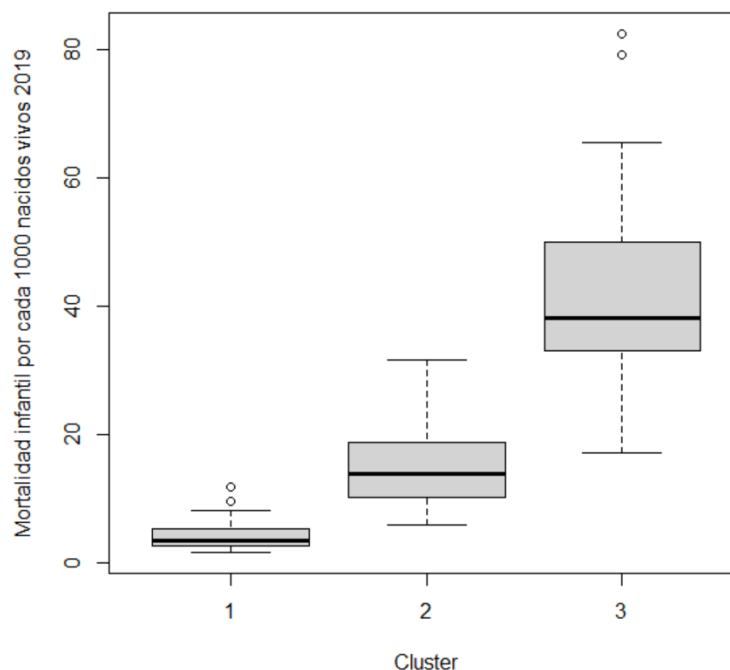
Las figuras **4-25** y **4-26** confirman el supuesto anterior, las tasas de mortalidad en la mujeres comparadas con las tasas de los hombres son inferiores en todos los clúster. Sin embargo, en países con bajo desarrollo las tasas de mortalidad de mujeres se cruzan y son similares con los valores de mortalidad para hombres.

En resumen, las tasas de mortalidad infantil y las tasas de mortalidad en hombres y mujeres adultas tienen un impacto directo en la esperanza de vida al nacer. Reducir la tasa de mortalidad infantil y abordar las diferencias en las tasas de mortalidad entre géneros pueden contribuir significativamente a aumentar la esperanza de vida en una población.

Otro factor que contrasta bastante en los grupos definidos es la incidencia de tuberculosis que calcula por cada 100 mil habitantes en las condiciones de vida. En este caso el clúster 1 de países desarrollados muestra una distribución cuya dispersión es muy baja y su mediana se aproxima a cero y se diferencia bastante de los clúster 2 y 3, que agrupan países de desarrollo intermedio y bajo desarrollo, estos 2 grupos presentan dispersiones altas, siendo el grupo 3 el más crítico con una mediana de 200 casos incidentes por cada 100 mil habitantes y países que pueden alcanzar hasta 700 casos incidentes de tuberculosis **4-27**, dejando entrever sistemas de salud precarios, con barreras de atención y gobiernos con carencia en la implementación políticas de salud que reduzcan el riesgo y la siniestralidad de estas poblaciones.

El acceso a tecnologías y combustibles limpios para cocinar tiene un impacto directo en la salud y la esperanza de vida de las poblaciones, especialmente en las regiones con recursos limitados. En muchas partes del mundo, el uso de combustibles sólidos como la madera, el carbón y el estiércol para cocinar, conlleva la emisión de contaminantes nocivos en el aire interior. Estos contaminantes incluyen partículas finas, monóxido de carbono y otros compuestos tóxicos que pueden causar enfermedades respiratorias, problemas cardiovasculares y afecciones pulmonares crónicas. La exposición continua a estos contaminantes está relacionada con una mayor incidencia de infecciones respiratorias agudas, así como con un mayor riesgo de padecer de enfermedades crónicas a largo plazo.

El uso estufas de cocina eficientes y energéticamente limpias, así como combustibles como el gas natural o la electricidad, puede reducir significativamente la exposición a estos contaminantes perjudiciales. Esto no solo mejora la calidad del aire interior en los hogares, sino que también disminuye el riesgo de enfermedades respiratorias y cardiovasculares. Además, al reducir la carga de enfermedades relacionadas con la contaminación del aire en la población, el acceso a tecnologías y combustibles limpios para cocinar puede contribuir a

Figura 4-24: *Boxplot tasa de mortalidad infantil*

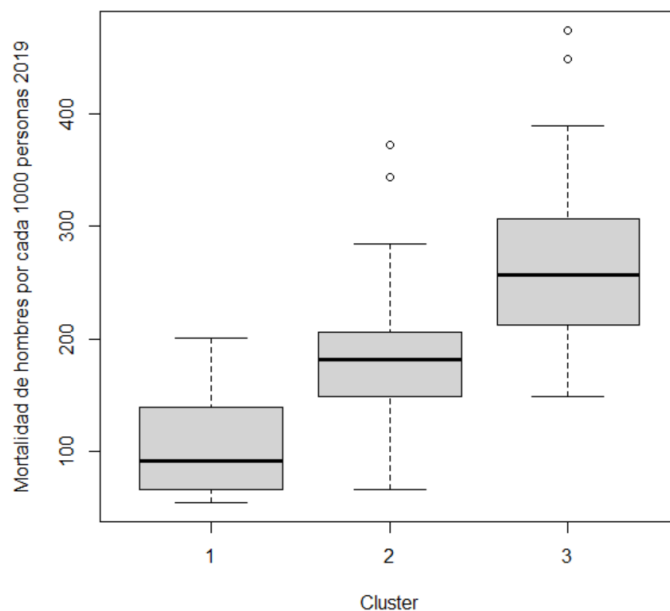
Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

un aumento en la esperanza de vida, permitiendo a las personas vivir vidas más saludables y productivas. Por lo tanto, la promoción de estas soluciones sostenibles no solo impacta positivamente en la salud de las personas, sino que también tiene implicaciones significativas para el desarrollo y el bienestar a largo plazo de las comunidades. Ver figura 4-28

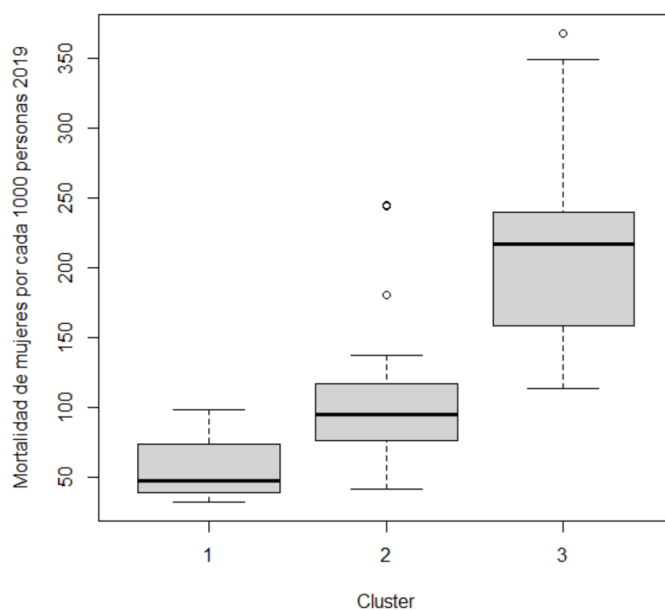
El boxplot relacionado con el índice de tecnologías limpias para cocinar permite inferir una relación estrecha entre el acceso a estas tecnologías y la longevidad de las personas. El clúster 1 muestra como los países desarrollados tienen en un 100 % solucionado el acceso a dichas tecnologías; en el clúster 2 la situación no está crítica pues su mediana es superior al 90 %; pero el clúster 3 si evidencia problema graves con un media menor a 40 %, que es bastante menor a la de los 2 primeros grupos y adicionalmente tiene una gran dispersión en este rubro, siendo desalentador el panorama para algunos países del grupo, donde el acceso a este tipo de tecnologías es casi nulo.

A continuación, se hace la revisión del comportamiento de las variables económicas en cada uno de los clúster. El propósito en este caso es detectar diferencias en los valores del PIB y sus implicaciones sobre el desarrollo humano. Ver figura 4-29

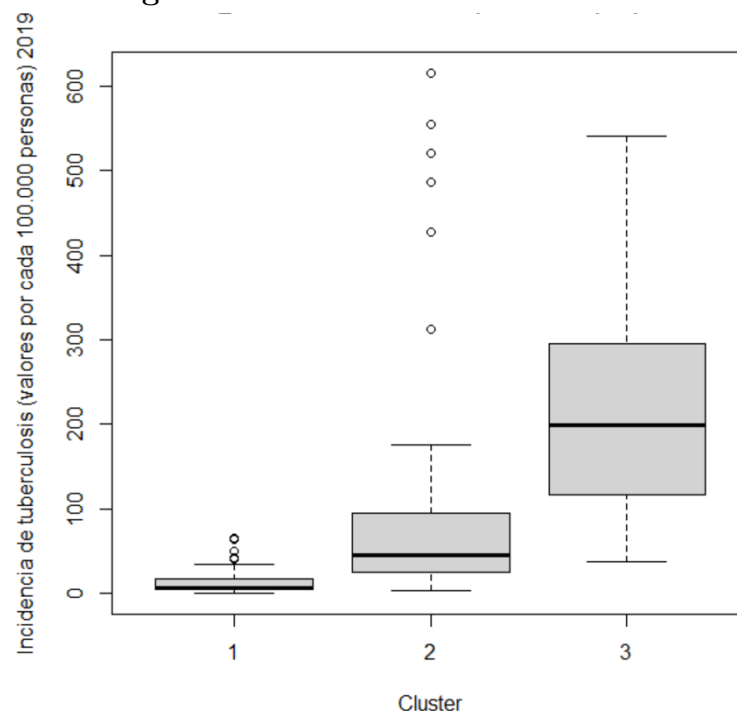
Para visualizar estas diferencias, se presentan los diagramas de caja para cada uno de los tres grupos determinados, lo que permitirá una comprensión de cómo el PIB influye en los diferentes niveles de desarrollo humano y cómo estas diferencias pueden estar relacionadas

Figura 4-25: *Boxplot tasa de mortalidad hombres adultos*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

Figura 4-26: *Boxplot tasa de mortalidad mujeres adultas*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

Figura 4-27: *Incidencia de tuberculosis*

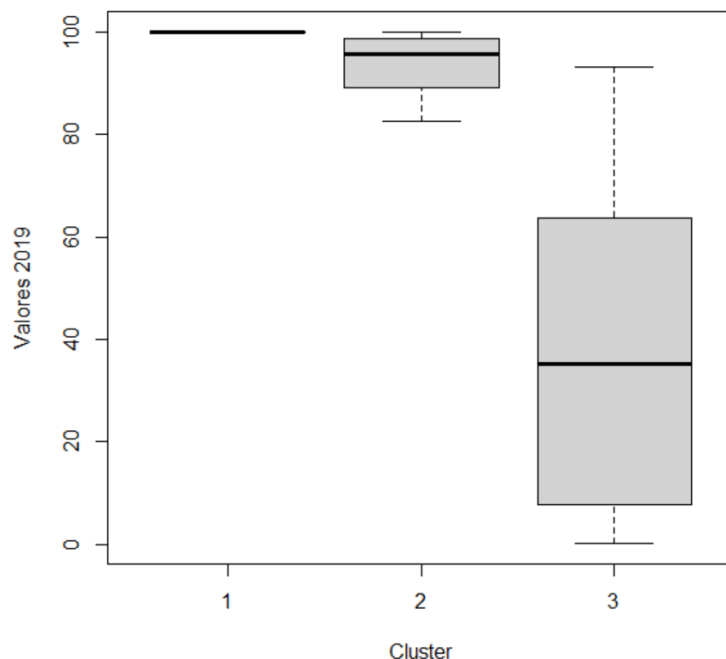
Nota: *Elaboración propia con información tomada de <https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>*

con diversos factores socioeconómicos y políticos.

El Clúster 1 se destaca por mostrar una variación amplia en sus valores de Producto Interno Bruto (PIB). Este grupo, conformado por países con las economías más robustas y desarrolladas, muestra una variación considerable en términos de cifras de PIB. Dicha variación puede ser atribuida a la diversidad de industrias, sectores y enfoques económicos adoptados por los países dentro de este grupo selecto, que a pesar de la variación en el PIB reflejan su fuerte posición económica a nivel global.

Por otra parte, el clúster 2 muestra un Producto Interno Bruto significativamente inferior al del grupo 1, pero se mantiene por encima de los valores del grupo 3. A pesar de esta diferencia en los valores de PIB; la brecha entre el clúster 2 y el clúster 3 no es sustancial. Este grupo podría estar conformado por países en vías de desarrollo con economías en crecimiento, pero que aún no han alcanzado el nivel económico de los países en el primer clúster. Aunque presentan una mejora relativa en términos económicos en comparación con el clúster 3, todavía enfrentan desafíos para cerrar la brecha con el grupo líder en términos de PIB. Este grupo podría estar en proceso de diversificar su base económica y mejorar su competitividad global.

Por otro lado, el clúster 3 se caracteriza por una escasa variación en los valores de PIB entre sus miembros, pero desafortunadamente reporta los menores valores de este indicador.

Figura 4-28: *Boxplot Acceso a tecnologías y combustibles limpios para cocinar*

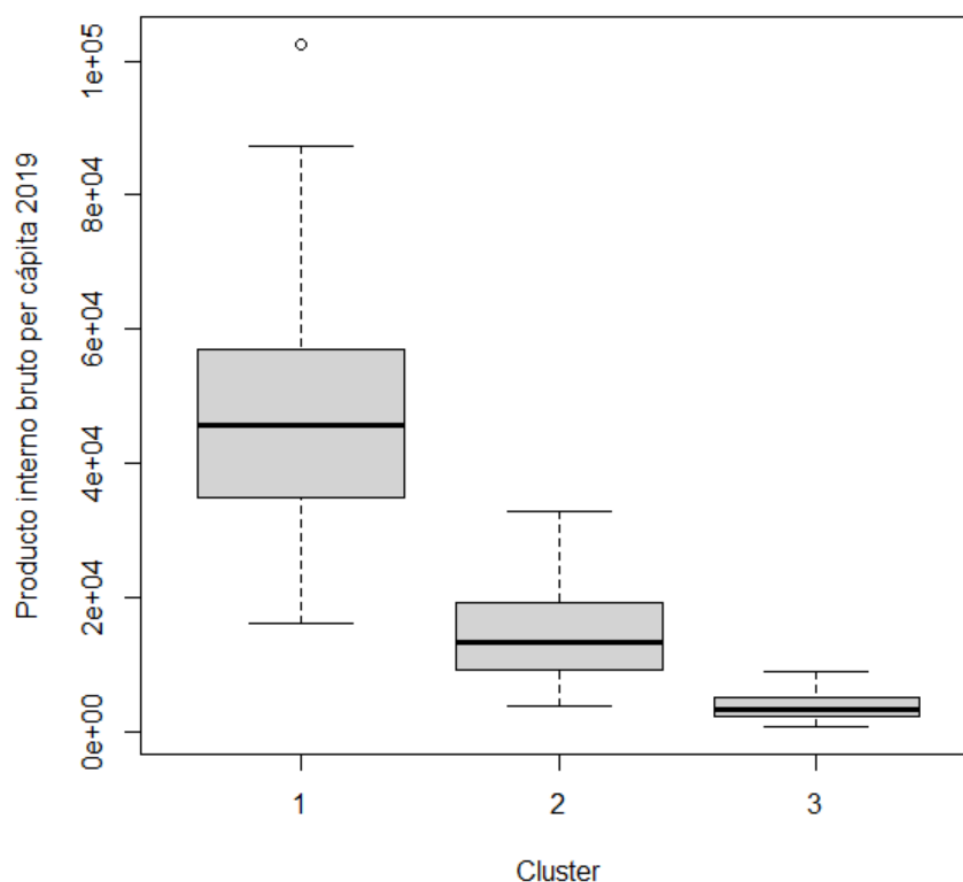
Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

Esta falta de diversidad económica puede derivar en una mayor vulnerabilidad ante cambios económicos adversos, ya que la dependencia en un rango estrecho de actividades económicas es limitante de oportunidades de crecimiento y resiliencia. Los países en este grupo podrían enfrentar desafíos como altas tasas de desempleo, y menor inversión en infraestructura y servicios públicos, así como dificultades para financiar programas de bienestar social.

En el análisis de estos clústeres en materia económica, es esencial considerar cómo el Producto Interno Bruto (PIB) se relaciona con otros indicadores clave. Uno de estos indicadores es el Ingreso Nacional Bruto (INB). En el primer clúster, donde se encuentran los países con los mayores valores de PIB, también se observa un alto INB. Esto indica que estos países no solo generan una gran cantidad de riqueza económica internamente, sino que también tienen una significativa entrada de recursos desde el extranjero. Este alto INB puede estar vinculado a inversiones extranjeras, remesas o ingresos por exportaciones, lo que contribuye al crecimiento económico sostenido.

En el clúster 2, donde el PIB es menor que en el grupo 1, pero mayor que en el grupo 3, el INB podría estar en un rango intermedio. Aunque estos países no disfrutan de la misma prosperidad económica que el grupo 1, su INB puede ser impulsado por un crecimiento económico relativamente constante y una diversidad económica en desarrollo.

En el tercer clúster, los valores de PIB son bajos y uniformes, el INB tiende a ser proporcionalmente bajo también. La falta de diversidad económica y las limitadas oportuni-

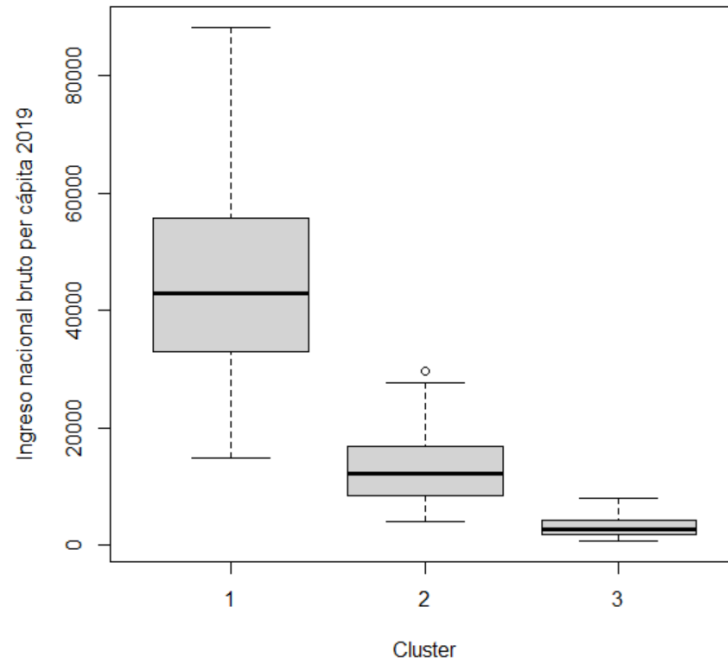
Figura 4-29: *Boxplot Producto Interno Bruto*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

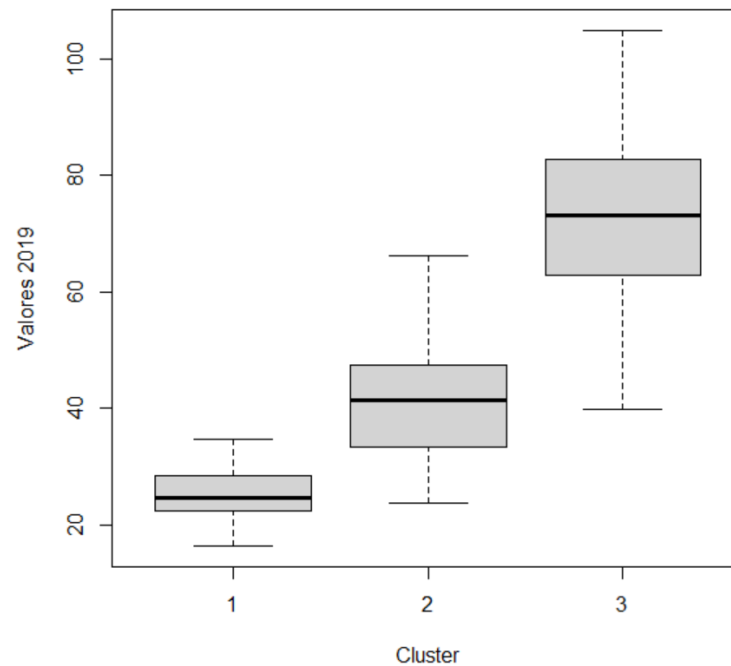
dades de ingresos internacionales pueden explicar esta correlación. Un INB bajo puede llevar a una menor inversión en educación, salud y otros servicios públicos, que a su vez puede aumentar la tasa de dependencia de los jóvenes en la población. La alta dependencia de los jóvenes en estos países puede convertirse en un obstáculo para el desarrollo económico a largo plazo, ya que se necesita inversión en la juventud para impulsar el crecimiento futuro.

Asimismo, es importante considerar el indicador de empleo vulnerable al analizar estos clústeres. En el clúster 3, donde los valores de PIB son bajos y la dependencia de los jóvenes es alta, es probable que también se encuentren altos niveles de empleo vulnerable. La falta de empleos estables y bien remunerados puede llevar a un aumento en la informalidad laboral y la falta de seguridad económica.

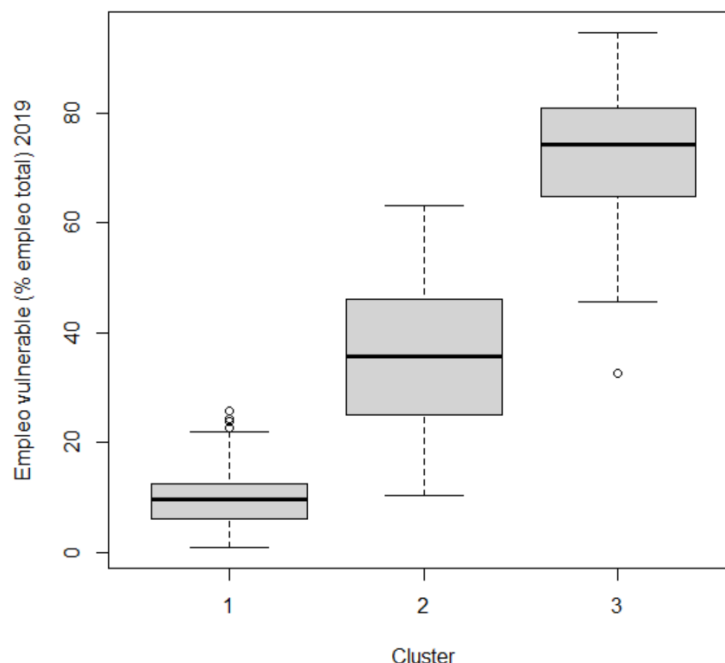
En resumen, el PIB de cada clúster está intrínsecamente relacionado con el INB, la tasa de dependencia de los jóvenes y el empleo vulnerable. Estos indicadores juntos proporcionan una imagen completa de la salud económica de cada grupo, y destacan las áreas que

Figura 4-30: *Boxplot Ingreso Nacional Bruto*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

Figura 4-31: *Boxplot Ratio de dependencia de jóvenes*

Nota: *Elaboración propia con información tomada de*
<https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>

Figura 4-32: *Boxplot Empleo Vulnerable*

Nota: *Elaboración propia con información tomada de <https://hdr.undp.org/data-center/human-development-index#/indicies/HDI/>*

requieren atención para lograr un crecimiento sostenible y equitativo.

Finalmente, se puede decir que este enfoque de análisis de clúster proporcionó una comprensión más profunda de las diferencias y similitudes del progreso humano, entre regiones del mundo.

4.7. Correlación de spearman

Para finalizar este apartado, dentro del proceso estadístico llevado a cabo en este estudio, se evaluó la correlación entre funciones empleando la correlación de Spearman 4-4. Esta técnica se permitió evaluar las relaciones entre las funciones de las 32 variables asociadas al Índice de Desarrollo Humano (IDH) seleccionadas anteriormente. El objetivo principal era identificar aquellas variables que mostraban correlaciones significativas, con valores absolutos superiores a 0.4, lo que indica relaciones fuertes y potencialmente influyentes en el resultado del IDH.

Este procedimiento permitió reducir el conjunto de variables funcionales a 20 covariables funcionales relevantes que demostraron estar intrínsecamente relacionadas con las diversas dimensiones del IDH, y se convirtieron en los elementos fundamentales para llevar

a feliz término los ejercicios de clasificación funcional para las curvas del IDH, garantizando que el proceso de clasificación se soportará en información de alta calidad y relevancia para la caracterización y diferenciación de los niveles de desarrollo humano por país.

Tabla 4-4: *Tabla comparativa de correlaciones*

Covariable funcional	Correlación spearman	Correlación Pearson
Desigualdad en la expectativa de vida	-0.96	-0.93
Esperanza de vida al nacer	0.93	0.93
Índice de educación	0.95	0.96
Expectativa de años de escolaridad	0.94	0.91
Emisiones de CO2	0.57	0.11
GDP per cápita	0.94	0.82
Ingreso nacional bruto per cápita	0.95	0.82
Índice de ingresos	0.95	0.13
Total de usuarios de internet	0.96	0.93
Suscripciones a telefonía móvil	0.84	0.56
Ratio de dependencia de jóvenes	-0.89	-0.88
Población urbana	0.77	0.69
Tasa de mortalidad infantil	-0.96	-0.89
Tasa de mortalidad mujeres adultas	-0.90	-0.82
Tasa de mortalidad hombres adultos	-0.80	-0.76
Incidencia de tuberculosis	-0.81	-0.57
Inmunización contra Sarampión	0.61	0.57
Acceso a electricidad en ciudades	0.89	0.69
Acceso a combustibles limpios y tecnologías limpias para cocinar en ciudades	0.89	0.80
Empleo vulnerable	-0.90	-0.87

Nota: *Elaboración propia*

El anterior proceso permitió identificar correlaciones fuertes del IDH principalmente con el total de usuarios internet, la tasa de mortalidad infantil, la desigualdad en la expectativa de vida, el ingreso nacional bruto per cápita y los ingresos, cuyas correlaciones de Spearman superaron en valores absolutos el 0.95. Además, se observa una notoria mejoría en las correlaciones del índice de ingresos y las emisiones de CO2 con el IDH, cuando se aplica la correlación Spearman frente a la correlación lineal de Pearson.

5 Modelado con datos funcionales

En la Sección 3.2.1 se abordó la teoría en cada uno de los principales modelos desarrollados en la clasificación de series de tiempo funcionales. Ahora bien, en este capítulo se llevará a cabo la explicación y ajuste de aquellos modelos implementados en este estudio, cuyas características se adaptaron a los datos previamente seleccionados y preparados. Asimismo, se determina la técnica más apropiada para hacer clasificación funcional del índice de desarrollo humano.

Los modelos de clasificación nos facilitan en este caso identificar grupos de funciones del IDH con características similares según la información proporcionada por una muestra. En el capítulo 4, mediante la aplicación de técnicas de análisis clúster se identificó plenamente los 3 grupos siguientes:

1. Clúster 1: Países desarrollados.
2. Clúster 2: Países desarrollo intermedio.
3. Clúster 3: Países con bajo desarrollo.

En términos funcionales, se dispone de conjuntos de datos de 20 covariables funcionales, y en cada una de ellas se tiene un tamaño $n = 119$ curvas funcionales (una curva por cada país), con datos discretizados d_j que van de 10 a 30 (uno por cada año). Por otra parte, las curvas del IDH y otros índices tienen un rango definido en $T = [0, 1]$, en razón a que su naturaleza es de carácter porcentual. Sin embargo, esta no es una regla general para todas las covariables, entendiendo que existen otras variables con escalas no porcentuales.

Todo lo anterior, permite definir un modelo de clasificación funcional **scalar on función**, dado que la variable de respuesta Y es escalar y la totalidad de covariables X a emplear en el modelo tienen naturaleza funcional (Ramsay and Silverman, 2005).

$$Y = m(Z, X) + \epsilon$$

donde $Y \in \mathbb{R}$ es la variable respuesta escalar con los 3 grupos clúster (1,2 y 3), m es la función desconocida de la combinación de covariables, $Z \in \mathbb{R}^P$ es la covariables del vector, $X = X_j(t)_{j=1}^q$ son las covariables funcionales y ϵ es el error de media cero. El vector de covariables es como sigue:

- X_1 : Desigualdad en la expectativa de vida.
- X_2 : Esperanza de vida al nacer.
- X_3 : Índice de educación.
- X_4 : Expectativa en años de escolaridad.
- X_5 : Emisiones de CO₂.
- X_6 Producto interno bruto (GDP) per cápita
- X_7 Ingreso nacional bruto per cápita.
- X_8 Índice de ingresos.
- X_9 Total de usuarios de internet.
- X_{10} Suscripciones a telefonía celular.
- X_{11} Ratio de dependencia juvenil.
- X_{12} Población urbana.
- X_{13} Tasa de mortalidad infantil.
- X_{14} Tasa de mortalidad mujeres adultas.
- X_{15} Tasa de mortalidad hombres adultos.
- X_{16} Incidencia de tuberculosis.
- X_{17} Inmunización contra Sarampión.
- X_{18} Acceso a electricidad en ciudades.
- X_{19} Acceso combustibles limpios y tecnología para cocinar en ciudades.
- X_{20} Empleo vulnerable.

Para $n = 119$ (países) con 20 covariables funcionales, se aplicó un total de 11 técnicas de clasificación funcional, buscando determinar cuál de estas es la más apropiada para predecir con precisión el grupo de desarrollo humano al que pertenece cada nación.

Los algoritmos de clasificación fueron implementados haciendo uso de las librerías **fda.usc**, **classiFunc**, y **FDboost** disponibles en R. En la tabla **5-1**, se mencionan las técnicas empleadas en este caso.

Tabla 5-1: Algoritmos de clasificación empleados

Paquete	Algoritmo	No. variables predictoras
classiFunc	classif.knn	Variable predictora única
	classif.kernel	Variable predictora única
fda.usc	classif.glm	Múltiples variables predictoras
	classif.gsam	Múltiples variables predictoras
	classif.gkam	Múltiples variables predictoras
	classif.rpart	Múltiples variables predictoras
	classif.svm	Múltiples variables predictoras
	classif.randomForest	Múltiples variables predictoras
	classif.nnet	Múltiples variables predictoras
	classif.cv.glmnet	Múltiples variables predictoras
	classif.gbm	Múltiples variables predictoras
Medidas	Precisión	

Nota: *Elaboración propia*

Las curvas funcionales en este apartado se refieren a comportamientos del desarrollo humano en países desarrollados, países con desarrollo intermedio y países con bajo desarrollo; durante un período t de 30 años. En total son 20 covariables funcionales que son curvas con datos anualizados para un total de 119 países. .

Las covariables son de interés en cada una de las dimensiones que mide el desarrollo humano, a saber, la Esperanza de vida, la expectativa de escolarización y el ingreso per cápita.

La respuesta escalar Y es el tipo de categoría de país según el desarrollo humano. La muestra n de tamaño igual a 119 países se ha dividido de manera aleatoria en 2 grupos:

1. Muestra de entrenamiento: La muestra de entrenamiento se constituye con una muestra aleatoria del 80 % de las curvas funcionales de n es decir, un tamaño $n_{train} = 95$.
2. Muestra de prueba: En este caso la muestra n_{test} se conformó con el 20 % de las funciones restantes con un total de $n_{test} = 24$. Finalmente, sobre estas funciones se calculan los errores de clasificación.

5.1. Modelo de clasificación KNN funcional

En línea con lo encontrado en la revisión de bibliografía se implementó un algoritmo de clasificación funcional K-Nearest Neighbors (KNN) en el entorno de programación R, se utilizó una parametrización con 1NN que significa que el algoritmo KNN consideró únicamente 1 vecino más cercano para realizar la clasificación. También se parametrizó la distancia Euclídea y deformación dinámica del tiempo de las trayectorias (dtwPath) del pa-

quete **dtwPath**. lo anterior con el fin de comprobar lo encontrado en la literatura que el modelo se comporta mejor con dtw.

La elección de N puede influir significativamente en la precisión y la sensibilidad del modelo, ya que un valor bajo como 1 puede hacer que el modelo sea muy sensible a pequeñas variaciones en los datos funcionales, además de tener que asumir un alto costo computacional que requiere el algoritmo DTW para el cálculo de las distancias. La descripción y el funcionamiento del algoritmo KNN-DTW puede consultarse en detalle en <https://developersbay.se/time-series-classification-an-overview/#ucr/>

La técnica de deformación dinámica de series temporales (Dynamic Time Warping, DTW) para medir la similitud entre las series funcionales es especialmente útil cuando las series temporales tienen una velocidad de cambio variable o cuando hay desplazamientos temporales entre ellas.

El método se aplicó con todas las covariables funcionales X_i con el fin de comparar los resultados y determinar las covariables más eficientes para clasificar las funciones del IDH.

El modelo se entrena utilizando el 80 % los registros en cada covariable funcional X_i . Con esta parametrización se obtienen las diferentes tasas de aciertos por cada una de las variables funcionales como se puede ver en la **5-2**.

Los resultados muestran mejores márgenes de precisión cuando se ejecuta el algoritmo con la distancia Ecuclidea y se emplean covariables funcionales como la expectativa en años de escolaridad, la tasa de usuarios de internet, el índice de educación y el acceso a combustibles y tecnologías limpias para cocinar; cuyos escenarios permitieron obtener precisiones que superan el 80

5.2. Modelo de clasificación Kernel funcional

De igual manera que en el modelo KNN, el modelo kernel se configuró un modelo de clasificación Kernel con **dtwPath** (Deformación dinámica del tiempo en las trayectorias) y distancia Euclidea, con esta configuración el modelo considera tanto las diferencias en la forma de las curvas funcionales como sus desplazamientos temporales. El modelo crea un estimador Kernel funcional eficiente y calcula la distancia por deformación dinámica temporal para cada función en un conjunto de datos de curvas de referencia (Pfisterer et al., 2019); resultando en una clasificación altamente sensible a las variaciones sutiles en los datos funcionales. **5-2**. En este caso, el algoritmo kernel aporta mejores resultados con DTW en gran parte de los casos, mostrando ser más eficiente para clasificar las curvas del desarrollo humano.

Las covariables funcionales que generan mejores resultados con el algoritmo Kernel son la tasa de acceso de usuarios de internet, el índice de educación, la expectativa de años de escolaridad, el acceso a combustibles y tecnologías limpias para cocinar y la tasa de mortalidad

infantil.

Tabla 5-2: Márgenes de precisión. Modelos funcionales KNN - NP Kernel por covariable funcional

Indicador / Covariable funcional	% precisión KNN (Dist. Euclidea)	% precisión Kernel(Dist. Euclidea)	% precisión KNN (DTW)	% precisión Kernel (DTW)
Total usuarios de internet	91,67	91,67	83,33	91,67
Índice de educación	87,5	87,5	79,17	87,50
Expectativa de años de escolaridad	83,33	83,33	83,33	83,33
Acceso en ciudades a combustibles limpios y tecnologías para cocinar	91,67	87,5	83,33	83,33
Tasa de mortalidad infantil	75	62,5	75,00	83,33
Ratio de dependencia de juvenil	79,17	79,17	79,17	79,17
Ingreso nacional bruto per cápita (constant 2017 PPP\$).xlsx	75	29,17	75,00	75,00
Indice de ingresos total	79,17	83,33	75,00	75,00
Suscripciones a telefonía móvil	75	45,83	75,00	75,00
Producto interno per cápita, PPP(current international \$)	70,83	29,17	70,83	75,00
Acceso a electricidad en ciudades	70,83	58,33	70,83	75,00
Desigualdad en la esperanza de vida	79,17	79,17	66,67	66,67
Incidencia en tuberculosis	50	33,33	62,50	54,17
Emisiones de CO2	58,33	29,17	58,33	58,33
Empleo vulnerable	62,5	62,5	58,33	54,17
Esperanza de vida al nacer	62,5	62,5	54,17	58,33
Población urbana.xlsx	54,17	54,17	54,17	54,17
Tasa de mortalidad de mujeres adultas	58,33	25	50,00	50,00
Inmunización contra sarampión	54,17	37,5	50,00	50,00
Tasa de mortalidad de hombres adultos	54,17	29,17	41,67	41,67

Nota: *Elaboración propia*

No obstante, los buenos resultados obtenidos con el modelo kernel, el modelo KNN iguala en resultados a kernel en 10 tipos de covariables, los mejores resultados al emplear el modelo KNN se obtiene con las covariables: total usuarios de internet, expectativa de años de escolaridad, acceso en ciudades a combustibles limpios y tecnologías para cocinar, índice de educación y ratio de dependencia juvenil.

Además, los resultados de la tabla 5-2 comprueban que la deformación dinámica temporal aunque proporciona buenos resultados en ambos modelos, no siempre garantiza las mejores predicciones de clasificación para las curvas de IDH principalmente en KNN, probablemente porque las curvas de IDH no presentan una alta velocidad de cambio.

5.3. Modelo Funcional Lineal Generalizado (FGLM)

Esta extensión de los modelos lineales generalizados permite trabajar con covariables funcionales, utilizando funciones de enlace funcionales para modelar la relación entre las variables predictoras y la variable de respuesta funcional.

De acuerdo con (Febrero-Bande and González-Manteiga, 2011), el modelo FGLM se consiste en reemplazar la combinación lineal de las covariables en por el producto interno de las mismas en el espacio funcional. Aquí $Z = \{X^i\}_{i=1}^p$ refiere el conjunto de covariables funcionales con valores en el producto de infinitas dimensiones en el espacio de Hilbert.

Dado que este método de clasificación permite la utilización de múltiples variables funcionales, se utilizó las 20 covariables seleccionadas a lo largo del documento, las cuales tienen asociadas al IDH, además se configura el modelo con una función de activación Logit de binomial, dado que de acuerdo a la literatura esta función permite relacionar las covariables con la variable de respuesta multinomial que clasifica los grupos de curvas del IDH.

El modelo se entrena utilizando el 80 % los registros en cada covariable funcional X_i .

Con todo lo anterior, la configuración del modelo y los resultados obtenidos son cómo se muestran en la tabla 5-3

Tabla 5-3: Parámetros modelo FGLM

Expresión	Parámetros
$E(Y Z) = \mu = g^1(\eta_z) = g^1(\beta_0 + (Z, \beta))$ con $Z = \{X^i\}_{i=1}^p$	Y: Clúster (1,2,3) $X : X_1, X_2, \dots X_{20}$ β : Coeficientes Familia: Binomial Función de enlace (g): Logit Precisión: 37,5 %

Nota: *Elaboración propia*

Siendo Z para este caso el conjunto de las 20 covariables funcionales del IDH, β es un parámetro funcional, y g la función de enlace que entre el valor esperado μ y la componente sistemática η .

5.4. Modelo Funcional Aditivo Generalizado (FGSAM)

Este modelo comparte las similitudes con el modelo funcional generalizado FGLM, con la diferencia que el GAM es más flexible y puede modelar relaciones no lineales y no paramétricas entre las variables predictoras funcionales y la variable de respuesta. Utiliza funciones suavizadas, como splines o funciones de base penalizadas, para capturar patrones no lineales en los datos.

El modelo se entrena utilizando el 80 % los registros en cada covariable funcional X_i .

La configuración del modelo se relaciona en la tabla 5-4

Tabla 5-4: Parámetros modelo FGSAM

Expresión	Parámetros
$E(Y Z) = g^{-1}(\eta_z) = g^{-1}(\beta_0 + \sum_{j=1}^p f_j(X^j))$ Siendo, $f_j(X^j) = \sum_{k=1}^p f_j^k(\varepsilon_j^k)$	Y: Clúster (1,2,3) $X : X_1, X_2, \dots, X_{20}$ f_j : Funciones suavizadas de las covariables X^j del IDH. Familia: Binomial Función de enlace (g): Logit Precisión: 41,67 %

Nota: *Elaboración propia*

En esta variante lo principal es la estimación de las funciones parciales f_j que permiten el uso de la suma de funciones suavizadas para el cálculo de una probabilidad posterior.

El modelo FGSAM de clasificación es especialmente útil cuando las relaciones entre las variables predictoras funcionales y las categorías de respuesta no pueden ser capturadas de manera efectiva mediante modelos lineales o cuando se necesita una mayor flexibilidad para modelar relaciones no lineales en datos funcionales. Una ampliación sobre implementación de este modelo puede encontrar consultar en Febrero-Bande and González-Manteiga (2011)

5.5. Modelo Funcional Kernel Aditivo Generalizado (FGKAM)

El FGKAM es un modelo aditivo, lo que significa que modela las relaciones entre las variables predictoras funcionales y la variable de respuesta como la suma de funciones suavizadas no paramétricas. En lugar de utilizar funciones suavizadas tradicionales, como

splines o funciones de base penalizadas, el GKAM utiliza kernels. Los kernels son funciones que permiten suavizar y modelar relaciones no lineales en datos funcionales.

La estimación de las funciones parciales f_j se realiza mediante kernel funcional y se considera la respuesta como continua. Febrero-Bande and González-Manteiga (2011) muestra una implementación detallada de la implementación de esta clase de algoritmos.

En este caso el modelo se entrena utilizando el 80 % los registros en cada covariable funcional X_i .

La configuración del modelo se relaciona en la tabla 5-5

Tabla 5-5: Parámetros modelo FGKAM

Expresión	Parámetros
$P(Y_i = k X_i) = \frac{\exp(f_k(X_i))}{\sum_{i=1}^k \exp(f_j(X_i))}$	Y: Clúster (1,2,3) $X_i : X_1, X_2, \dots, X_{20}$ Variables predictoras del IDH Familia: Binomial fk: función suavizada basada en kernels fj: función suavizada de X_i Precisión: 75 %

Nota: *Elaboración propia*

La función de enlace logit es una elección común en problemas de clasificación debido a sus propiedades que relacionan las probabilidades con la combinación lineal de los predictores y las funciones suavizadas, y también es ampliamente utilizada en el contexto de modelos GAM y FGKAM para problemas de clasificación funcional.

5.6. Modelo de clasificación funcional de partición Recursiva (RPART)

El modelo de clasificación funcional rpart”se basa en el algoritmo CART (Classification and Regression Trees) y se utiliza para problemas de clasificación, donde el objetivo es predecir la pertenencia a una de varias categorías o clases. Aunque el enfoque original de CART se desarrolló para datos tabulares, se puede adaptar para problemas de clasificación funcional, donde las observaciones son curvas o funciones en lugar de valores individuales. Aquí se explica cómo funciona el modelo de clasificación funcional RPART:

Datos Funcionales: En lugar de tener datos tabulares con filas y columnas, en el caso de la clasificación funcional, los datos consisten en observaciones que son curvas o funciones

$X_1, X_2, \dots, X_{20}$. Cada observación se asocia con una etiqueta de clase o categoría que se busca clasificar la curva Y_i del grupo de desarrollo al que pertenece un país en este caso.

El algoritmo de partición recursiva utiliza un enfoque de división recursiva para dividir las funciones en grupos más homogéneos con respecto a las categorías de respuesta. La idea es encontrar las divisiones en las funciones que minimizan la impureza o el error de clasificación en la curvas del IDH.

Para este caso

$$G(x_i) = \sum_{j=1}^M p(x_{ij}) G(x_{ij})$$

Donde x_{ij} es la covariable del IDH empleada para ramificar el árbol e i representa el grupo del IDH para determinado país.

$p(x_{ij})$ es la probabilidad de que X_i tome su j -ésimo valor.

El modelo se entrena utilizando el 80 % los registros en cada covariable funcional.

La parametrización del modelo es como sigue:

Cada nodo del árbol representa una región o conjunto de funciones, y las divisiones se realizan en función de una variable predictora funcional y un umbral específico. Se limita la profundidad del árbol a 20.

Para decidir cómo se dividen las funciones, el algoritmo utiliza criterios como la ganancia de información, la reducción de la impureza Gini o el error de clasificación, dependiendo de la configuración y los objetivos del usuario.

El árbol de decisión resultante tiende a ser complejo y puede estar sobre ajustado a los datos de entrenamiento. Por lo tanto, a través del parámetro CP se aplica un proceso de poda para simplificar el árbol y reducir el riesgo de sobreajuste.

Teniendo en cuenta los parametros más importantes para este algoritmo, con la ayuda de la librería **FDA.USC** se configura un modelo `classif.RPART` con las siguientes características:

Método tipo `class` para partición de problemas de clasificación, un `minsplit` de 20 el cuál permite definir el mínimo de observaciones requeridas para dividir un nodo en dos subnodos. Esto controla la complejidad del árbol y evita divisiones en nodos con muy pocas observaciones. El número mínimo de observaciones en un nodo hoja (`minbucket=7`), un parámetro de complejidad `cp= 0.1` (parámetro que controla la poda del árbol); un valor más alto de `cp` conduce a árboles más pequeños y simples.

La profundidad máxima permitida del árbol `maxdepth=20`, con el objetivo de controlar la profundidad máxima del árbol y ayudar a prevenir el sobreajuste. Finalmente, el número de validaciones cruzadas a realizar para seleccionar el tamaño óptimo del árbol y

controlar el sobreajuste `xval` se ajusta en un valor igual a 10.

La ejecución de este algoritmo mostro un buen comportamiento con una precisión del 70.83

5.7. Modelo Funcional Random Forest

El algoritmo que vamos a utilizar en esta ocasión es Random Forest, se plantea un bosque aleatorio con 100 árboles, con tamaño de nodo 1, el modelo se entrena con una muestra `ntrain` de tamaño igual a 95 (países) con 20 covariables funcionales presentadas durante el desarrollo del trabajo.

La función `classif.randomForest` del paquete `fda.usc` en R, se utiliza para ajustar modelos de clasificación de este tipo. Los parámetros a ajustar al configurar un modelo de clasificación funcional con `randomForest` son similares a los de una clasificación estándar.

El modelo se entrena utilizando el 80 % de los registros en cada covariable funcional X_j .

La parametrización del modelo es como sigue:

Como en los otros métodos el modelo se configuró para datos funcionales con las variables

$$Y_i, f_j(X_j).$$

Donde:

Y_i es la variables de respuesta. $f_j(X)$ es una la función suavizada $X : X_1, X_2, \dots X_{20}$ hace referencia a las 20 covariables funcionales asociadas al IDH.

Este modelo pertenece a la familia de los métodos de ensamble, por lo tanto, el modelo se parametrizó para que cada árbol dentro del bosque este compuesto por covariables distintas en cada caso. El modelo se parametrizó con `ntree = 100` árboles un mayor número de árboles puede mejorar la precisión, pero también aumenta el tiempo de entrenamiento y el costo computacional, 20 covariables funcionales.

Asimismo, el modelo también incorpora la importancia de las covariables (`importance=TRUE`) que pueden ayudar a identificar qué covariables son más informativas. En este caso las covariables corresponden al intervalo del dominio de los predictores.

Adicionalmente se utiliza el parámetros **proximity** cuyo valor lógico que indica si se deben calcular las matrices de proximidad entre las observaciones; el parámetro **nodessize = 3** para indicar el número mínimo de observaciones en un nodo terminal (hoja) de un árbol que controla la complejidad del árbol y puede ayudar a prevenir el sobreajuste; y finalmente ,

maxnodes para configurar el número máximo de nodos terminales (hojas) en un árbol. Los parámetros adicionales se dejaron con los valores que trae la librería **fda.usc** de (Febrero-Bande and Oviedo de la Fuente, 2012) por defecto. Una explicación más detallada de cómo funciona la versión funcional de Random Forest puede consultarse en (Möller et al., 2016)

Los resultados obtenidos al clasificar funciones del IDH con Random forest son los mejores obtenidos entre todas la técnica empleadas con una tasa de aciertos del 100 % al clasificar las curvas de los países contenidos en la muestra de test.

5.8. Modelo de clasificación Funcional SVM

El modelo de clasificación funcional multiclase SVM (Support Vector Machine) es una extensión de los SVM para problemas de clasificación comunes, en los cuales hay más de dos clases y los datos se representan de manera funcional, es decir, cada observación es una función continua en lugar de un vector de características.

El modelo se entrena con el 80 % de los registros en cada covariable funcional X_j .

La parametrización del modelo es como sigue:

El modelo se configuró con datos funcionales con las variables

$$Y_i, f_j(X_j).$$

Donde:

Y_i es la variables de respuesta. $f_j(X)$ es una la función suavizada $X : X_1, X_2, \dots, X_{20}$ hace referencia a las 20 covariables funcionales asociadas al IDH.

En este modelo, se utiliza un núcleo que puede ser lineal, radial, sigmoide, entre otros, que se utiliza para transformar las funciones originales en un espacio de características de mayor dimensión. Este núcleo lineal se utiliza para llevar a cabo la clasificación en el espacio de características lineales entra las covariables. El objetivo principal sigue siendo el de encontrar hiperplanos de separación óptimos en el espacio de características. Sin embargo, en un problema multiclase, hay varios hiperplanos de separación, uno para cada par de clases y también existen vectores de soporte asociados a cada hiperplano de separación.

Adicionalmente, los modelos de clasificación funcional SVM multiclase incluyen un parámetro de costo (C) que controla el equilibrio entre maximizar el margen de separación y permitir errores de clasificación. A menudo, se utiliza una estrategia "uno contra el resto" (One-vs-Rest) para resolver problemas multiclase en SVM, donde se entrena un clasificador binario para cada clase.

El modelo se ejecutó utilizando la función **classif.SVM** del paquete '**fda.usc** de (Febrero-Bande and Oviedo de la Fuente, 2012). Los parámetros principales que fueron

configurados son los que se mencionan a continuación.

kernel: Especifica el tipo de kernel a utilizar en el SVM, se permite elegir entre varios tipos, como "linear" (lineal), radial" (RBF, función de base radial), "poly" (polinómico), entre otros. Según la bibliografía, el kernel radial es una elección común; sin embargo, se utilizó el valor el kernel lineal posteriormente de evidenciar mejores resultados con este valor.

cost: Controla el costo de las penalizaciones por errores de clasificación. El valor predeterminado es 1 y fue empleado de esta manera para la configuración del SVM

gamma: Parámetro para el kernel radial. Por lo tanto, no aplica para la parametrización de este ejercicio que se decanto por incluir un kernel lineal para mejorar la clasificación.

los resultado son efientes siendo la segunda técnica con los mejores resultados alcanzado un 95,83 % de curvcas bien clasificadas según el comportamiento del IDH.

(Rossi and Villa, 2006) presenta un paso a paso de esta tecnica en datos funcionalesy una guia completa sobre el desarrollo de SVM.

5.9. Modelo de clasificación funcional Fit Neural Networks (nnet)

Este modelo de clasificación funcional se basa en redes neuronales artificiales (NN). Los parámetros de un modelo de clasificación funcional configurado con nnet son similares a los de un modelo de clasificación estándar.

Se configura el modelo con un máximo de 1000 iteraciones para controlar al número de veces que se actualizan los pesos en la red, el parámetro de entropía se deja por defecto y asigna la entropía como función de perdida con el concepto de mínimos cuadrados, se parametriza una capa oculta de tamaño 6 y la tasa de aprendizaje se deja por defecto igual (decay = 0)

Uno de los aspectos fundamentales en la configuración es el número de neuronas en la capa oculta, que se establece mediante el parámetro size. La elección de este valor implica un compromiso entre la capacidad del modelo para capturar patrones complejos en los datos y la prevención del sobreajuste. Un valor mayor de size aumenta la capacidad del modelo, pero también puede aumentar el riesgo de sobreajuste si no hay suficientes datos de entrenamiento.

Una vez configurada la función se entrena el modelo y se evaluar su eficiencia para clasificar funciones del IDH, la precisión generada no es buena con 37,5

5.10. Modelo Lineal Generalizado Lasso and Elastic-Net Regularizado (GLMNET)

El modelo `cv.glmnet` en R es una extensión de la regresión logística regularizada que se utiliza para resolver problemas de clasificación con regularización L1 (Lasso) o L2 (Ridge). Cuando se aplica a datos funcionales, los parámetros son similares a los de la regresión logística regularizada, pero se adapta al contexto de datos funcionales.

Se parametrizó un modelo con intercepto, con un máximo de 1000 iteraciones, `nfolds=10` y parámetros `alpha=0` y `lambda = 100`. En este caso, Alpha controla la mezcla entre la regularización L1 (Lasso) y L2 (Ridge). Un valor de 1 indica Lasso puro, 0 indica Ridge puro, y valores entre 0 y 1 representan una mezcla de ambos tipos de regularización. Por otra parte, lambda controla la fuerza de la regularización; este algoritmo realiza una búsqueda automática para encontrar el mejor valor de lambda utilizando validación cruzada.

El número de particiones en las que se divide el conjunto de datos durante la validación cruzada `nfolds = 10`, ayuda a evaluar la capacidad de generalización del modelo.

Type.measure: La métrica utilizada para evaluar el rendimiento del modelo en cada iteración de la validación cruzada. En este caso usa la opción `class` que permite evaluar precisión del modelo.

Standardize: Un valor lógico que indica si debes estandarizar las covariables funcionales antes de ajustar el modelo. La estandarización puede ser importante cuando las covariables tienen diferentes escalas, para este caso se permite la estandarización de los datos, dado la diferencia de escalas entre covariables.

5.11. Modelo de clasificación funcional Gradient Boosting Machine (GBM)

Un modelo de Gradient Boosting Machine (GBM) para datos funcionales es una extensión del GBM tradicional que se utiliza para el análisis de datos funcionales, donde los datos se consideran como curvas funcionales en lugar de características tabulares estáticas.

El modelo se parametrizó con un total de 500 árboles, una tasa de aprendizaje o shrinkage de 0.1 y profundidad del árbol igual a 3. La tasa de aprendizaje (shrinkage) controla la contribución de cada árbol al modelo final. Un valor más bajo significa que cada árbol contribuye menos al modelo final y, por lo tanto, puede requerir más árboles para lograr un buen rendimiento.

La Profundidad del árbol controla la profundidad máxima de cada árbol en el ensemble. Árboles más profundos pueden capturar relaciones más complejas en los datos, pero

también pueden llevar al sobreajuste.

La precisión obtenida con este método es bastante buena con 95,83 % de éxito en la clasificación.

5.12. Análisis comparativo de técnicas empleadas

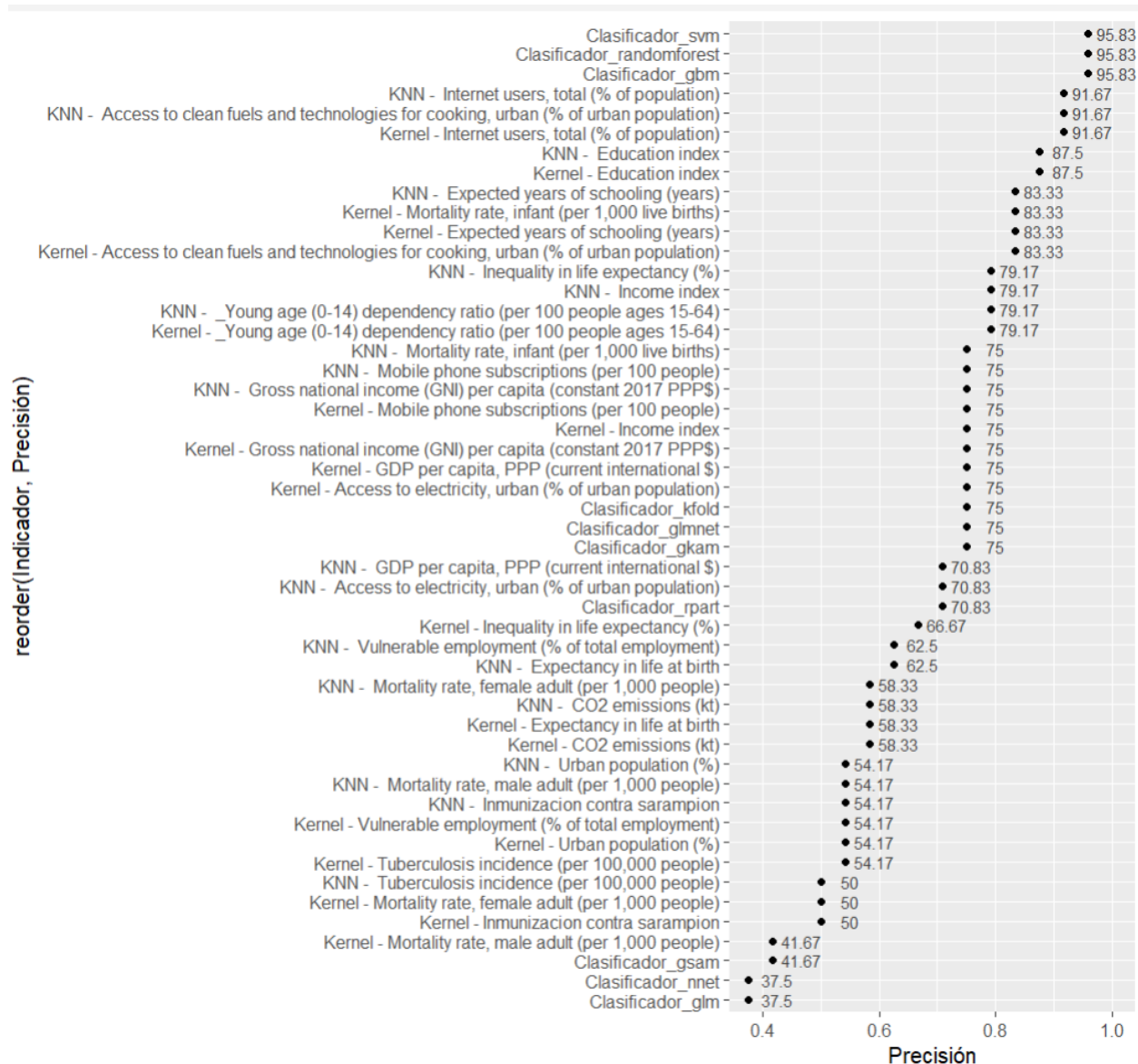
Para terminar, se procedió a la creación de un diagrama comparativo, con el fin de mostrar los resultados obtenidos después de aplicar 11 técnicas de clasificación de series de tiempos que comprenden métodos funcionales y métodos de Machine Learning funcionales a las funciones del Índice de Desarrollo Humano (IDH). Este diagrama comparativo permitirá visualizar de manera efectiva cómo se desempeñan las diferentes técnicas de clasificación funcional en la tarea de agrupar y caracterizar las curvas funcionales del IDH. Cada técnica puede destacar aspectos particulares de las curvas, como cambios de tendencia, divergencias regionales o similitudes temporales, proporcionando una comprensión más profunda y completa de la evolución del desarrollo humano a lo largo de los años.

En este estudio para la clasificación del IDH se abarcó diversas técnicas funcionales y de machine learning; para evaluar el rendimiento de cada ejercicio se comparó las proyecciones obtenidas contra los valores de la muestra de test, para esto se construyó la matriz de confusión de cada ejercicio y posteriormente se calculó la tasa de precisión de cada uno. Como parte de los hallazgos, el 24 % de los ejercicios mostraron resultados notables, con precisiones superiores al 80 %.

Dentro de las técnicas empleadas, se destacaron tres clasificadores en particular: Random Forest, Máquina de Soporte de Vectores (SVM) y Gradient Boosting. Dichas técnicas demostraron un rendimiento excepcional con resultados notables frente a los demás ejercicios con una precisión del 95.83 %.

Adicionalmente es importante destacar que las técnicas univariadas algoritmo 1NN con distancia Euclídea y Kernel con DTW, logran resultados de precisión del 91.67 % cuando se utilizan las covariables como total de usuarios de internet y acceso a tecnologías y combustibles limpios para cocinar. Además estas técnicas mostraron precisiones por encima del 80 % en 9 casos (18 %), lo que demuestra una buena capacidad y solidez para clasificar curvas de IDH, principalmente cuando se emplean las covariables como total usuarios de internet, índice de educación, años de escolaridad, acceso a tecnologías y combustibles limpios para cocinar y las tasas de mortalidad infantil.

Para terminar, se evidencia que los clasificadores con resultados más bajos en este estudio fueron el clasificador de red neuronal, el clasificador FGLM y el clasificador FGSAM. Los modelos univariados KNN y Kernel no mostraron buenos resultados al emplearse covariables como las tasas de mortalidad masculina y femenina, tasas de inmunización contra sarampión e incidencia de tuberculosis.

Figura 5-1: *Comparativo de eficiencia en técnicas clasificación empleadas*

Nota: Elaboración propia

6 Conclusiones y recomendaciones

6.1. Conclusiones

Los resultados obtenidos en este estudio respaldan de manera sólida y concluyente la eficacia de la clasificación de curvas del Índice de Desarrollo Humano (IDH). El análisis de datos funcionales es una técnica prometedora de buenos resultados en ejercicios de clasificación de curvas del IDH y demás variables asociadas a este índice.

La capacidad de los métodos de análisis de datos funcionales para modelar las curvas del IDH, no es poderosa únicamente en ejercicios de clasificación, este enfoque también aplica con buenos resultados en modelos aprendizaje no supervisado. En este trabajo se aplicó técnicas de reducción de dimensionalidad (FPCA) y análisis clúster cuyos resultados permitieron determinar 3 grupos socioeconómicos de interés (desarrollo alto, medio y bajo) y la caracterización de las naciones según su nivel de desarrollo y aspectos relevantes por diferentes dimensiones del IDH.

En los modelos univariados de clasificación funcional 1NN funciona mejor al parametrizarse con la distancia Euclídea para clasificar curvas de IDH en este caso DTW no es la mejor opción. Sin embargo, el algoritmo Kernel funciona mejor al combinarse con la deformación dinámica temporal DTW (tabla 5-2). De igual manera se comprobó que los modelos univariados KNN y Kernel funcionan mejor para clasificar las curvas de IDH, cuando se toma como variable explicativa el total de suscripciones de internet.

Los modelos multivariados con mejores resultados para clasificar funciones del IDH son los clasificadores Random forest, las Máquinas de soporte vectorial y el clasificador Gradient Boosting Machine, puesto que los resultados obtenidos con estas técnicas mostraron precisiones que superaron el 90 %.

Las técnicas de listas anidadas en R mejoran la eficiencia en la ejecución de los procesos, logrando una reducción significativa del costo computacional al optimizar los tiempos de respuesta de máquina cuando se trabaja con altos volúmenes de información y gran cantidad de bases de datos de manera simultánea. Lo anterior, se evidenció en la implementación del código desarrollado para este trabajo el cual permitió la carga inicial de 52 bases de datos y su análisis en paralelo.

6.2. Recomendaciones

Los resultados obtenidos en este estudio brindan una base sólida que se espera motive para continuar empleando técnicas de análisis de datos funcionales para la explicación y predicción de los resultados del Índice de Desarrollo Humano (IDH) a nivel local y global, y por extensión a las demás variables que se relacionan con el progreso de la humanidad. La eficacia demostrada en la clasificación de curvas del IDH, así como la capacidad de capturar y modelar la complejidad inherente a las dimensiones del IDH, revelan el potencial transformador de este enfoque estadístico avanzado.

Estos hallazgos pueden alentar a la comunidad académica y a los responsables de la toma de decisiones a continuar explorando y aplicando técnicas de análisis de datos funcionales en el estudio de los indicadores socioeconómicos.

Referencias

- Akinbobola, T. O. and Saibu, M. O. O. (2004). Income inequality, unemployment, and poverty in nigeria: a vector autoregressive approach. *The Journal of Policy Reform*, 7(3):175–183.
- Bagnall, A., Lines, J., Bostrom, A., Large, J., and Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660.
- Bagnall, A., Lines, J., Hills, J., and Bostrom, A. (2015). Time-series classification with cote: The collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2522–2535.
- Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting.
- de Boor, C. (1978). *A Practical Guide to Spline*, volume Volume 27.
- de las naciones unidas, O. (2021). Población.
- Deprez, P., Shevchenko, P. V., and Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal*, 7(2):337–352.
- DEVBAY. (2020). Time series classification – an overview.
- Faouzi, J. (2022). Time series classification: A review of algorithms and implementations. *Machine Learning (Emerging Trends and Applications)*.
- Febrero-Bande, M. and González-Manteiga, W. (2011). *Generalized Additive Models for Functional Data*, volume 22, pages 91–96.
- Febrero-Bande, M. and Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28.
- Golovkine, S. (2021). *Méthodes statistiques pour données fonctionnelles multivariées*. PhD thesis.
- Grenander, U. (1950). Stochastic processes and statistical inference. *Arkiv för matematik*, 1(3):195–277.

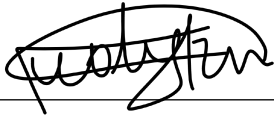
- Hernández-Sampieri, R. and Torres, C. P. M. (2018). *Metodología de la investigación*, volume 4. McGraw-Hill Interamericana México^eD. F DF.
- Hyndman, R. J. and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24(3):323–342.
- Hyndman, R. J. and Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3):199–211.
- Khan, Y. A., Abbas, S. Z., and Truong, B.-C. (2020). Machine learning-based mortality rate prediction using optimized hyper-parameter. *Computer Methods and Programs in Biomedicine*, 197:105704.
- Kokoszka, P. and Reimherr, M. (2017). *Introduction to functional data analysis*. Chapman and Hall/CRC.
- Lee, R. (2000). The lee-carter method for forecasting mortality, with various extensions and applications. *North American actuarial journal*, 4(1):80–91.
- Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting u.s. mortality. *Journal of the American Statistical Association*, 87(419):659–671.
- López, J. F. (2019). Guía para calcular e interpretar el IDH.
- Maierhofer, T. and Pfisterer, F. (2017). Classifunc: classification of functional data. *R package version 0.1. 0*, URL <https://CRAN.R-project.org/package=classiFunc>.
- Mark, R. (2018). Python implementation of KNN and DTW classification algorithm.
- Medina, F. (2001). Consideraciones sobre el índice de gini. Technical report, Naciones unidas.
- Möller, A., Tutz, G., and Gertheiss, J. (2016). Random forests for functional covariates. *Journal of Chemometrics*, 30(12):715–725.
- Mundial, B. (2021). Esperanza de vida al nacer, total (años).
- Mundial, G. B. (2023). Datos de libre acceso del banco mundial.
- Pfisterer, F., Beggel, L., Sun, X., Scheipl, F., and Bischl, B. (2019). Benchmarking time series classification—functional data vs machine learning approaches. *arXiv preprint arXiv:1911.07511*.
- PNUD (2019). Informe sobre desarrollo humano 2019. Technical report, Programa de las Naciones Unidas para el Desarrollo.
- PNUD (2020). Informe sobre desarrollo humano 2020. Technical report, Programa de las Naciones Unidas para el Desarrollo.
- Programme, U. N. D. (2023). HUMAN DEVELOPMENT DATA.

- Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47(4):379–396.
- Ramsay, J. O. and Silverman, B. W. (2005). Functional data analysis. *Springer Series in Statistics*, New York: Springer Verlag.
- Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17.
- RCODER. (2023). Listas en R.
- Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification. *Neurocomputing*, 69(7):730–742. New Issues in Neurocomputing: 13th European Symposium on Artificial Neural Networks.
- Shang, H. L. (2015). Statistically tested comparisons of the accuracy of forecasting methods for age-specific and sex-specific mortality and life expectancy. *Population Studies*, 69(3):317–335.
- Shang, H. L. (2016). Mortality and life expectancy forecasting for a group of populations in developed countries: a multilevel functional data method. *The Annals of Applied Statistics*, 10(3):1639–1672.
- Shang, H. L., Smith, P. W., Bijak, J., and Wisniowski, A. (2013). A functional data analysis approach for forecasting population: a case study for the united kingdom.
- Swamynathan, M. (2019). *Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python*. Apress.
- Wang, J.-L., Chiou, J.-M., and Muller, H.-G. (2015). Review of functional data analysis. *arXiv preprint arXiv:1507.05135*.

Declaración

Me permito afirmar que he realizado la presente tesis de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en la propia tesis. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido en el presente trabajo. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de tesis.

Bogotá, D.C., 10.10.2023

A handwritten signature in black ink, appearing to read 'Fredy Alexander Rivera Gómez', written over a horizontal line.

(Fredy Alexander Rivera Gómez)